



TECHNICAL UNIVERSITY OF KOŠICE
Faculty of Electrical Engineering and Informatics



Electrical Engineering and Informatics II

**Proceeding
of
the Faculty of Electrical Engineering and Informatics
of
the Technical University of Košice**

Editor: Dušan Kocur

March, 2011

Košice, Slovak Republic

ISBN 978-80-553-0611-7

Electrical Engineering and Informatics II

Proceeding of the Faculty of Electrical Engineering and Informatics
of the Technical University of Košice

Košice, Slovak Republic

March, 2011

ISBN 978-80-553-0611-7

Contents

Implementácia pásmového $\Sigma\Delta$ AČP na programovateľnom obvode PSoC <i>Martin SEKERÁK, Ján ŠALIGA, Linus MICHAELI, Marek GODLA</i>	1
New Approach for Measurement of Static Characteristics D/A Converter <i>Martin SEKERÁK, Linus MICHAELI, Ján ŠALIGA, Marek GODLA</i>	6
Modeling of Some Analog-to-Digital Converter Properties in LabVIEW <i>Ján ŠALIGA, Martin SEKERÁK, Linus MICHAELI, Marek GODLA, Michal ROSINA</i>	9
Experimentálny systém FSO pre 625nm RONJA <i>Pavol MIŠENČÍK, Ján TURÁN, Ľuboš OVSEŇÍK</i>	15
Experimentálny systém FSO pre 850 nm LightPointe Fligh Strata 155E <i>Pavol MIŠENČÍK, Ján TURÁN, Ľuboš OVSEŇÍK</i>	18
Experimentálny systém FSO pre 1550nm Fsona SONAbeam™ 155-E <i>Pavol MIŠENČÍK, Ľuboš OVSEŇÍK, Ján TURÁN</i>	21
Projektovanie a analýza FSO systémov s využitím programového balíka „FSO System Simulator” <i>Pavol MIŠENČÍK, Ľuboš OVSEŇÍK, Ján TURÁN</i>	25
Experimentálne pracovisko pre meranie hustoty hmly <i>Matúš TATARKO, Ľuboš OVSEŇÍK, Ján TURÁN</i>	29
Experimentálne pracovisko s optickým korelátorom <i>Tomáš HARASTHY, Ján TURÁN, Ľuboš OVSEŇÍK</i>	35
Teoretické základy práce s optickým korelátorom <i>Tomáš HARASTHY, Ľuboš OVSEŇÍK, Ján TURÁN</i>	39
Experimenty s transformačným systémom na rozpoznávanie dopravných značiek <i>Tomáš HARASTHY, Ján TURÁN, Ľuboš OVSEŇÍK</i>	44
Projektovanie OVKS s využitím programového balíka „OptSim” <i>Ľuboš OVSEŇÍK, Ján TURÁN</i>	48
Projektovanie OVKS s využitím programového balíka „ModeSYS” <i>Ján TURÁN, Ľuboš OVSEŇÍK</i>	54
Videodohľadový systém pre detekciu opustenia batožiny <i>Anna KAŽIMÍROVÁ KOLESÁROVÁ, Ľuboš OVSEŇÍK, Ján TURÁN</i>	58
Architektúra videodohľadového systému na detekciu opustenej batožiny <i>Anna KAŽIMÍROVÁ KOLESÁROVÁ, Ján TURÁN, Ľuboš OVSEŇÍK</i>	62

Experimentálne pracovisko pre zváranie optických vlákien <i>Euboš OVSEŇÍK, Ján TURÁN, Ľudovít HINTOŠ</i>	67
Experimentálne pracovisko pre meranie na optických vláknach pomocou OTDR <i>Ján TURÁN, Euboš OVSEŇÍK, Ľudovít HINTOŠ</i>	71
DCT vizuálneho objektu s predspracovaním hraničných blokov <i>Iveta GLADIŠOVÁ, Ján MIHALÍK</i>	75
CAN zbernica vo vložených aplikáciách na báze ARM procesorov STM32 <i>Martin PETRVALSKÝ, Miloš DRUTAROVSKÝ</i>	79
Syntetický procesor NIOS s galvanicky oddeleným sériovým rozhraním <i>Peter HUSNAJ, Miloš DRUTAROVSKÝ</i>	85
Vývojová doska na báze Altera FPGA Cyclone III <i>Jozef MACEJ, Miloš DRUTAROVSKÝ</i>	91
Radarové systémy v automobiloch <i>Ján GAMEC, Iveta GLADIŠOVÁ, Mária GAMCOVÁ</i>	97
Elektronické systémy automobilov a trendy ich vývoja <i>Ján GAMEC, Iveta GLADIŠOVÁ, Mária GAMCOVÁ</i>	103
OPNET Modeler: Testing of the New Integration Model of QoS and Security for MANET <i>Ján PAPAJ, Lubomír DOBOŠ, Anton ČIŽMÁR</i>	107
Prediktívne prepínanie videotokov v internetovej televízii <i>Branislav HRUŠOVSKÝ, Pavol KOČAN, Stanislav MARCHEVSKÝ, Martin LIPTAJ</i>	113
Metódy maskovania chýb založené na filtrácii pohybových vektorov <i>Branislav HRUŠOVSKÝ, Stanislav MARCHEVSKÝ, Bohumil FRANKO, Martin LIPTAJ</i>	119
Prehľad základných metód maskovania stratených paketov v priestorovej oblasti pri prenose videa zašumeným prostredím <i>Branislav HRUŠOVSKÝ, Stanislav MARCHEVSKÝ, Martin LIPTAJ</i>	125
Pracovisko so vzdialeným prístupom pre testovanie návrhov na báze rekonfigurovateľných obvodov <i>Juraj TALÁN, Miloš DRUTAROVSKÝ, Ľudmila MACEKOVÁ</i>	131
Klasifikácia metód digitálnej vodotlače <i>Peter GOČ – MATIS, Tomáš KANÓCZ, Patrik GALLO, Zita KLENOVIČOVÁ, Dušan LEVICKÝ</i>	136
Súčasný metódy digitálnej vodotlače vo videu v reálnom čase <i>Peter GOČ - MATIS, Tomáš KANÓCZ, Patrik GALLO, Dušan LEVICKÝ</i>	142
Násobička frekvencie troma pre UWB radarový systém v technológii SiGe BiCMOS <i>Martin LIPTAJ, Pavol GALAJDA, Branislav HRUŠOVSKÝ</i>	147
Parallel Processing Applied on Elliptic Curve Arithmetic in Affine Coordinates <i>Patrik GALLO, Peter GOČ-MATIS, Dušan LEVICKÝ</i>	153
Parallel Processing Applied on Elliptic Curve Arithmetic in Jacobian Projective Coordinates <i>Patrik GALLO, Peter GOČ-MATIS, Dušan LEVICKÝ</i>	159

Obrazová steganografia v komunikačných systémoch <i>Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ, Zita KLENOVIČOVÁ, Dušan LEVICKÝ . . .</i>	165
Slepá steganografia na báze DWT <i>Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ, Zita KLENOVIČOVÁ, Dušan LEVICKÝ . . .</i>	170
Steganografia s vyhľadáváním perceptívne vhodných oblastí v krycích obrazoch <i>Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ, Zita KLENOVIČOVÁ, Dušan LEVICKÝ . . .</i>	176
Steganografia v statických farebných obrazoch <i>Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ, Zita KLENOVIČOVÁ, Dušan LEVICKÝ . . .</i>	182
Tvorba modelu slovenského jazyka predoménovo orientovanú úlohu v systéme automatického rozpoznávania reči <i>Ján STAŠ, Jozef JUHÁR</i>	188
Detekcia akustických udalostí na báze rečových parametrizácií a GMM <i>Eva VOZÁRIKOVÁ, Anton ČIŽMÁR.</i>	193
Klasifikácia akustických udalostí na báze SVM <i>Jozef VAVREK, Eva VOZÁRIKOVÁ, Peter VISZLAY, Anton ČIŽMÁR.</i>	199
Porovnanie dekorelačných metód príznakových vektorov v akustickom modelovaní založenom na HMM <i>Peter VISZLAY, Matúš PLEVA, Jozef JUHÁR</i>	205
Bezdrôtové senzorové siete v automobiloch <i>Kamil ŠINDLERY, Stanislav MARCHEVSKÝ</i>	210
Manažment spotreby energie v inteligentných budovách pomocou multi-agentov <i>Kamil ŠINDLERY, Stanislav MARCHEVSKÝ</i>	215
Formal Description of Behaviour for Object-based Programs <i>Marián JENČÍK, Valerie NOVITZKÁ</i>	223
Integration Framework for Common Knowledge Representation from Models <i>Maroš TYRPÁK, Iveta TONHAUSEROVÁ, Zdeněk HAVLICE.</i>	229
The Process of Expressing Web Services by Terms of Category Theory <i>Csaba SZABÓ, Viliam SLODIČÁK, Viliam MEDE</i>	233
Action Graphs and Their Properties <i>Martina LAJOVÁ, Valerie NOVITZKÁ</i>	239
The Resource Character of Linear Logic <i>Pavol MACKO, Valerie NOVITZKÁ.</i>	244
3D Buildings Modeling and Visualization <i>Branislav SOBOTA, František HROZEK, Iveta REPICKÁ</i>	251
Stereoscopic Image Processing <i>Branislav SOBOTA, František HROZEK, Martin FEČKANIN.</i>	255
An Anomaly-Based Intrusion Detection System <i>Eva DANKOVÁ, Norbert ÁDÁM, Peter JAKUBČO</i>	260
Aplikačné prostredie pre správu systémov a služieb v počítačových sieťach <i>Eva DANKOVÁ, Norbert ÁDÁM, Peter JAKUBČO</i>	265

Simplification of the Real-Time Network Traffic Monitoring <i>Adrián PEKÁR, Juraj GIERTL, Martin RÉVÉS, Peter FECILAK</i>	272
Adaptive Platform for Video Streaming to Mobile Devices <i>Miroslav MICHALKO, Miroslav BIŇAS, František JAKAB</i>	278
Distributed Emulation Using GPGPU <i>Peter JAKUBČO, Eva DANKOVÁ</i>	284
SIP Based Multimedia Mobility in IP Networks <i>Jozef JANITOR, František JAKAB</i>	288
Approaches to a Problem of Anomaly Detection in User Behaviour in a Smart Home Environment <i>Marek NOVÁK, František JAKAB</i>	291
Modelovanie povrchových štruktúr <i>Norbert ÁDÁM</i>	296
Zvýšenie návštevnosti webu na základe Long Tail analýzy <i>Adela TUŠANOVÁ, Ján PARALIČ</i>	301
Rozpoznávanie slovných povelov s robotom NAO <i>Zlatko FEDOR, Peter SINČÁK</i>	305
Mobile Robot Navigation Using Fuzzy Cognitive Maps <i>Martin PALA, Ján VAŠČÁK</i>	309
Modifikácie siete MF ARTMAP pre zvýšenie presnosti klasifikácie a zvýšenie generalizačnej schopnosti <i>Peter SMOLÁR, Peter SINČÁK</i>	313
Riadenie pohybu robota NAO pri kope do lopty v robotickom futbale <i>Jaroslav TUHÁRSKY, Ján VAŠČÁK</i>	317
Control of Mobile Robot Joints Using Fuzzy Logic <i>Ján VAŠČÁK</i>	320
Nastavovanie parametrov fuzzy regulátora pre navigáciu robotov pomocou evolučných prístupov <i>Ján VAŠČÁK</i>	326
Využitie neurónových sietí pri extrakcii pravidiel pre fuzzy regulátor <i>Ján VAŠČÁK</i>	333
From Imitation Learning to Innovation in Designing Behaviors for Humanoid Robots <i>Mária VIRČÍKOVÁ, Peter SINČÁK</i>	338
Analýza historických dát z kolaboratívneho prostredia KP-LAB <i>Matúš FEDOR, František BABIČ, Stanislav PETRULA</i>	345
Vytvorenie B2B aplikácie postavenej na platforme .NET <i>Marek FURIMSKÝ, František BABIČ</i>	349
Predikcia predajnosti na knižnom trhu prostredníctvom vybraných metód objavovania znalostí <i>Eudmila LIZÁKOVÁ, František BABIČ</i>	355
Vytvorenie modelu pre predikciu úspešnosti internetových aukcií <i>Vladimír GAŠPAR, Adrián DRINGUŠ, František BABIČ</i>	360

Aplikácia DDE a OPC komunikačných protokolov v identifikácii a riadení reálneho systému <i>Štefan JAJČIŠIN, Anna JADLOVSKÁ</i>	365
Automatická analýza sentimentu textových zdrojov <i>Peter KONCZ, Ján PARALIČ</i>	371
Návrh automatického dopĺňania Wordnetu <i>Viliam ROČKAI</i>	377
Structural Analysis of Social Groups in Collaboration Network <i>Gabriel TUTOKY, Martin REPKA, Ján PARALIČ</i>	380
Využitie agentových simulácií v socioekonomickej analýze a ich možné rozšírenia <i>Peter BUTKA, Martin SARNOVSKÝ</i>	386
Analýza množín textových dokumentov v distribuovaných prostrediach <i>Martin SARNOVSKÝ, Peter BUTKA, Peter BEDNÁR</i>	392
Niektoré vlastnosti sociálnej siete európskych výskumných organizácií <i>Gabriel LUKÁČ</i>	397
Matlab-Based Tools for Analysis and Control of Inverted Pendula Systems <i>Slávka JADLOVSKÁ, Ján SARNOVSKÝ</i>	403
Extrakcia informácií pre podporu rozhodovania v oblasti obchodného práva <i>Peter SMATANA</i>	409
Solder Paste Dispensers and Their Parameters <i>Alena PIETRIKOVÁ, Michal KRAVČÍK, Juraj BANSKÝ</i>	413
Vplyv vlhkosti na funkčnosť súčiastok <i>Eubomír LIVOVSÝ, Juraj BANSKÝ, Alena PIETRIKOVÁ</i>	417
Ultrazvukovo kontaktované spoje pre elektrotechniku a ich kvalita <i>Igor VEHEC, Juraj BANSKÝ, Alena PIETRIKOVÁ</i>	422
Animated Components and Their Utilization in the Education of the Electronic Materials <i>Slavomír KARDOS, Igor VEHEC, Juraj BANSKÝ</i>	428
Melting and Solidification of Solder Alloys <i>Alena PIETRIKOVÁ, Juraj ĎURIŠIN</i>	431
Implementácia kamerového systému do výučby na Katedre technológií v elektronike <i>Dominik DEMETER, Juraj BANSKÝ, Eubomír LIVOVSÝ</i>	435
Harmonické potenciálové pole pri orientácii robotov v priestore <i>Michal KALAVSKÝ</i>	439
Neurónový estimátor rýchlosti v štruktúre vektorového riadenia <i>Peter NGUYEN, Jaroslava ŽILKOVÁ</i>	443
Nepriamy jednosmerný menič s mäkkým spínaním so sekundárnym riadeným usmerňovačom <i>Marcel BODOR, Jaroslav DUDRIK</i>	449
Obojsmerný menič pre hybridnú batériu v elektromobiloch <i>Tomáš BÉREŠ, Jaroslav DUDRIK, Erik EÖTVÖS</i>	455

Jednosmerný rezonančný menič pre obnoviteľné zdroje <i>Erik EÖTVÖS, Jaroslav DUDRIK, Tomáš BÉREŠ</i>	460
Riadenie viacúrovňového meniča pre fotovoltické aplikácie pomocou programu RT-LAB <i>Marek PÁSTOR, Jaroslav DUDRIK</i>	465
Vektorové riadenie AM s využitím UI <i>Jaroslav TIMKO, Jaroslava ŽILKOVÁ, Michal KOVÁČ</i>	471
Analýza kumulatívnej chyby autonómneho vozidla pri inkrementálnom mapovaní priestoru <i>František BANÍK, Pavol FEDOR, Daniela PERDUKOVÁ</i>	476
Analýza vlastností elektromagnetického nástroja na gravírovavnie kameňa <i>Mišél BATMEND</i>	481
Dielektrická relaxačná spektroskopia 6 kV statorových cievok bez polovodivej ochrany <i>Eudóit CSÁNYI, Roman CIMBALA</i>	485
Jednosmerné charakteristiky nanokeramik <i>Juraj KURIMSKÝ</i>	491
Modelovanie výbojovej činnosti na vn cievke 6kV asynchrónneho motora <i>Milan KVAKOVSKÝ, Irida KOLCUNOVÁ, Juraj KURIMSKÝ</i>	495
Analysis of Voltage Variation in Low Voltage Networks <i>Bystrík DOLNÍK</i>	500
Príspevok k štatistickému vyhodnoteniu prepätí v sieti nízkeho napätia <i>Bystrík DOLNÍK</i>	504
Meranie výbojovej činnosti na cievkach s termosetickou izoláciou a s polovodivou ochranou <i>Marián HRINKO, Irida KOLCUNOVÁ, Juraj KURIMSKÝ</i>	509
Použitie kapacitných senzorov pre snímanie čiastkových výbojov <i>Jaroslav PETRÁŠ, Jozef BALOGH, Jaroslav DŽMURA</i>	513
Morphological Complexity of State Space Quaternary Memory <i>Milan GUZAN, Radoslav BUČKO</i>	518
The Crossing Numbers of Join of Special Graphs <i>Marián KLEŠČ, Daniela KRAVECOVÁ, Jana PETRILLOVÁ</i>	522
Kombinovaná segmentácia - inicializácia aktívnej kontúry pomocou Graph cut metódy <i>Peter KARCH, Marián BAKOŠ, Iveta ZOLOTOVÁ</i>	528
Vzdialené monitorovanie a supervízne riadenie na báze aplikačného a webového servera <i>Iveta ZOLOTOVÁ, Miloš PAVLÍK, Roman MIHAL, Miroslav KOLLÁR</i>	532

Implementácia pásmového $\Sigma\Delta$ AČP na programovateľnom obvode PSoC

Martin SEKERÁK, Ján ŠALIGA, Linus MICHAELI, Marek GODLA

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

martin.sekerak@tuke.sk, jan.saliga@tuke.sk, linus.michaeli@tuke.sk
marek.godla@tuke.sk

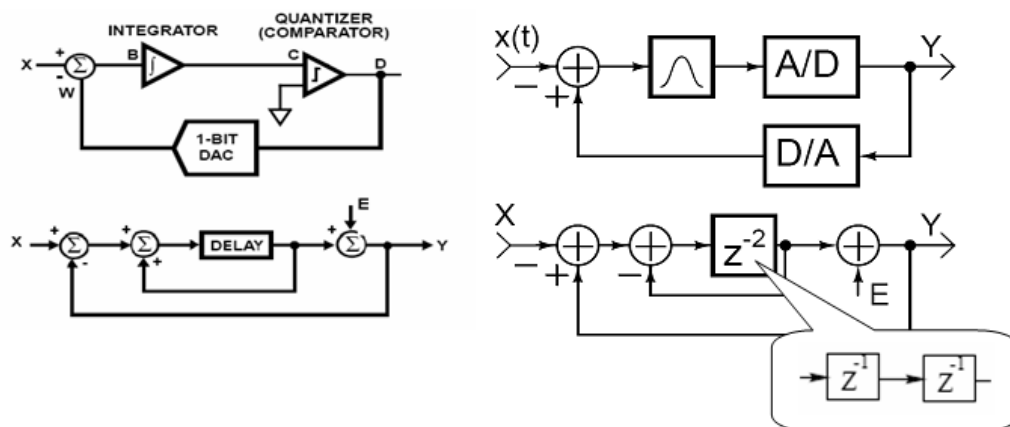
Abstrakt — Masívne využívanie digitálneho spracovania signálov vo všetkých oblastiach techniky, ale najmä v komunikačných systémoch vytvára vysoký dopyt po lacných programovateľných A/Č prevodníkoch pre spracovanie digitálne modulovaných úzkopásmových signálov. Implementácia prevodníkov na VLSI technológii umožňuje použitím najmenšieho možného počtu analógových funkčných blokov realizovať vhodný analógovo číslicový prevodník využívajúci vo svojej štruktúre princíp SigmaDelta modulácie. Článok sa zaoberá realizáciou $\Sigma\Delta$ modulátora ako základného stavebného prvku pásmového $\Sigma\Delta$ prevodníka na dostupnom programovateľnom obvode System on Chip (SoC). Použitý obvod SoC (CY8C27443-24PVI) bol vyvinutý firmou CYPRESS MICROSYSTEMS.

Kľúčové slová — analógovo-číslicový prevodník, Sigma-Delta prevodník, PSoC

I. INTRODUCTION

Analógovo číslicové prevodníky (AČP) predstavujú elektronické bloky, ktoré sú základom analógovo digitálnych rozhraní v súčasnosti mimoriadne široko využívaných v rôznych oblastiach techniky napr. v senzorovej technike [1] komunikačných systémoch [2], [3], počítačoch [4] a inde. Problematika AČP je veľmi rozsiahla a zložitá: existuje množstvo architektur A/Č prevodníkov s rôznymi vlastnosťami a parametrami [5] a mimoriadne zložitou otázkou je ich testovanie [6] - [11]. AČP v súčasnosti prenikajú stále viac aj do vysokofrekvenčných komunikačných systémov a zariadení, ako napr. softvérové rádio, kde veľmi významnú úlohu zohrávajú pásmové sigma -delta AČP [2], [3].

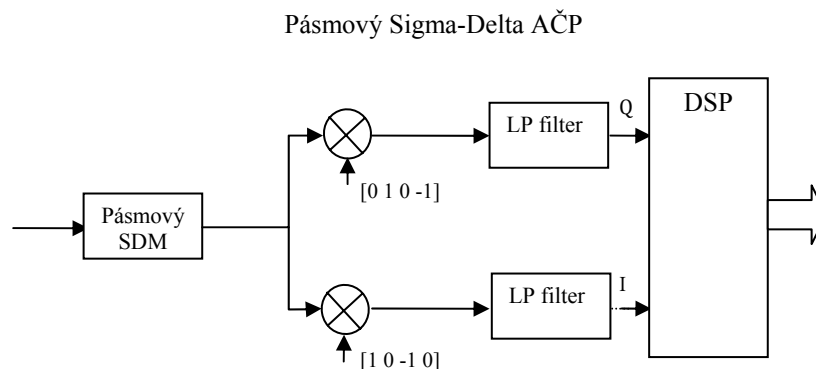
Základom štruktúry pásmového $\Sigma\Delta$ prevodníka (BP $\Sigma\Delta$) sa stala štruktúra LP $\Sigma\Delta$ modulátora (obr. 1). LP $\Sigma\Delta$ modulátore sa skladá z rozdielového zosilňovača s dolnopriepustným filtrom (integrátor), ktorý spracováva rozdiel medzi vstupným signálom a signálom zo spätnej väzby tvorenej komparátorom a jednobitovým číslicovým analógovým prevodníkom (ČAP). Vnútroštruktúru BP $\Sigma\Delta$ modulátora a jeho diskretný model je zobrazený na obr. 1. Náhrada dolnopriepustného filtra rezonátorom spôsobí, že dolnopriepustný modulátor sa zmení na pásmový.



Obr. 1 LP $\Sigma\Delta$ modulátor a jeho diskretný model + transformácia LP $\Sigma\Delta$ modulátora na pásmový

Rezonátor je realizovaný časovo diskretizovanou štruktúrou v okolí rezonančnej frekvencie, danou vzorkovacou frekvenciou f_s . To predstavuje transformáciu prenosovej funkcie LP modulátora v z -oblasti. Transformácia $z^{-1} \Rightarrow -z^{-2}$ je výhodná, pretože vlastnosti LP modelu, ako stabilita a dynamika, zostávajú zachované. Nevýhodou je, že prenosová funkcia šumu (NTF) dosahuje minimum na pevne stanovenej frekvencii $f_s/4$ vzhľadom k vzorkovacej frekvencii [2].

Binárnym násobením výstupu z modulátora binárnym priebehom s frekvenciou $4x f_{IN}$ dôjde k frekvenčnému posunutiu signálu zo vstupu modulátora do počiatku frekvenčnej osi. Fázovým posunutím signálu v číslicovom demodulátore sa dá získať fázovo posunutá zložka vstupného signálu. Po LP číslicovej filtrácii sa získa číslicový údaj úmerný reálnej a imaginárnej zložke vstupného signálu. Prevodník sa správa ako vektorový A/Č prevodník. Výsledná štruktúra $\Sigma\Delta$ A/D prevodníka je na obr. 2.



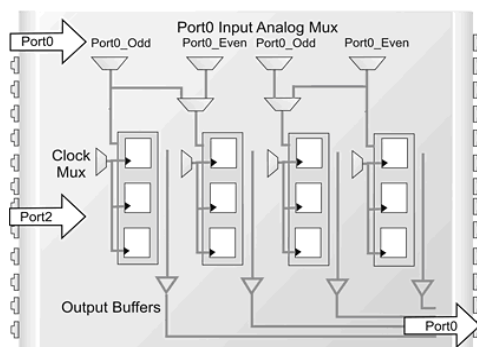
Obr. 2 Výsledný AČ prevodník na dekodovanie I/Q zložky harmonického signálu

Hlavnou výhodou $\Sigma\Delta$ prevodníkov z pohľadu konverzie je možnosť priamej konverzie vstupného analógového signálu na vektor s dvomi zložkami (reálna a imaginárna) a tiež možnosť tvarovania kvantizačného šumu. Digitálne $\Sigma\Delta$ A/D prevodníky môžu byť vstavané v rôznych digitálnych zariadeniach, keďže sú kompatibilné s technológiou VLSI a navyše sú zostavené z minimálneho počtu jednoduchých analógových prvkov. Z jednoduchej štruktúry vyplýva vysoká linearita prevodu.

Tieto vlastnosti prispeli k tomu, že sa stali veľmi obľúbenými v komunikačných systémoch využívajúcich technológie ako GSM, WCDMA, UMTS používaných pre mobilné telefóny, tiež pre FM/AM prijímače, pre bezdrôtovú digitálnu televíziu a mnohé ďalšie podobné aplikácie. Možnosť získania I a Q zložky signálu sa používa pri demoduláciách typu QAM, 16QAM, QPSK, ale dá tiež využiť v senzorovej technike [1]. Tu sa získa signál z dvoj parametrického senzora, kde jedna veličina ovplyvňuje reálnu zložku a druhá veličina imaginárnu zložku výstupnej impedancie pasívneho senzora, alebo môže predstavovať diagnostickú informáciu o funkčnosti senzora.

II. OBVOD PROGRAMABLE SYSTEM ON CHIP (PSOC)

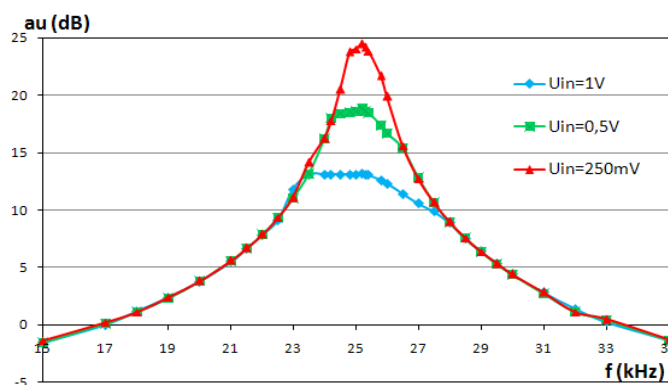
Obvody Programmable System On Chip (PSoC) firmy Cypress Microsystems [12] umožňujú integráciu analógových číslicových programovateľných blokov do jedného integrovaného obvodu. System On Chip (SoC) pozostáva z hardvérovej a softvérovej časti, ktorá riadi mikroprocesor a periférie. Použitý obvod CY8C27443-24PVI je vytvorený SONOS technológiou. Mikroprocesor nazývaný M8C je postavený na 8-bitovej harvardskej architektúre a umožňuje efektívne programovanie v programovacom jazyku nízkej úrovne. Digitálna časť pozostáva z programovateľných blokov usporiadaných v rovinách. Analógová časť reprezentujú tri typy analógových blokov: Continuous Time (CT) poskytujúce v čase spojité analógové funkcie a Switch Capacitor (SC) typu C a D, poskytujúce analógové funkcie s využitím techniky spínaných kapacitorov. Bloky sú voľne programovateľné čím poskytujú širokú paletu funkcií a sú usporiadané do stĺpcov pričom každý má k dispozícii niekoľko vstupov a výstupov, ako to možno vidieť na obr. 3, umožňujúcich prepojenie blokov navzájom, pripojenie pracovných hodín, referenčného napätia a tiež pripojenie ako na vstup tak aj na výstup obvodu. Analógové bloky sa používajú pre vývoj analógových elementov, ako sú analógové filtre, komparátory, zosilňovače a dobre známe ADC a DAC prevodníky. Pre konfiguráciu a emuláciu obvodu slúži integrované grafické vývojové prostredie PSoC Designer.



Obr. 3 Usporiadanie analógových blokov v obvode PSoC

III. NÁVRH BP $\Sigma\Delta$ MODULÁTORA NA OBVODE PSoC

Rezonátor $\Sigma\Delta$ modulátora je vytvorený doplnením dvojitého oneskorovacieho obvodu o zápornú spätnú väzbu a sčítavacieho obvodu v analógovej časti obvodu PSoC. Oneskorovací obvod tvorený zosilňovačom s jednotkovým zosilnením oneskorí vstupný signál o jeden hodinový takt. Vstupný analógový signál je privedený na A vstup prvého zosilňovača nastavený ako analógový vstup. Signál potom prechádza dvomi zosilňovačmi, ktoré vytvárajú potrebné dvojité oneskorenie. Prvý blok na vstupe štruktúry rezonátora bol doplnený len pre časovú diskretizáciu priebehu a nastavenie požadovaných vlastností rezonátora postaveného na obvode PSoC. Výstup z tretieho zosilňovača nastavený ako analógový výstup cez zásobník predstavuje výstup obvodu dvojitého oneskorenia. Tento výstup je pomocou externého prepojenia privedený na invertujúci B vstup prvého zosilňovača a predstavuje spätnú väzbu rezonátora. Prenosovú charakteristiku vytvoreného rezonátora možno vidieť na obr. 4.



Obr. 4 Prenosová charakteristika rezonátora

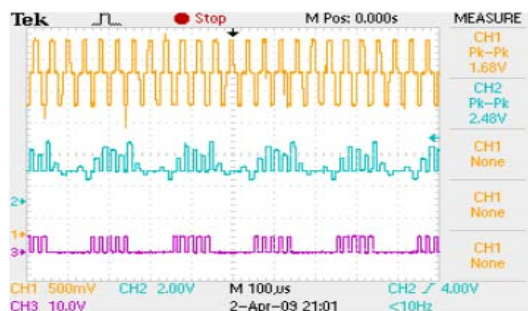
Komparátor je pripojený k rezonátoru, cez výstup druhého bloku zosilňovača. Výstup komparátora je digitálnou zbernicou pripojený na blok zásobníka nachádzajúceho sa digitálnej časti obvodu. Výstup predstavuje pin P_2_4, ktorý je nastavený ako GlobalOut (vnútorná zbernica pre pripájanie blokov k vonkajšiemu okoliu) a v móde Strong (tento mód sa používa pre digitálne vstupy/výstupy). Výstup komparátora je následne privedený na vstup modulátora, pričom obvod sumátora je realizovaný diskretnými prvkami (rezistory). Tento spôsob realizácie súčtového obvodu na vstupe bol zvolený pre nedostatok ako aj nevhodnosť použitia zvyšných blokov obvodu, ktoré majú obmedzené možnosti pre výber vhodnej pracovnej frekvencie ako aj dodatočné oneskorenia vznikajúce medzi A a B vstupom. Konfigurácia BP $\Sigma\Delta$ modulátora v analógovej časti obvodu PSoC bola vytvorená v grafickom vývojovom prostredí PSoC Designer.

Spracovanie získanej digitálnej informácie bolo realizované použitím meracej karty NI USB-6521 a aplikačného programu vytvoreného pre tento účel vo vývojovom prostredí LabView. Meracia karta spoločne s vytvoreným aplikačným programom umožňuje načítavanie digitálneho signálu a jeho násobenie binárnym sledom, následné spracovanie, vyhodnotenie a zobrazenie výsledkov v číselnej aj grafickej forme.

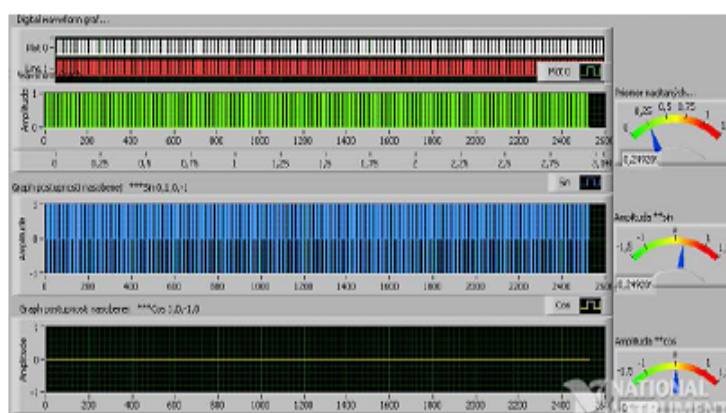
IV. EXPERIMENTÁLNE VÝSLEDKY

Meranie vytvorenej štruktúry BP $\Sigma\Delta$ modulátora bolo v oboch prípadoch vykonávané súčasne osciloskopom aj meracou kartou DAQ a programom pre registráciu binárných údajov

z modulátora vytvoreného v LabView. Nameraný priebeh z BP $\Sigma\Delta$ modulátora pri jednej hodnote vstupného napätia $U_{in}=0,5V$ je ukázaný na osciloskopickom zázname obr. 5 (oranžový priebeh je výstupom z rezonátora, fialový priebeh je výstup komparátora pripojeného na rezonátor a modrým výsledný priebeh modulátora) kým výsledný binárny sled získaný DAQ kartou a aplikačným programom je na obr. 6.

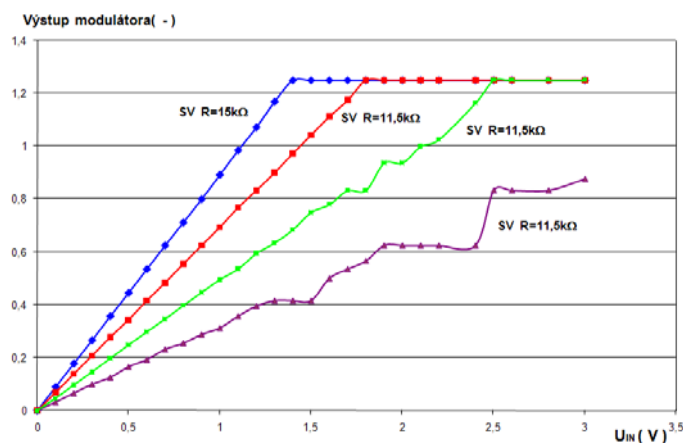


Obr. 5 Priebeh z osciloskopu na výstupe BP $\Delta\Sigma$ modulátora pre amplitúdu vstupného signálu $U_{in}=0,5V$



Obr. 6 Priebeh a číselné vyhodnotenie z programu pre registráciu binárnych údajov na výstupe BP $\Sigma\Delta$ modulátora pre amplitúdu vstupného signálu $U_{in} = 0,5V$

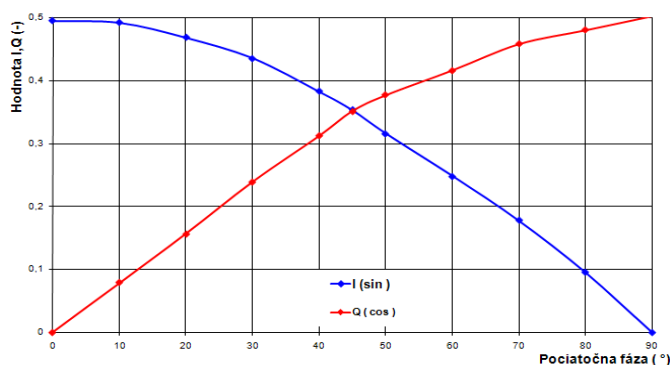
Po binárnom násobení, ktorému odpovedá posun signálu do nuly frekvenčnej osi, je výsledný signál filtrovaný dolnofrekvenčným filtrom typu FIR. Rozsah, kedy sa hodnota na výstupe modulátora mení priamo úmerne amplitúde vstupného signálu, je obmedzený. Strmosť prevodovej charakteristiky a teda aj rozsah amplitúd vstupného signálu je možné ovplyvniť zmenou odporu v spätnej väzbe komparátora. Prevodovú charakteristiku ako aj chovanie sa navrhnutého prevodníka zmenou spätnoväzobného odporu môžeme vidieť na obr. 7.



Obr. 7 Prevodová charakteristika finálneho BP SD modulátora

Jednou z významných vlastností tohto prevodníka je citlivosť na fázu vstupného signálu, čo sa prejaví zmenou hodnoty I a Q zložky získanej spracovaním binárnej informácie

z modulátora. Na obr. 8 môžeme vidieť nameranú závislosť I a Q zložky v závislosti od počiatočnej fázy prevádzaného analógového signálu.



Obr. 8 Závislosť I a Q zložky od počiatočnej fázy prevádzaného analógového signálu

V. ZÁVER

Sigma-Delta ($\Sigma\Delta$) modulácia poskytuje veľa výhod medzi ktoré patrí hlavne jednoduchá realizovateľnosť a tiež fakt, že je vhodná pre programovateľné komunikačné rozhrania. Použitý programovateľný obvod PSoC patriaci do kategórie System on Chip obvodov umožňuje jednoduchý návrh úplného systému na spracovanie zmiešaných signálov za pomoci jedného integrovaného obvodu. Hlavným nedostatkom tohto riešenia je malý rozsah linearity prevodovej charakteristiky. Tento je spôsobený nutným posunutím nuly prevodníka za pomoci jednosmerného predpätia na úroveň 2,5V. Okrem obmedzenia rozkmitu spracovávaného signálu problémom bolo aj tepelný drift blokov obvodu. Výhodou implementácie na SoC obvode je aj možnosť číslicovej filtrácie z výstupu modulátora na tom istom čipe.

POĎAKOVANIE

Táto práca bola podporovaná realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (70%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (30%).

LITERATÚRA

- [1] HAZE, J. - VRBA, R. – FUJCIK, L. – FOREJTEK, J. – ZAVORAL, P. – PAVLIK, M. – MICHAELI, L.: The Novel Band-Pass Sigma-Delta Modulator Utilised for Capacitive Pressure Sensing, ICONS '7 Second International Conference, 22 – 28 april 2007.
- [2] SALO, T.: A 80-MHz bandpass $\Delta\Sigma$ modulator for a 100-MHz IF receiver, Solid-State Circuits, IEEE Journal of, vol. 37, pp. 798-808, 2002.
- [3] H. K. C. ISSAC, "A 70 MHz CMOS band-pass sigma-delta analog-to-digital converter for wireless receivers", M.S. thesis, Hong Kong University of Science and Technology, August 1999.
- [4] ŠALIGA, Ján - MICHAELI, Linus: Software for metrological characterisation of PC sound cards. In: Computer standards and interfaces. vol. 25, no. 1 (2003), p. 45-55.
- [5] MICHAELI, Linus - ŠALIGA, Ján - KOLLÁR, Martin: Parameters of band pass sigma delta ADC and the comparison with the standard ones. In: Measurement. vol. 40, no. 5 (2007), p. 473-478.
- [6] ŠALIGA, Ján: Testovanie parametrov analógovo-číslicových prevodníkov dynamickými metódami. In: Slaboproudý obzor. roč. 59, č. 1 (2002), s. 6-11.
- [7] CHOVANEC, Marián - MICHAELI, Linus - ŠALIGA, Ján: Noise shaping structures of band pass sigma delta ADC and their impact on metrological parameters. In: 16th IMEKO TC4 International Symposium, Florence, Italy, 2008, p. 977-982.
- [8] MICHAELI, Linus - ŠALIGA, Ján - MICHALKO, Peter: Triangular testing signal for identification of unified error model parameters. In: Measurement. vol. 40, no. 5 (2007), p. 491-499.
- [9] MICHAELI, Linus - MICHALKO, Peter - ŠALIGA, Ján: A new ADC fast testing method based on the unified error model. In: Measurement. vol. 41, no. 2 (2008), p. 192-197.
- [10] SERRA, António Cruz et al.: Combined spectral and histogram analysis for fast ADC testing. In: IEEE transactions on instrumentation and measurement. vol. 54, no. 4 (2005), p. 1617-1623.
- [11] CORRADO, Marco et al.: A critical analysis of alternative stimulus signals for histogram based testing of ADCs. In: 2008 IEEE International Instrumentation and Measurement Technology Conference Proceedings: I2MT,C 2008, Victoria, Canada, p. 320-352.
- [12] <http://www.cypress.com>

New Approach for Measurement of Static Characteristics D/A Converter

Martin SEKERÁK, Linus MICHAELI, Ján ŠALIGA, Marek GODLA

Dept. of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

*martin.sekerak@tuke.sk, linus.michaeli@tuke.sk, jan.saliga@tuke.sk,
marek.godla@tuke.sk*

Abstract — This paper describes new approach for measurement of the static characteristic of Digital-to-Analog Converters (DACs) under dynamic conditions. The main principle of proposed method is conversion of the DAC analog output voltage to time interval corresponding the equality of DAC output voltage to the reference by using a fast comparator. The main advantage of this approach is that time can be measured in the electronic domain much more accurately than amplitude. Time interval is measured by the registration of the input DAC code. From the obtained information a Personal Computer (PC) evaluates functional parameters Integral-nonlinearity (INL) and Differential-nonlinearity (DNL).

Keywords — analog-to-digital converter, comparator, differential and integral-nonlinearity

I. INTRODUCTION

Nowadays, electronic devices based on digital signal processing integrate more and more features, each of them presenting higher requirements for the further increase of the volume of processed data, speed but also their accuracy and quality. Despite a significant increase in performance and accuracy of today's digital electronic devices, most signals in the natural environment are analog, and therefore there is a need for equipment to enable the transfer of signal from digital to analog form. Digital to analog conversion requires the use of highly accurate, fast and linear DAC converters. The implementation of high quality DAC converters can be achieved by using suitable DAC architectures with their associated properties and technology used in production.

An important question arises for testing and measuring static and dynamic characteristics of such high quality converters when it is necessary, because their accuracy and speed has been higher than ADC converters. In literature, many suggestions have been proposed for measuring DACs characteristics ([1], [2] [3], - [9]), but most of them have been analyzed only from a theoretical point of view in simulations and their experimental implementation is often difficult. This paper proposes a method based on the conversion of the output voltage DAC tested level into a time interval. The time interval is measurable with higher precision than dynamic output voltage.

II. THE PROPOSED METHOD TO MEASURE DAC CONVERTERS

The principle of the proposed methods is based on conversion of DAC analog output voltage corresponding tested quantization level to time using a fast comparator. The time instant when the DAC output signal achieves the chosen quantization level on its output carries information about nonlinearity of transfer characteristics of tested DAC. Time interval will be estimated with a high precision by using fast voltage comparators, precision DC voltage sources and further auxiliary circuits. Moreover, the output of the comparator is not directly represented as a time, which will be subsequently measured, but the output of the comparator will be used for setting the conditions for the entry digital code value of the input buffer memory. Adding dithering voltage increases the resolving power of the method distinguishes error below 1LSB. It is necessary to synchronize the entire measurement system, and also to remove possible error sources in the testing stand and look for ways how to improve precision.

III. METHOD OF GENERATION AND MEASUREMENT OF PULSES

Digital code word k on input tested DAC introduces one digital input code of the digital sawtooth generated by the counter, which counts clock impulses. The output of the counter is connected to the input of the test converter and the buffering memory. The analog output voltage U_{DAC} from tested DAC is then connected to the input of the fast comparator, which compares it with sum of reference DC voltage U_a and auxiliary dithering voltage U_{DITH} . The reference voltage shall be adjusted as the decision-making level for the test quantization level, and can be generated precisely calibrate source or simply by using DC source and measure analog output value by using an accurate digital voltmeter. The required amplitude of dithering voltage is achieved by using an appropriate resistive divider connected to the output of the generator of dithering voltage. Triangular dithering voltage is used as dithering voltage with the peak value of the amplitude $U_{DITH\ MAX} = 1\text{LSB}$ and the period for which applies that ratio of the period of dithering voltage to the period of saw is an integer. The block diagram of the proposed method, which was described above, can be seen in Fig. 1.

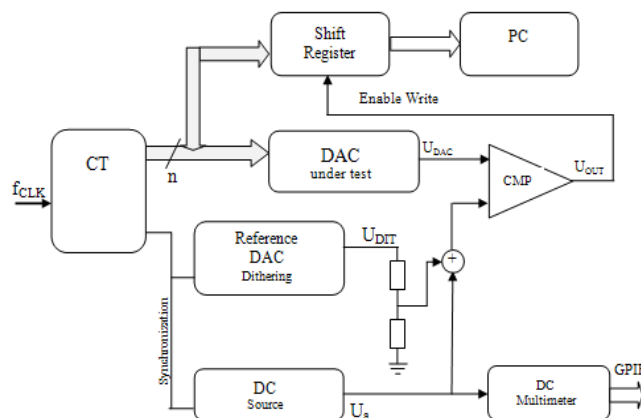


Fig. 1 Block diagram of the method of generation and measurement pulse

The output of the comparator is not directly represented as a time which will be subsequently measured, but the output of the comparator will be used for setting the conditions for entry code value of the input buffer memory. Buffering memory is used for storing values of code words when the condition (1) is met that the test voltage output DAC converter is greater than the sum of the DC voltage and dithering voltage:

$$U_{CAP} \geq (U_{ak} + U_{DITH}) \Rightarrow k \quad (1)$$

where the value k is the value of digital input code of the test DAC, when the analog output voltage U_{DAC} reaches the set reference level U_{ak} . The recorded sequence of value k for individual quantization level is then loaded from buffer memory into the PC, where will be processed. The main mentioned signals of proposed method U_{DAC} , U_{ak} , U_{DITH} along with obtained sequence of k value and a detail view of DAC real transfer characteristics along with mentioned signal for the third quantization level are for better understanding shown in following figure (Fig. 2)

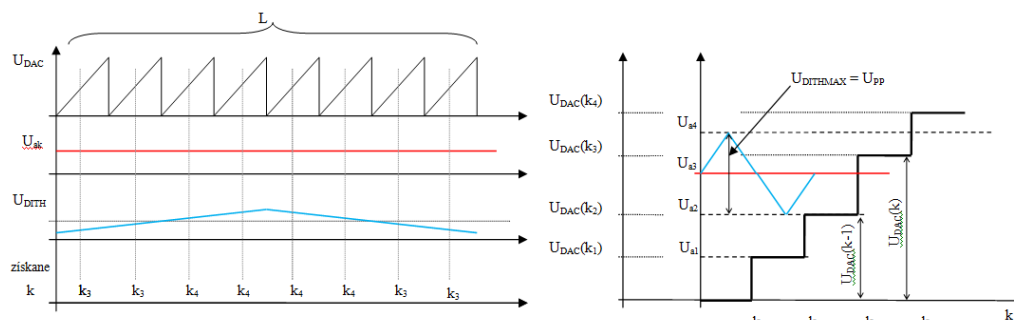


Fig. 2 Waveform of particular signals U_{DAC} , U_{ak} , U_{DITH} along with sequence of the obtained value k and detail view signals of method of generation and measurements of pulses for DAC real transfer characteristics for the third quantization level

The increase of the measurement accuracy will be achieved by the repeated measurements of one value and thus a sequence of k value for a certain quantization level will be obtained, which will then be averaged.

$$T_k = \frac{1}{L} \sum_{i=1}^L k_i \quad (2)$$

The average value of registered digital codes will be equal to a digital linear interpolated value T_k between discrete value k and $(k-1)$. Further processing results in obtaining the value of variations output voltage of the corresponding tested quantization level from the value of the ideal transfer characteristics. Value T_k multiplied by ideal quantization step Q can be interpreted as a linearly interpolated value of the analog output voltage of DAC under test for the tested quantization level, which can be expressed by the following statement:

$$U_{MER}(T_k) = T_k * Q \approx U_{IDEAL}(k) = k * Q \quad (3)$$

The difference between the linearly interpolated value of the analogue output voltage $U_{MER}(T_k)$ and exact value of DC voltage U_{ak} represent integral nonlinearity according to:

$$INL(T_k) = \frac{U_{IDEAL}(T_k) - U_{ak}}{Q} \quad (4)$$

Assuming that the integral non linearity in the quantization step changing at least, we can write the following statement:

$$INL(T_k) = INL(k) \quad (5)$$

Proposed methods have been simulated in LabView environment and have not been yet fully implemented experimentally.

IV. CONCLUSION

This paper describes a new method of measuring functional characteristics of D/A converter, based on the conversion of voltage to time, which is precisely than measurement of amplitude.

The advantage of the proposed method in addition to the converting amplitude of analog voltage to time is a relative simplicity and low demands for the used laboratory equipment. Using the speed memory, whose contents is loaded into a PC where it will run processing of data and in that case we do not need any fast counters or quick scopes. The addition of dithering voltage allows more accurate measurement of small errors. Dithering voltage added to DC voltage suppresses the systematic error in the digital interval measurement caused by the synchronism of stating instant. On the other hand, the weak point of the method is a fast comparator, which causes a measurement error due to its own delays. Assuming that this delay is still the same, it is possible to correct this error by appropriate techniques.

ACKNOWLEDGMENT

This work is the result of the project implementation Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120020) supported by the Research & Development Operational Program funded by the ERDF (50%).

This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

REFERENCES

- [1] D. L. Cari, D. Grimaldi, "Static characterization of high resolution DAC based on over sampling and low resolution ADC", Proc. Of IEEE Instrum. And Measur. Techn. Conf., IMTC 2007, May 1-3, 2007, Warshaw Poland.
- [2] C. W. Lin, S. F. Lin and S. F. Luo, "A new approach for nonlinearity test of high speed DAC," in Proc. IEEE Int. Workshop Mixed-Signals, Sens., Syst., 2008, pp. 1-5.
- [3] B. Vargha, J. Schoukens and Y. Rolain, "Static nonlinearity testing of digital-to-analog converters," IEEE Trans. Instrum. Meas., vol. 50, no. 5, pp. 1283-1288, Oct. 2001.
- [4] A. Baccigalupi, Mauro D'Arco, A. Liccardo, M. Vardusi, "Implementation of high resolution DAC test station: A contribution to draft standard IEEE P1658", XIX IMEKO World Congress Fundamental and Applied Metrology, September 6-11, 2009, Lisbon, Portugal.
- [5] J. Savoj, Ali-Azam Abbasfar, A. Amirkhany, B. W. Garlepp, M. A. Horowitz, "A new technique for characterization of Digital-to-Analog Converters in high-speed systems" Design, Automation & Test in Europe Conf.& Exhibition, 16-20 April 2007, pp.1-6.
- [6] J. Le, H. Hosam, R. Geiger, C. Degang, "Testing of precision DACs using low-resolution ADCs with dithering", Proc. Of IEEE Intern. Test Conf., Oct. 2006, pp.1-10.
- [7] L. Jin, H. Haggag, R. Geiger and D. Chen, "Testing of precision DACs using low-resolution ADCs with wobbling," IEEE Trans. Instrum. Meas., vol. 57, no. 5, pp. 940-946, May 2008.
- [8] A. Baccigalupi, M. D'Arco, A. Liccardo, M. Vardusi, "Test equipment for DAC's Performance Assessment: Design and Characterization" IEEE Trans. Instrum. Meas., vol. 59, no. 5, May 2010.

Modeling of Some Analog-to-Digital Converter Properties in LabVIEW

*Ján ŠALIGA, Martin SEKERÁK, Linus MICHAELI,
Marek GODLA, Michal ROSINA*

Dept. of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

{jan.saliga, martin.sekerak, linus.michaeli, marek.godla}@tuke.sk,
michal.rosina@gmail.com

Abstrakt — This paper presents modelling of the non-ideal analogue-to-digital converter with hysteresis in its transfer characteristics. The phenomenon of hysteresis is not very well explored in practice as well as its effect on test methods of ADC. The proposed model of nonlinear ADC with hysteresis enables simple and fast evaluation of newly developed non standardized ADC test method by simulations. The model was developed in LabVIEW and realized as a library with a variety of signal processing and ADC modelling functions.

Keywords — ADC tests, hysteresis, model of ADC, LabVIEW

I. INTRODUCTION

Nowadays, the advantage of digital signal processing along with the rapid development of semiconductors caused, that almost everything is processed digitally. There is a tendency to get a continuous analogue signal from the physical environment to the digital form in the shortest possible way. The devices that perform conversion of the analogue signal to the sequence of digital codes are the analogue-to-digital converters (ADC). There are a lot of different structures of ADCs with different quality and properties. The error analysis of ADCs by means of models is becoming more and more popular.

The objective of this work is to create a model of ADC with non-ideal properties (nonlinearity) including hysteresis. Functions of ADC individual characteristics such as nonlinearity and hysteresis errors were developed so that they are similar to a real ADC. The model of converter changes the input waveform to the output code sequence. In conclusion, the developed model was tested. The created model will enable to perform different tests from the area of ADC measurement. Comparing the test results with the known simulated properties it will be able to determine the efficiency of examined method.

II. ANALOG-TO-DIGITAL CONVERTERS AND THEIR PROPERTIES

ADC is a device which converts the natural continuing input waveform from the physical environment to the discrete output values in accordance with the corresponding transfer characteristic. After the conversion all input analog values are represented by the definite number of digital codes while each code represents a part from the input range [1].

The transfer characteristic maps ADC input values from the certain range of V_{FS} to the definite number K of code words k which create the definite set of all possible output code words. The transfer characteristic is definite by plotting where the x axis is divided into a set of contiguous not overlapping intervals and each of them has a certain co-ordinate output code on the y axis [1].

Code bin (k) is a digital output that corresponds to a particular set of input values [2]. Code transition level $T[k]$ is the value of the converter -input parameter at the transition point between two given adjacent code bins. The transition point is defined as the input value that causes 50% of the output codes to be greater than or equal to the upper code of the transition, and 50% to be less than the upper code of the transition. It is level of output code, where the probability of occurrence of neighboring code words on ADC output is the same [1].

The difference of the code transition levels, $T[k+1]$ and $T[k]$, that delimit the k -th bin, that

we can write following formula (1) [2] is code bin width ($W[k]$),

$$W[k] = T[k + 1] - T[k] \quad (1)$$

Gain and offset are the values by which the input values are multiplied and then to which the input values are added, respectively, to cause the deviations from the output values to be zero at the terminal points, that is, at the first and last codes [2]. It is the deviation of actual width of code bin with $W[k]$, from its nominal width, usually refers to its nominal value and is usually expressed in LSB. (2).

$$DNL[k] = \frac{W[k] - W_{nom}}{W_{nom}} \quad (2)$$

Before determining of $DNL[k]$ it is important to make correction of gain error. Sometimes is instead of all $DNL[k]$ function expressed only absolute value of DNL [2]. It is the deviation of actual transition level $T[k]$ from its nominal value, usually refers to nominal quantization step and is expressed in LSB units (3).

$$INL[k] = \frac{T[k] - T_{nom}[k]}{W_{nom}} \quad (3)$$

Before determining of $INL[k]$ it is important to make correction of gain and offset error. In that case are the end points of $INL[k]$ function equal to zero and the non-linearity of ADC makes felt by nonzero values of internal points. There is relation between $INL[k]$ and $DNL[k]$ parameters, which can be expressed by formula (4).

$$INL[k] = INL[k - 1] + DNL[k - 1] = \sum_{i=1}^{k-1} DNL[i] \quad (4)$$

where $k = 2, 3, \dots, 2^N - 1$, whereby is assumed a correction $INL[1]=0$ [2].

III. OBJECTIVE ANALYSIS

ADC hysteresis is described as an effect of the ADC transfer curve that it depends on the direction by which the transfer curve is traversed (i.e., increasing or decreasing signal). The reported hysteresis of the ADC, if any, is the maximum of such differences. In other words, it is a deviation between the transition levels $T_{up}[k]$ measured by the rising waveform and $T_{down}[k]$ measured by the falling waveform on the ADC input. Only a maximum is usually expressed on the given ADC.

In practice, a smaller emphasis is put on this phenomenon compared to other errors of ADC. We decided to verify the sensitivity of actual test methods used for ADCs on hysteresis effect using the model of ADC which allows us to simulate hysteresis along with other error properties. Here we are able to perform standard test methods on the output codes of the model and compare the properties of ADC obtained from the tests with the known properties set up in the model.

Simulation of ADC was created in the development environment LabVIEW and consists of two main parts. The first one serves to create the desired transfer characteristic, the second one uses the created transfer characteristic and converts a simulated waveform on a sequence of codes. These two parts are not connected directly. The transfer characteristic modelled by the first part of the model can be saved to the memory. The saved transfer characteristic can be loaded from the memory in the second part of the model and used for conversion of a waveform in this part. On the output of ADC model, we added the functions designed to histogram testing to evaluate the developed model. The all parts of the model will be described in the following chapter.

IV. DESIGN OF ADC MODEL

As mentioned above, created model of ADC consist of two parts, the model of the ADC transfer characteristic and the model of the converter in action.

Designed model of transfer characteristic has a few options to set the ADC resolution, the magnitude and shape of nonlinearity and the magnitude and shape of hysteresis. The transfer characteristic is simulated by means of its INL because such an approach allows us to see simulated nonlinearity and hysteresis on the transfer characteristic directly. The disadvantage of this approach lies in modelling of gain and offset errors because these are not manifested in

INL. In this case we had to add these errors to the simulation later after the basic transfer characteristic was modelled by adding the real value of transition levels. Simulator of transfer characteristic has option to set two different shapes of nonlinearity, the first one is one period of sine, and the second one half period of sine. In simulation of hysteresis a possibility of the adjustment of other magnitudes of nonlinear transfer characteristic for a rising and a falling waveform was involved.

Both the modelled hysteresis and the modelled nonlinearity directly present the ADC transfer characteristic and after their addition a final transfer characteristic originates. We would like to emphasize that we will finally create two different transfer characteristic that is caused by hysteresis. The first one is given for rising waveform conversion and the second one is given for falling waveform conversion.

The simulated transfer characteristic may be written to the memory, which allows us to simulate and save more different transfer characteristic under optional names to the memory. In the second part of the model we programmed a function to upload the transfer characteristic from the memory by means of a dialog box. After the transfer characteristic is read from the memory, its INL expression is changed to quantization levels $T[k]$ and gain and offset errors are added to the quantization levels.

The second part of the model allows conversion of the transfer characteristic from INL to $T[k]$, a simulation of some basic types of a waveform and modelling of the analogue-to-digital converter. The model of ADC converts simulated waveform by means of the created transfer characteristic to the sequence of codes. There was created waveform simulator which simulates a sine wave, a triangle, a sawtooth and a square signal with optional amplitude, offset, frequency and a phase and besides a sampling frequency and a number of samples per a record. Moreover, there is possibility to add Gaussian noise to the simulated signal in the waveform simulator. The output of the waveform simulator as well as the transfer characteristic for a rising and falling signal is connected to the inputs of ADC simulator. The simulator decides about each input sample of the waveform if that sample will be converted by the transfer characteristic for a rising or falling signal. The decision is performed by means of internal logic which compares each particular input sample with the last previous input sample. According to the fact if the particular sample is greater, smaller or equal to the previous sample, the logic chooses appropriate transfer characteristic. The output code of a converted sample is from the interval $0-(2^N-1)$. The code represents the serial number of a transition level from the transfer characteristic at which the value of a converted sample occurs. After the conversion, the sequence of codes is displayed on the graph. Data flow in the final model is shown on following block diagram (Fig. 1).

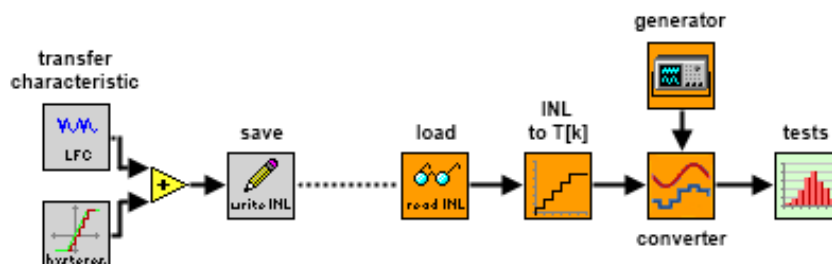


Fig. 1 The data flow in the final model; grey icons are the first part of the model, where the nonlinearity (LFC) and hysteresis create the transfer characteristic which could be saved to memory by means of "write INL"; orange icons are the second part of the model, where "read INL" uploads the transfer characteristic from memory, then the transfer characteristic is changed from INL expression to quantization levels $T[k]$. The $T[k]$ enters to "converter" as well as a waveform which is simulated by means of "generator". The waveform is converted according to the $T[k]$ into a code sequence inside the "converter"; green icon manifests the tests executed on output code sequence.

V. MODEL EVALUATION

The created model of ADC has been evaluated by means of some standard tests. For the reason, were used already prepared standardised test functions of a sine wave and ramp histogram test methods. The principle of the histogram test method is in building of the code histogram from the output codes of ADC that is then compared with an ideal one ([2]). With a quite long record it is assumed that a number of occurrences of individual code words in a record is directly proportional to corresponding code bin width. It is important to set up the actuating signal in a way that its frequency and the sampling frequency will correspond to the coherency condition (5)

$$M = J \cdot f_s / f_i \quad (5)$$

where M is a number of samples per a record, J is an integral number of periods of a waveform, f_s is a sampling frequency and f_i is the frequency of the actuating signal. It is essential that the amplitude of the actuating signal is equal to the full scale range of ADC.

We simulated 10-bits ADC with nonlinearity of the transfer function in the form of one period of sine and with the hysteresis modelled by means of a window function in tests. The maximum of both errors is 1 LSB. At first we simulated a ramp actuating signal and then a sine-wave actuating signal on the model input of the converter, both of them with frequency 3Hz, sampling frequency 100 kHz and with 100.000 samples per a record. We tested the output codes by means of an appropriate test method (sine and ramp histogram tests). INL of originally simulated transfer characteristic is shown on following Fig. 2.

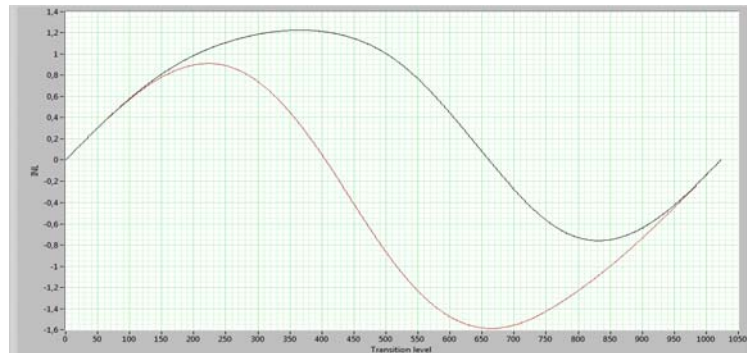


Fig. 2 Simulated INL of transfer characteristic of 10-bits ADC with nonlinearity and hysteresis. The black curve is the transfer characteristic for conversion of a rising waveform; the red curve is the transfer characteristic for conversion of a falling waveform.

The transfer characteristic shown on the picture above was uploaded into the model of the converter and the ramp waveform was simulated on the input of the converter. The output code sequence was tested by means of a ramp histogram test method. The INL measured from the output codes of ADC is shown on Fig. 3. Finally was compared the measured INL with both simulated INLs in a way that the measured INL was subtracted from the first one modelled INL and also subtracted the measured INL from the second modelled INL. The results are shown pictures Fig. 4 and Fig. 5.

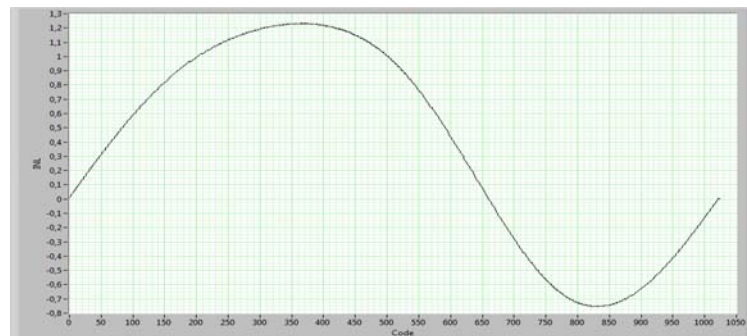


Fig. 3 INL of ADC measured by a ramp histogram method

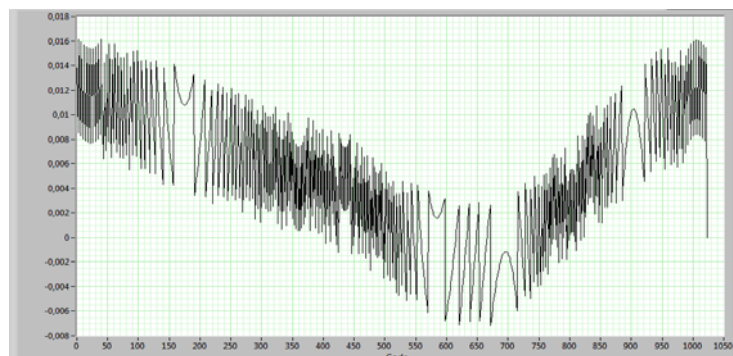


Fig. 4 The difference between INL of ADC measured by a ramp histogram method and the simulated INL for a rising signal

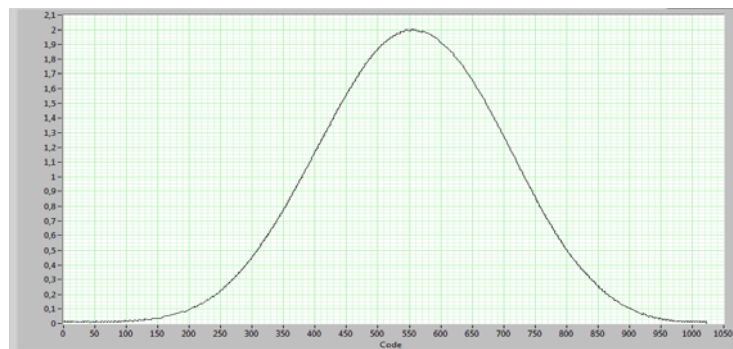


Fig. 5 The difference between INL of ADC measured by a ramp histogram method and the simulated INL for a falling signal

Similar measurement was performed by means of a sine waveform on the input of ADC. There was used the same transfer characteristic like in the previous measurement (Fig. 2). This time was tested the output codes by a sine-wave histogram test method and result was compared again. The measured INL by this method is on Fig. 6 and the comparison of the measured INL with both modelled INLs is on Fig. 7, and Fig. 8.

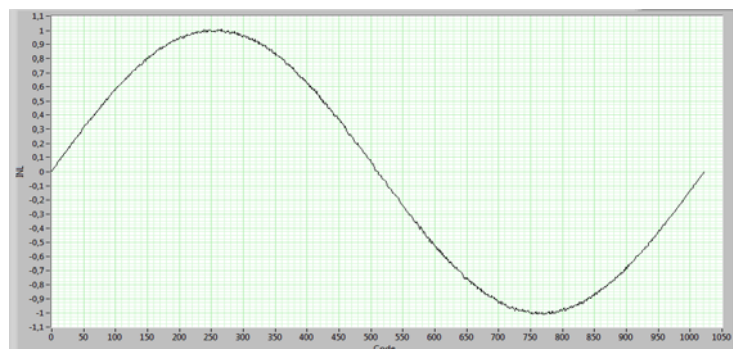


Fig. 6 INL of ADC measured by a sine-wave histogram method

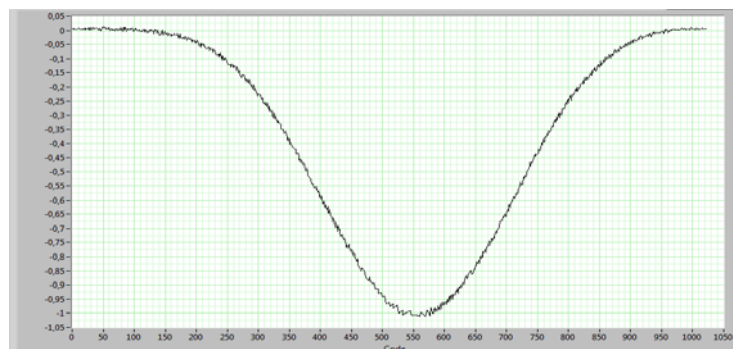


Fig. 7 The difference between INL of ADC measured by a sine-wave histogram method and the simulated INL for a rising signal

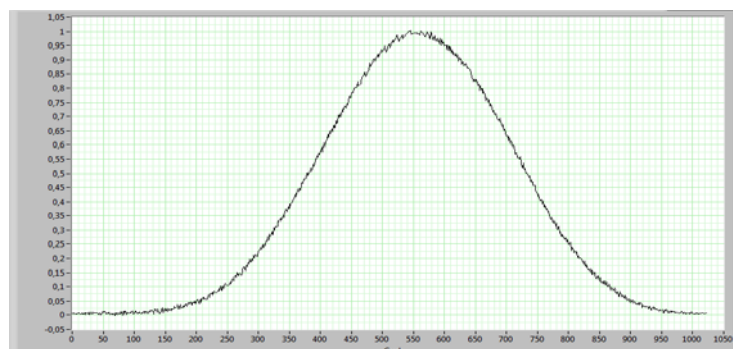


Fig. 8 The difference between INL of ADC measured by a sine-wave histogram method and the simulated INL for a falling signal

The advantage of the ramp histogram is obvious. The difference between INL measured by the ramp histogram and INL simulated for the conversion of a rising signal is relatively small, in a range from -0,006 LSB to 0,016 LSB. We suppose that it is possible to measure INL for the conversion of a falling signal when we use a ramp histogram method and a code sequence created from the decreasing ramp waveform. A sine-wave histogram method proved to be improper to find hysteresis on ADC. The measured INL by a sine-wave histogram method occurs among the modelled INLs for a rising and falling signal. In conclusion, when we tried to discover hysteresis of ADC, the ramp histogram test is more suitable than the sine-wave histogram test.

VI. CONCLUSION

A model of non-ideal ADC with hysteresis was developed. It allows finding out the efficiency of various ADC test methods on condition that hysteresis is on ADC. This model was tested by using histogram test. The ramp histogram method was successful and revealed hysteresis on ADC while the sine-wave histogram method failed.

The created model of ADC can be set up according to the needs of a user thanks to the fact that the individual functions of the model are inserted into the independent function.

ACKNOWLEDGMENT

This work is the result of the project implementation Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120020) supported by the Research & Development Operational Program funded by the ERDF (50%).

This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

REFERENCES

- [1] ŠALIGA J., "Metódy testovania analógovo-číslicových prevodníkov", Habilitačná práca, Košice 2001, KEMT, FEI TU.
- [2] IEEE Std. 1241 – 1998, "IEEE Standard for Terminology and Test Methods for Analogue-to-Digital Converters", Institute of Electrical and Electronics Engineers, Inc. New York, USA 1998.

Experimentálny systém FSO pre 625nm RONJA

Pavol MIŠENČÍK, Ján TURÁN, Ľuboš OVSENÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

pavol.misencik@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — Táto publikácia sa podrobne zaoberá experimentálnym systémom FSO pre 625 nm RONJA, ktorý aplikujeme do praxe. V článku je uvedený popis tohto systému, jeho vlastnosti a pripojenie k sieti.

Kľúčové slová — FSO, free space optics, RONJA, Reasonable Optical Near Joint Access

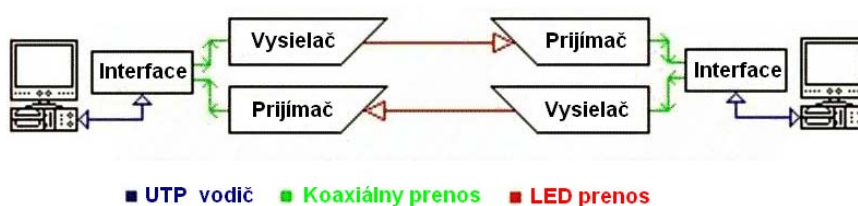
I. ÚVOD

V dnešnej modernej dobe veľmi rýchlo vzrastá ako dopyt po informáciách, tak aj rýchlosť, s ktorou sa informácie môžu dostať od poskytovateľa ku užívateľovi. Nadviazať spojenie nikdy nebolo také ľahké ako v dnešnej dobe. Veľmi obľúbenými sa stávajú optické komunikačné systémy voľným prostredím (FSO – Free Space Optics). Riešenia, tzv. poslednej míle je možné vyriešiť v rámci niekoľkých hodín a to bez akejkoľvek nutnej administratívy, ukladania kabeľáže do zeme alebo zlepšenia infraštruktúry. Lasery v súčasnej dobe sú natoľko výkonné, že umožňujú vysielat' signál spoza skla okna, čím sa zväčšujú možnosti umiestnenia a zabezpečenia FSO systémov. V tomto článku sa budeme zaoberať experimentálnym systémom RONJA implementovaným na KEMT FEI TUKE, jeho technickými parametrami a možným využitím [3].

II. POPIS SYSTÉMU A JEHO ZÁKLADNÉ PARAMETRE

RONJA je názov optického zariadenia, ktoré umožňuje optické bezvláknové pripojenie dvoch bodov počítačovej siete až na vzdialenosť 1300 metrov pri zachovaní konštantnej prenosovej rýchlosti 10 Mb/s. Pripája sa do ethernetovej karty alebo switch-u pomocou konektora RJ-45. Technológia je voľne dostupná bez nutnosti licencie, a bola vyvinutá na MFF Karlovej univerzity v Prahe. Názov je skratka z anglických slov „Reasonable Optical Near Joint Access“, čo je vo voľnom preklade „dostupné optické prepojenie na krátku vzdialenosť“.

Komunikačný prenosový systém RONJA sa skladá z dvoch úplne rovnakých zariadení, pričom každé z nich pozostáva z rozhrania (interface), vysielачa a prijímača. Ako je vidieť z obr. 1, interface je umiestnený neďaleko počítača a jeho úlohou je transformovať napät'ové úrovne a generovať signál. Vysielač a prijímač sú umiestnené v ochranných rúrach (obr. 2), ktoré sa pripevňujú na streche budov a ich úlohou je prenášať lúč svetla atmosferickým prenosovým prostredím (APP). Čo sa týka prepojenia jednotlivých častí systému RONJA, tak medzi počítačom a interfacem sa využíva na komunikáciu UTP vodič a k vysielачu a prijímaču vedie koaxiálny vodič [1,2,3,4].



Obr. 1 Bloková schéma RONJA [2]

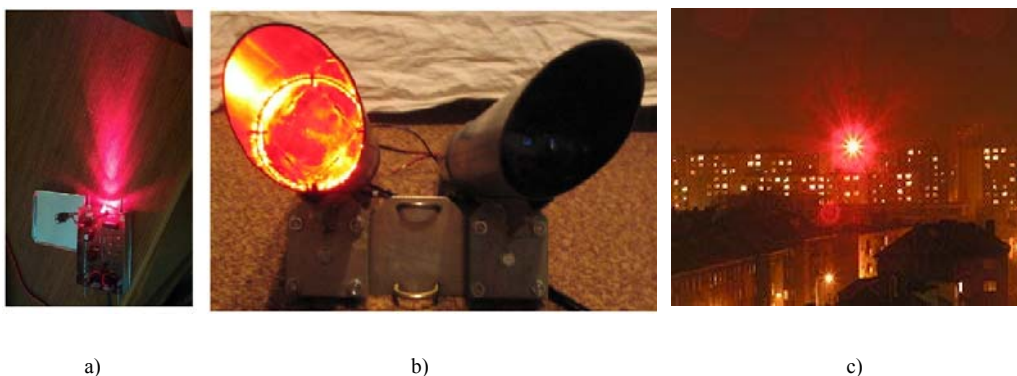
A. Základné časti systému RONJA

Interface:

Úlohou tohto interface je upraviť signálovú úroveň a impedancie pre optický prenos. Je to koncové zariadenie pre pripojenie k PC pomocou konektora RJ45. Taktiež generuje prídavný signál potrebný pre bezchybovú funkciu zariadenia s okolitými rušivými vplyvmi. Na zariadení je nainštalovaná sústava LED diód, ktoré indikujú funkčnosť dátového spojenia [3].

Vysielač:

Modul optického vysielača je zložený zo stabilizátora napätia z 12 V na 5 V a to kvôli použitiu logických obvodov. Ďalšiu časť tvorí vstupný zosilňovač, ktorý impedančne prispôbuje signál na vstupe z koaxiálneho vedenia. Poslednú časť tvorí digitálny zosilňovač. Vysielacia dióda je rýchla LED dióda s vysokým jasom, pretože slúži k prenosu dát cez APP (takéto vysoko svietivé diódy sa používajú napríklad v zadných svetlách niektorých automobilov). LED pri prenose dát bliká na 10MHz [2]. Celé zariadenie, ktoré je zobrazené na obr. 2a, sa umiestni do rúry so šošovkou, ktorá svetelný kužeľ usmerní do kužeľa svetla s minimálnym rozptylom obr. 2b, c.



Obr. 2 Vysielač systému RONJA [1]

Prijímač:

Úlohou prijímača je zosilniť a upraviť signál z vysielača. Je umiestnený v ochrannnej rúre rovnako ďaleko ako vysielač. Šošovka v tomto prípade zaostruje prichádzajúce svetlo do bodu, ktorý svieti priamo na prijímaciu diódu. Týmto mechanickým riešením sa zvýši účinnosť prenosu dát. Prijímač je najviac citlivý na šum a okolité rušenie, ako aj na silné slnečné svetlo (vysoká jednosmerná zložka signálu). Je napájaný 12 V napätím, pričom napájanie je filtrované a blokované niekoľkými kondenzátormi, aby sa modul mohol napájať z ďalšieho vodiča. Ako prijímací prvok v zapojení je použitá PIN fotodióda [2]. Je zapojená do série s odporom pomocou ktorého nastavujeme pracovný bod a prepätie pre fotodiódu. Modul prijímača je zobrazený na obr. 3.



Obr. 3 Prijímač systému RONJA [1]

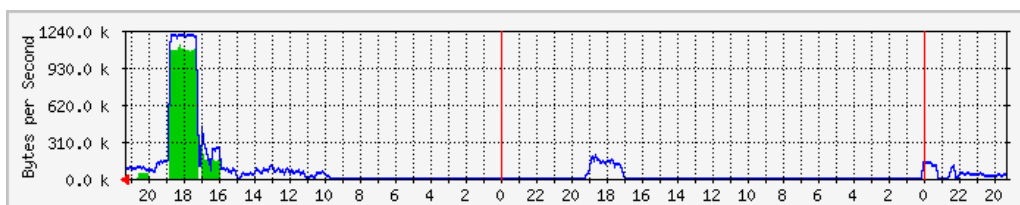
B. Technické parametre komunikačného systému RONJA:

- Rýchlosť prenosu: 10 Mbit/s, full-duplex;
- Maximálna pracovná vzdialenosť: 700 m s 90 mm šošovkou (1300 m so 130 mm šošovkou);
- Minimálna pracovná vzdialenosť: 1/15 maximálnej pracovnej vzdialenosti;
- Dátové rozhranie: prepojovacie rozhranie 10BaseT (UTP);
- Prikon: 260 mA, 12 V (3,1 W) z PC zdroja;

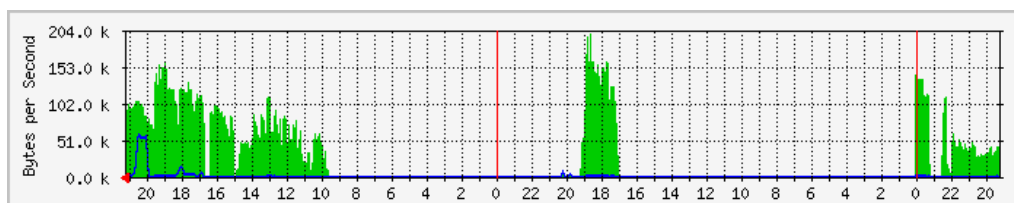
- Operačná vlnová dĺžka: viditeľná, 625 nm, 100 nm šírka spektra (červená farba);
- Odhadovaný optický výkon: 12 mW
- Rozptyl svetelného kužela - polovica uhlu: 1,9 mrad;
- Operačná vlhkosť: 100 % s vyhrievacím výkonom šošoviek 1 W;
- Viditeľnosť: musí byť k dispozícii priama optická viditeľnosť.

C. Vyťaženie prenosového systému RONJA

Z grafických priebehov je zrejmé, že takto vytvorený spoj má stále dostatočnú šírku pásma pre prístup do internetu aj v čase zaťaženia tohto spoja. To dokumentuje aj fakt, že pri vyťažení príkazom ping rýchlosť prenosu z internetu takmer neklesla, stále tam bola prevádzka približne 100 kB/s. Ďalej na grafe vyťaženia optického spoja (obr. 4) vyplýva, že spoj je skutočne realizovaný v plne duplexnom režime, teda max. 10 Mbit/s na strane vysielania a max. 10 Mbit/s na strane prijímu. Pri meraní bolo dosiahnuté 1206,3 kB/s v smere vysielania a 1137,5 kB/s v smere prijímu, spolu v oboch smeroch to tvorí 2343,8 kB/s, čo je 18,75 Mbit/s. Takýto výsledok je možný vďaka použitiu full-duplex sieťových kariet, pri použití sieťových kariet half-duplex by súčet prevádzky vysielateľa a prijímateľa mohol byť najvyššie 10 Mbit/s (čo je stále takmer dvojnásobne viac oproti Wireless LAN spojeniam, ktoré sú kompatibilné so štandardom IEEE 802.11b a pracujú na nelicencovanej frekvencii 2,4 GHz).



Obr. 4 Graf vyťaženia na strane optického spoja (zelenou farbou je prijím, modrou farbou je vysielanie)



Obr. 5 Graf vyťaženia na strane pripojenia na internet (zelenou farbou je prijím, modrou farbou je vysielanie)

III. ZÁVER

RONJA komunikuje s dvoma počítačmi pomocou optického dvojbodového dátového spojenia na veľkú vzdialenosť. Maximálny dosah tohto spojenia je až 1300 m a to za podmienky priamej viditeľnosti. RONJA dokáže prenášať dáta konštantnou rýchlosťou 10 Mb/s v plne duplexnom režime. V súčasnej dobe je to asi najlacnejšie zariadenie pre optickú komunikáciu atmosferickým prenosovým prostredím (APP) dvoch počítačov s takou veľkou priepustnosťou. Je kompatibilná k pripojeniu so switch-om alebo ethernetovou kartou s RJ45 konektorom. Prenos je ťažko odpočúvať alebo rušiť pretože vysielá veľmi úzky lúč svetla. Najväčšou nevýhodou je, že prenos nefunguje za veľmi zlého počasia ako je hmľa, veľmi hustý dážď alebo silné sneženie.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] <http://www.chrudim2000.cz/ronja/>
- [2] <http://ronja.twibright.com/>
- [3] <http://www.root.cz/clanky/ronja-reloaded/>
- [4] <http://www.root.cz/clanky/ronja-dcera-matfyzaka/>

Experimentálny systém FSO pre 850 nm LightPointe Fligh Strata 155E

Pavol MIŠENČÍK, Ján TURÁN, Ľuboš OVSENÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

pavol.misencik@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — Táto publikácia sa podrobne zaoberá experimentálnym systémom FSO pre 850nm LightPointe Fligh Strata 155E, ktorý je implementovaný na KEMT FEI TUKE. Venujeme sa popisu tohto systému, analýze jeho vlastností a pripojeniu k sieti.

Kľúčové slová — FSO, free space optics, LightPointe Fligh Strata systém, Flight Strata 155E

I. ÚVOD

FSO je plne duplexná optická linka, ktorá využíva optickú nosnú vlnu na prenos dát atmosférickým prenosovým prostredím (APP). Optický výkon je sústredený do jedného alebo viacerých úzkych optických lúčov. FSO systémy predstavujú dobrú alternatívu k optickým vláknovým systémom a to hlavne vďaka jednoduchosti implementácie. Tento systém je veľmi užitočný v prípade, keď uloženie optických vlákien do zeme je nevyhovujúce (dočasné využívanie, nepriaznivý terén, atď.) [1].

FSO prenosové linky sú ovplyvňované APP (počasím), čo môže znížiť dostupnosť a môžu nastať chyby prenosu, ktoré sa nevyskytujú pri vláknovom prenose.

Tento článok opisuje experimentálny systém FSO implementovaný na KEMT FEI TUKE pracujúci na vlnovej dĺžke 850nm, t.j. systém LightPointe Flight Strata 155E, jeho možné využitie a pripojenie k sieti.

II. LIGHTPOINTE SYSTÉM

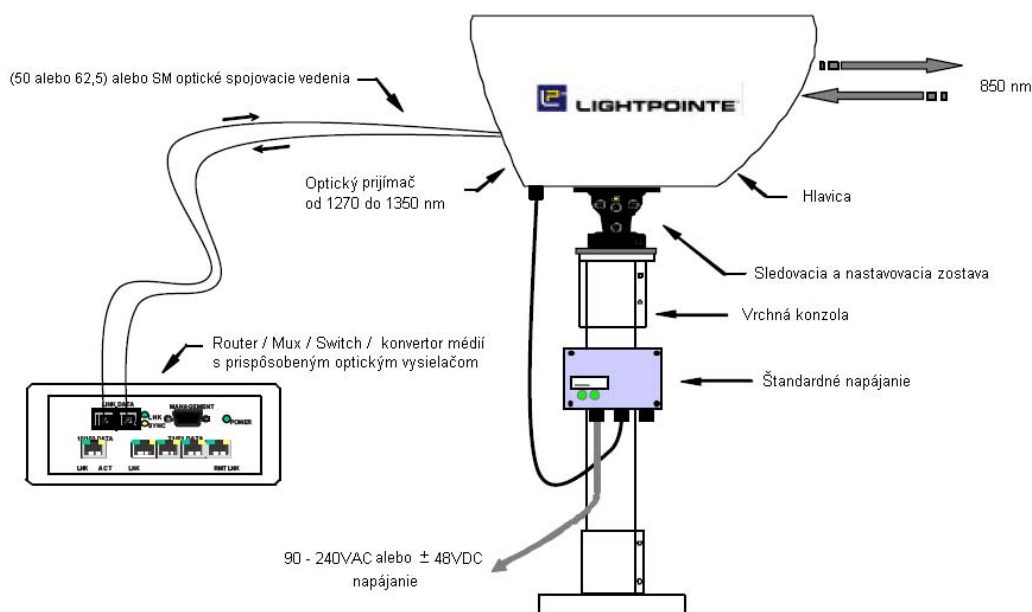
FlightStrata 155E je jediný produkt na trhu ktorý kombinuje smerovú stabilitu so štvorlúčovým systémom. Tento model s osvedčenou technológiu viacnásobných lúčov dal LightPointe na ďalšiu úroveň tým, ako zabezpečiť najvyššiu možnú dostupnosť a výkon samostatného optického bezdrôtového riešenia. Objektív prijímača kombinuje tvarovanie optického lúča (OBS - Optical Beam Shaping) s automatickou kontrolou výkonu (APC – Automatic Power Control) a zameriava zmeny APP a pohyb budov.

LightPointe FSO systém komunikuje pomocou viacnásobného infračerveného lúča, ktorý je neviditeľný pre ľudské oko. Obidve hlavice tohto FSO systému vyžadujú priamu viditeľnosť medzi sebou, takzvané line-of-sight. Systémy pracujú s vlnovými dĺžkami, pre ktoré nie je potrebná rádiový frekvenčná licencia ani licencia pre spektrum [2].

Na obr. 1 je znázornený FSO systém LightPointe Fligh Strata 155E. Hlavica vysiela štyri infračervené lúče svetla, ktoré prenášajú dáta zo siete. Hlavica na opačnej strane prijíma lúče na štyri prijímače, odfiltruje ich na filtroch optického signálu a vyšle ich do optického vlákna, ktoré spája hlavicu so sieťou. Systém vždy pracuje v režime plný duplex preto obidve hlavice vysielať a prijímať celý tok dát súčasne v oboch smeroch.

Typické možnosti aplikácie systému:

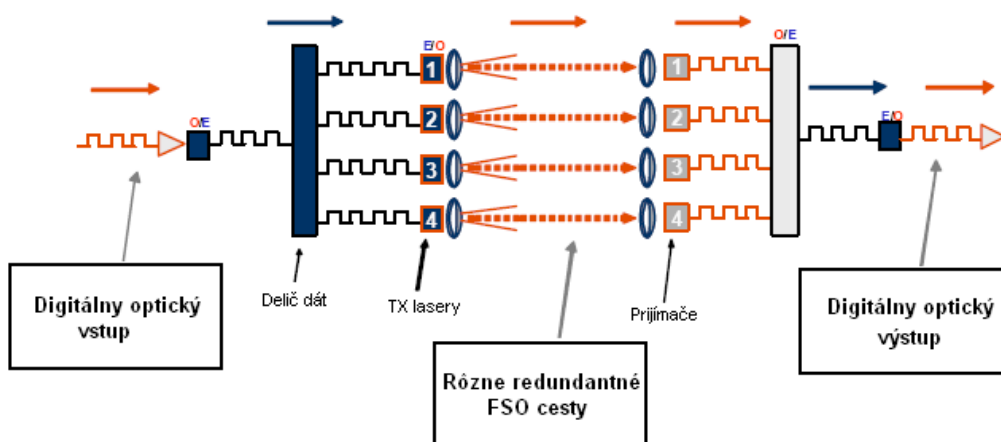
- Podnikové LAN a rozšírenie PABX;
- Metropolitné siete;
- Mobilná bezvláknová sieť.



Obr. 1 FSO systém [3]

A. Princíp prenosu

Systém FlightStrata využíva štyri vysielacie lúče a štyri prijímacie šošovky ako zobrazuje obr. 2. Každý vysielací lúč je plne zaťažený a poskytuje ochranu zlyhania linky voči blokovaniu preletom vtákov, APP a ďalšími vplyvmi.



Obr. 2 Blokový diagram štvor lúčového prenosového FSO systému [3]

B. Vysielacie komponenty

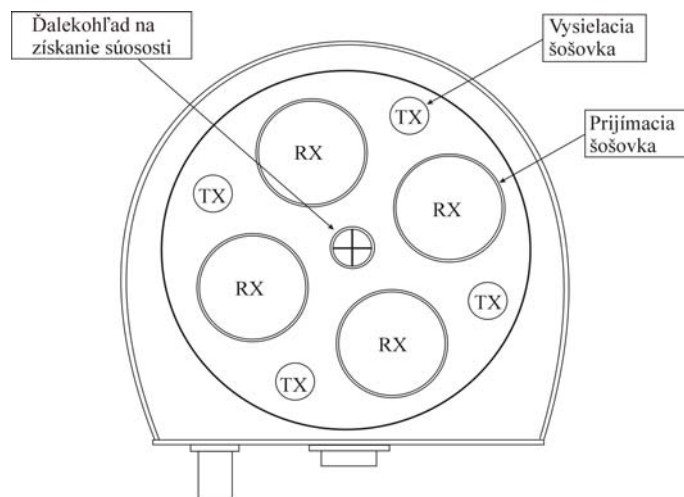
Optické signály z dátovej siete sú prijaté na dátový vysielateľ a prevedené na štyri rovnaké dátové toky. Štyri VCSEL lasery konvertujú elektrické signály na optické, ktoré sú prenášané ako štyri infračervené lúče vo voľnom prostredí [3].

C. Prijímacie komponenty

Prijímací zosilňovač prevádza infračervené signály prichádzajúce z voľného prostredia na elektrický signál, ktorý riadi optický vysielateľ- prijímač (transceiver) posielajúci pôvodný signál do dátovej siete na vzdialenom konci.

D. Geometria objektívu

Vysielacie (malé, priemer 2,5 cm) a prijímacie (veľké, priemer 8 cm) šošovky (zobrazené na obr. 3) sú dokonale umiestnené a musia byť centrované pre zladenie medzi vysielateľom a prijímačom na všetky vzdialenosti.



Obr. 3 Orientácia šošoviek na hlavicí [3]

E. Pripojenie k sieti

Prístroje Lightpointe FlightStrata 155 sú protokolovo transparentné a môžu podporovať hociaký bod sieťovej architektúry. Vstupná požiadavka je digitálny signál privedený na hlavicu cez jedno módové alebo viac modové vlákno (v závislosti od zakúpeného systému) využívajúci štandardné SC optický spojovacie vlákna [2].

Možnosti pre Local Area Network (LAN):

- 10/100 Ethernet;
- FDDI.

Možnosti pre ATM/SONET Network:

- STM-1, SDH;
- OC-1, OC-3, SONET.

Možnosti pre T1/E1 (Single/Multiple T1/E1):

- PABX, Mobilné Wireless, Digitálne Video.

III. ZÁVER

Základnými výhodami Lightpointe FlightStrata 155 technológie sú relatívne malé náklady, vysoká bezpečnosť a dobrá prenosová rýchlosť, ktorá je od 1,5 do 155 Mbps. RF prenosový systém v porovnaní s FSO Lightpointe FlightStrata systémom trpí interferenciami, je málo odolná voči rušeniu, a je často licencovaná. LightPointe optické systémy využívajú polovodičové lasery ako vysielacie zdroje. Pri využívaní tohto systému musíme brať na vedomie príslušné predpisy o bezpečnosti laserového zariadenia a prijať potrebné opatrenia, aby nedošlo k priamemu vystaveniu očí laserovým lúčom. Infračervený laserový lúč je neviditeľný a má potenciál preniknúť na sieťnicu a spôsobiť tepelné poškodenie. FlightStrata optický laserový komunikačný systém využíva laser triedy M v súlade s medzinárodnými laserovými bezpečnostnými normami IEC / EN 60825-1 A2:2001.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] Hovořák P., Wilfert O. Free Space Optical Link For Measurement Of Transmission Quality Link Parameters. In *5th International Conference of PhD Students*. Miskolc: University Of Miskolc, 2005. s. 325-659. ISBN: 963 661 6736.
- [2] Lightpointe, FlightStrata™ 155E/ FlightStrata™ G, dostupné na internete: http://www.lightpointe.com/downloads/datasheets/FlightStrata155E_G.pdf
- [3] FlightTransport™ Systems, Instalation and Maintenance manual, (505-002000-0001M Rev L), 2006-01.

Experimentálny systém FSO pre 1550nm Fsona SONAbeamTM 155-E

Pavol MIŠENČÍK, Ľuboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

pavol.misencik@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — Tento článok sa zaoberá experimentálnym systémom FSO pre 1550nm Fsona SONAbeamTM 155-E, implementovaný na KEMT FEI TUKE. Venujeme sa popisu tohto systému, analýze jeho vlastností a pripojeniu k sieti.

Kľúčové slová — FSO, free space optics, Fsona SONAbeam systém, Fsona SONAbeamTM 155-E

I. ÚVOD

FSO (Free Space Optics) je bezvláknová, laserom riadená technológia, ktorá podporuje vysokorýchlostnú komunikáciu a veľkú šírku pásma s jednoduchou inštaláciou. Tieto systémy preto začínajú získavať prioritu na súkromnom trhu, ako rýchle a výhodné riešenia schopné nahradiť drahé optické vláknové systémy [1,5].

FSO je často implementovaná na predĺženie existujúcej širokopásmovej siete, kde iné technológie majú problémy ako sú čas, náklady, licenčné alebo infraštruktúrne obmedzenia. SONAbeamTM FSO systémy sú kompatibilné so všetkými sieťami a protokolmi – Ethernet, Fast Ethernet, Gigabit Ethernet, SONET/SDS, ATM, FDDI siete a pod. Tento transparentný protokolový prístup je dôležitý rys pre zákazníkov. Poskytuje oveľa väčšiu flexibilitu nasadenia, ktorá udržiava nízke náklady a umožňuje organický rast siete.

Tento článok opisuje experimentálny systém FSO pre 1550nm Fsona SONAbeamTM 155-E implementovaný na KEMT FEI TUKE, jeho možné využitie a pripojenie k sieti.

II. SONABEAMTM SYSTÉM

Produktová rada SONAbeam zahŕňa niekoľko modelov rozlíšených podľa bitovej rýchlosti a prevádzkovým rozsahom. Linky môžu byť navrhnuté tak, aby ich dostupnosť bola 99,999 % v rozsahu od 50m do 5000m. SONAbeam ponúka prenosové rýchlosti od 1,5 Mb/s až do 2,5 Gb/s. Medzi hlavné výhody technológie SONAbeam oproti konkurenčným produktom patria:

Oku bezpečná vlnová dĺžka 1550 nm, t.j. tieto systémy sú úplne bezpečné pre ľudské oko, pretože používajú laser s 1550 nm technológiou. To umožňuje zvýšiť výstupný optický výkon až 50 krát viac, ako je to povolené pre vlnovú dĺžku 850nm.

Vyšší výstupný výkon lasera umožňuje navrhnuť systémy pre prevádzku na väčšie vzdialenosti, lepší prienik lúča cez hustejšiu hmlu, väčšiu dostupnosť a samozrejme vyššie prenosové rýchlosti.

Škálovateľnosť - vzhľadom na to, že vlnová dĺžka 1550 nm je najčastejšie stanovená pre pozemné optické vláknové komunikácie, podporujúca technická infraštruktúra pre tento spektrálny rozsah je obrovská a rýchlo rastie vďaka každoročným investíciám do komponentov a podsystémov.

Vysoká spoľahlivosť v nadväznosti na bohaté skúsenosti firmy fSONA v oblasti komunikačných systémov pre vojenské laserové a kozmické aplikácie, sú všetky podsystémy vyrobené z vysoko spoľahlivých komponentov. Spoľahlivosť je zabezpečovaná aj testovaním systémov pre rôzne podmienky prevádzky v extrémnom rozmedzí od -50°C až do +75°C.

A. SONAbeam systém série „E“

SONAbeam série „E“ je veľmi univerzálny a kompaktný systém aj napriek svojmu

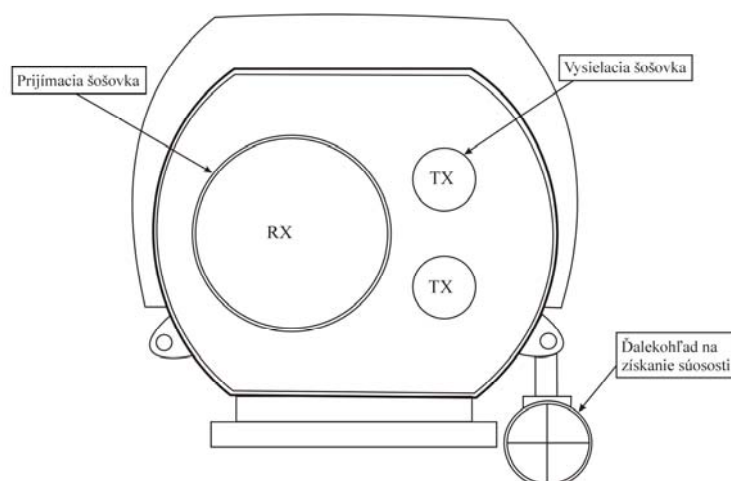
robustnému hliníkovému telu, ktoré ho predurčuje aj na využívanie v náročných poveternostných podmienkach. Je možné ho ľahko prepravovať a je ideálny pre situácie, ktoré vyžadujú rýchle nasadenie.

Terminál SONAbeam série „E“ je veľmi spoľahlivá, vysokorýchlostná a plne duplexná optická linka, ktorá využíva optickú nosnú vlnu na prenos dát atmosférickým prenosovým prostredím (APP). Pre komunikáciu využíva infračervené, pre zrak bezpečné laserové vysielacie a optický prijímač s veľkou prijímacou apertúrou a vstavaným mikroprocesorom. Terminály sú konštruované na prevádzku ako v budovách (in-door), tak aj vo voľnom prostredí (out-door) za podmienok priamej viditeľnosti medzi dvoma bodmi v sieti. Pracujú na vzdialenosť od niekoľkých metrov až po niekoľko kilometrov v prípade čistého vzduchu. Systém je navrhnutý tak, aby akceptoval rôzne užívateľské rozhrania a napájacie moduly. Skutočná využiteľná rýchlosť prenosu dát je závislá od podmienok prenosu. Terminály série „E“ môžu byť použité na prenosové rýchlosti od 1,5 až po 160 Mbps a podporujú množstvo komunikačných štandardov vrátane: T1/E1, E3/DS3, 10/100 Ethernet, OC-1/STM-0, OC-3/STM-1, rovnako ako SMTPE a HDTV štandardy na prenos videa [2,3,4,5].

S integrovanou správou siete (INM - Integrated Network Management) môžu byť terminály pripojené k podnikovým LAN sieťam a umožnia nám tým sledovať a spravovať terminály z kancelárskeho počítača.

B. Geometria objektívu

Predná časť hlavice obsahuje dve vysielacie šošovky, ktoré sú umiestnené pod sebou a obsahujú lasery s výkonom 2×160 mW. Objektív pozostáva iba z jednej prijímacej šošovky s priemerom 10 cm (obr. 1).



Obr. 1 Orientácia šošoviek na hlavici [3]

C. Popis systému a pripojenie k sieti

Na obr.2 sú vyobrazené a popísané všetky porty a svetelné indikátory, ktoré sa nachádzajú na zadnej strane hlavice FSO systému SONAbeam™ 155-E.

Linková slučka (LINK LOOPBACK THIS END): LED indikuje konfigurovanie terminálu tak, že FSO prijatý signál je priamo vysielaný späť do ďalšieho terminálu.

Vstupná slučka (INPUT LOOPBACK): LED indikuje konfigurovanie terminálu tak, že prijatý vláknový signál terminálovým RX portom, je vysielaný späť priamo do jeho TX portu.

RCV časové blokovanie (RCV CLOK LOCKED): LED indikuje, že terminálom prijímaný signál je nastavený na kryštálom riadenú dátovú rýchlosť, ktorá je nastavená pre terminál. (Ak je terminál nastavený pre prácu v režime Bypass alebo inú dátovú rýchlosť, táto LED je vždy OFF).

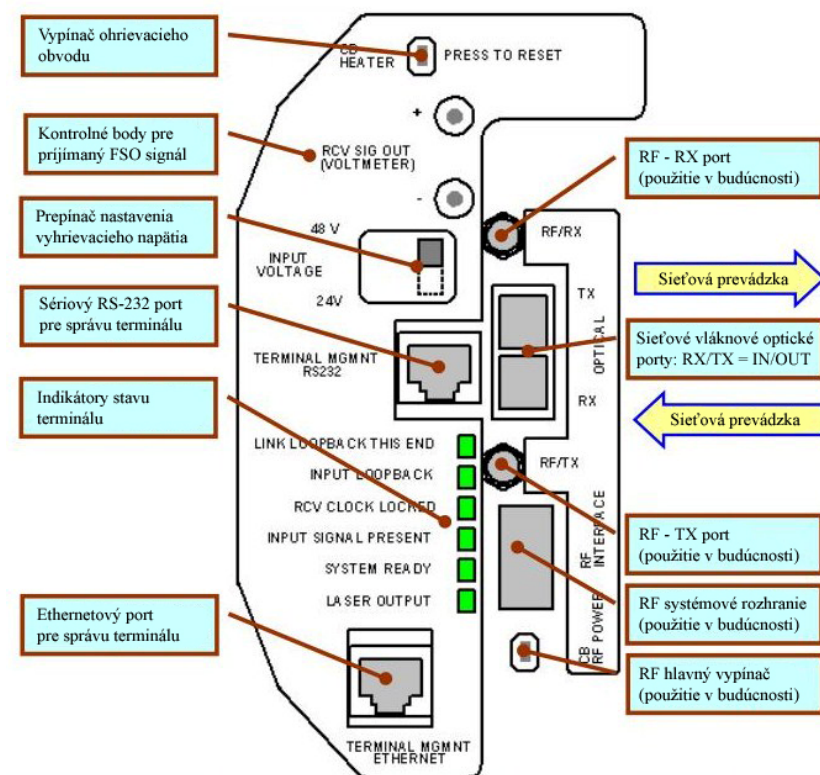
Vstupný signál pripravený (INPU SIGNAL PRESENT): LED indikuje že terminál detekuje vstupný signál zo siete na RX vlákne alebo drôtovom vstupe. (vláknovom alebo drôtovom – závisí od toho akú kartu terminálového rozhrania využívame).

Systém pripravený (SYSTEM READY): LED z počiatku bliká, pokiaľ terminál nedokončí svoj systémový test pripojenia a potom svieti stabilne a indikuje, že terminál je pripravený na konfiguráciu alebo na používanie.

Výstup z lasera (LASER OUTPUT): LED indikuje, že jeden alebo oba lasery sú aktívne.

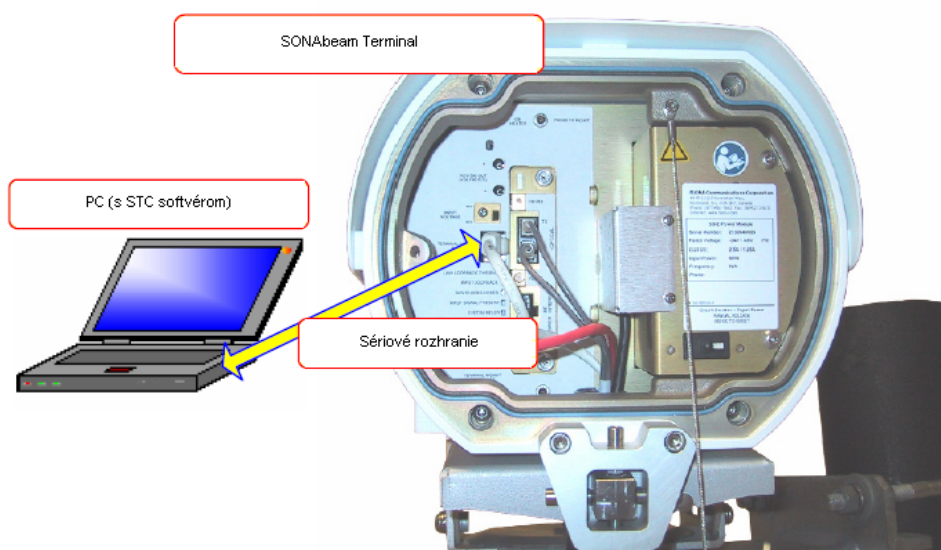
Vnútné ohrievače SONAbeam E-série (INPUT VOLTAGE): Vnútné ohrievače SONAbeam série „E“ terminálu sú konštruované na prevádzku 24 V alebo 48 V DC zdrojov.

Optické komunikačné vlákna (OPTICAL Tx a Rx): Rozhranie umožňuje pripojiť MM alebo SM optické vlákna ukončené SC konektorom pre komunikáciu v sieti. Jedno vlákno pre dáta vysielané (Tx) do ďalšieho terminálu v sieti a jedno pre príjem dát (Rx) z terminálu v sieti.



Obr. 2 Označenie portov a indikátorov na hlavici FSO systému SONAbeamTM 155-E [3]

Ethernetové a sériové pripojenie (TERMINAL MGMT ETHERNET, RS232): Na pripojenie siete LAN sa využíva Ethernetový vodič CAT5 s RJ45 koncovkou, ktorá slúži pre pripojenie vodiča do „TERMINAL MGMT ETHERNET“ portu na termináli SONAbeam (obr. 2). Pre aplikovanie sériového pripojenia sa využíva prípojka a koncovka sériového vodiča CAT5 a „TERMINAL MGMT RS232“ port na SONAbeam termináli. Dočasné nastavenie spojenia medzi PC a SONAbeam je prostredníctvom jednoduchého, troj-vodičového rozhrania. Toto rozhranie sa pripája na COM1 PC port a na „TERMINAL MGMT RS232“ SONAbeam (obr. 3).



Obr. 3 Pripojenie PC k terminálu SONAbeam [3]

D. Technická špecifikácia

- Prenosové rýchlosti: Fast Ethernet: 125 Mb/s obojsmerne,
OC-3/STM-1: 155 Mb/s obojsmerne,
- Prevádzkový dosah: útlm 3 dB/km pre čistý vzduch, dosah 50-3500 m,
útlm 10 dB/km pre extrémny dážď, dosah 50-1700 m,
- Výstupný výkon lasera: 320 mW špičky 2×160 mW,
- Veľkosť prijímacej apertúry: 10 cm,
- Vlnová dĺžka: 1550 nm,
- Pracovná teplota: od -40°C do +60°C.

III. ZÁVER

Komunikácia využívajúca fSONA systém bola založená v roku 1997 na poskytovanie jednoduchých bezvláknových riešení pre dopravcov, poskytovateľov služieb a podnikových sietí po celom svete. SONAbeamTM využíva nelicencované pásmo a pracuje s rýchlosťou prenosu dát až 1,5 Gbps na vzdialenosť do 5km. Tento systém využíva pre vysokorýchlostnú bezdrôtovú komunikáciu neviditeľné, ale zároveň bezpečné infračervené optické žiarenie.

SONAbeamTM šetrí náklady na zriadenie spojenia (nie je nutná tvorba optickej vláknovej infraštruktúry a získanie licencie pre rádiové frekvencie). Je imúnna voči rušeniu a prenos informácií je veľmi bezpečný. Je to veľmi kompaktný a univerzálny systém a aj napriek robustnému hliníkovému krytu je rovnako výhodná vonku v náročných poveternostných podmienkach, ako aj v prevádzkach, kde optický lúč prechádza cez okno. Systém môže byť ľahko prepravený do miesta inštalácie a je ideálny pre situácie, ktoré vyžadujú rýchle nasadenie.

Každý SONAbeam terminál série „E“ používa viacnásobne vyšší výkon, väčšiu šírku pásma, oku bezpečné 1550 nm lasery s patentovaným optickým prijímacím systémom na dosiahnutie vyššej výkonnosti za všetkých poveternostných podmienok.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] Hovořák P., Wilfert O. Free Space Optical Link For Measurement Of Transmission Quality Link Parameters. In *5th International Conference of PhD Students*. Miskolc: University Of Miskolc, 2005. s. 325-659. ISBN: 963 661 6736.
- [2] Fsona SONAbeam E, http://www.fsona.com/prod/SONAbeam_E.pdf
- [3] SONAbeamTH E Series Installation Manual
- [4] fSONA Announces the SONAbeam 155-E, First in a New Series of Free Space Optical Wireless Products, <http://www.thefreelibrary.com/fSONA+Announces+the+SONAbeam+155-,+First+in+a+New+Series+of+Free...-a0102752117>
- [5] SONABEAM FROM FSONA, http://www.multicap.biz/files/product_download/1271777530/SONAbeam-Datasheets%5B1%5D.pdf

Projektovanie a analýza FSO systémov s využitím programového balíka „FSO System Simulator”

Pavol MIŠENČÍK, Ľuboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

pavol.misencik@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — Táto publikácia sa podrobne zaoberá modelovaním optických komunikácií voľným prostredím (FSO - Free Space Optics) pomocou programového balíka „FSO System Simulator“, ktorý bol navrhnutý a implementovaný na KEMT FEI TUKE. Simulácia FSO linky má veľký význam pri navrhovaní a pochopení súvislostí takýchto spojov v závislosti od najrôznejších parametrov, či už technických alebo neustále sa meniacich parametrov prenosového kanála. Článok preto stručne popisuje tento programový nástroj a popisuje experimenty realizované týmto systémom a ich výsledky.

Kľúčové slová — FSO, FSO System Simulator, statický model FSO, štatistický model FSO

I. ÚVOD

FSO systémy pracujú pri meniacich sa podmienkach, ktoré nedokážeme presne odhadnúť a ani sa im vyhnúť – meniace sa počasie. Preto je veľmi potrebné poznať správanie sa FSO prenosovej linky pri akýchkoľvek podmienkach a vedieť tak zvoliť vhodné parametre na prenos v konkrétnom prostredí, či už ide o 200 metrovú linku ponad mestský park alebo 2 kilometrový záložný spoj. Správanie sa FSO prenosovej linky v určitých oblastiach, kde poznáme štatistické parametre atmosférického prenosového prostredia (APP), počas dostatočného dlhého obdobia, popisuje tzv. štatistický model FSO. To na akú vzdialenosť bude optický prenosový systém fungovať s ohľadom na parametre prijímacích a vysielacích zariadení, ako sú vysielací výkon, citlivosť prijímača, priemer vysielacej a prijímacej šošovky, smerovosť laserového zväzku, vzdialenosť medzi vysielacom a prijímačom nám popisuje statický model FSO [1,3].

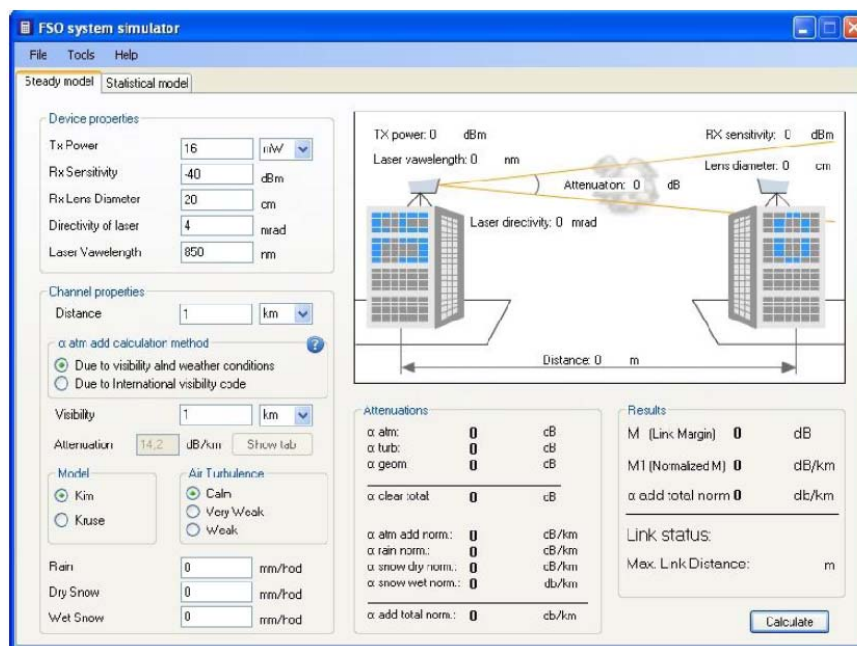
Počítačový program „FSO System Simulator“ obsahuje všetky prvky a zariadenia FSO spoja. Úlohou tohto programu je pomocou vstupných technických parametrov zariadení, vzdialenosti a podmienok prenosového kanála vypočítať dostupnosť, resp. nedostupnosť linky v danom prostredí.

II. FSO SYSTEM SIMULATOR

A. Opis programu

Nami vytvorený programový balík „FSO System Simulator“ je vytvorený vo vývojovom prostredí Microsoft Visual C# 2008 Express Edition, nástroji pre tvorbu grafických aplikácií. Po otvorení programu sa zobrazí grafické okno, ktoré je rozdelené na dve polovice. Ľavá polovica okna pozostáva z roliet, kde zadávame a volíme parametre prenosu a pravá polovica okna slúži na zobrazenie vypočítaných výsledkov. Na obr. 1 je úvodné okno, ktoré sa objaví pri spustení programu. Veľkosť okna je 800 x 600 o.p. (obrazových prvkov), čo je optimálne pre dané množstvo ovládacích prvkov a taktiež kvôli dostupnosti zobrazenia na všetkých typoch monitorov. Čo sa týka manipulácie s oknom, tá je klasická, ako pri ostatných oknových aplikáciách, ale možnosť maximalizácie alebo zmeny veľkosti okna je zakázaná z dôvodu zachovania konzistentnosti a čitateľnosti programu [2,4].

Graficky, ale aj tematicky je program rozdelený na dve časti a to: Statický model (Steady model) a Štatistický model (Statistical model). Kliknutím na danú záložku si vyberieme daný model a s tým súvisiace grafické prvky, vstupy a výstupy programu.



Obr. 1 Úvodné grafické okno programu „FSO System Simulator“

Ľavú stranu okna (obr. 1) tvoria polia pre zadávanie vstupov užívateľom. Prvé pole je pole „Device properties“, slúžiace na zadanie parametrov vysielačieho zariadenia (program predpokladá, že prijímacie zariadenie má rovnaké parametre) a to: stredná hodnota vysielačieho výkonu lasera, citlivosť prijímača, priemer prijímacej šošovky, smerovosť laserového zväzku a vlnová dĺžka použitého lasera. Všetky textové okienka pri prekročení povoleného rozsahu príslušnej veličiny zobrazia chybové hlásenia, v ktorých upozorňujú užívateľa na interval, z ktorého je potrebné daný parameter zadať.

Ďalším poľom v oblasti zadávania vstupov je pole „Channel properties“, kde má užívateľ možnosť zadať podmienky prenosového kanála, t.j. APP: vzdialenosť medzi vysielačom a prijímačom, možnosť voľby výpočtu aditívneho útlmu spôsobeného zhoršenými atmosférickými podmienkami, hodnotu útlmu, ktorú vyberie z Medzinárodnej tabuľky viditeľnosti „Show tab“ (obr. 2) [2], ďalej má užívateľ možnosť voľby modelu potrebného pri výpočte útlmu spôsobeného rozptylom a možnosť voľby určitého modelu turbulentnosti atmosféry.

Weather conditions	Precipitation		Visibility	Attenuation (dB/km) at $\lambda = 785\text{nm}$
		mm/hod		
Dense fog	Snow	Storm	0 m	350
Thick fog			50 m	339,6
Moderate fog			200 m	84,9
Light fog			500 m	34,0
Very light fog		Strong rain	1 km	20,0
Light mist			1,9 km	14,2
Very light mist			2 km	6,7
Clear air			2,8 km	4,6
Very clear air		Average rain	4 km	3,0
			5,9 km	1,8
		Light rain	10 km	1,1
			18,1 km	0,6
			20 km	0,53
			23 km	0,46
		Drizzle	50 km	0,21

Obr. 2 Medzinárodná tabuľka viditeľnosti

Pravú časť okna programu tvoria vypočítané a zobrazené výstupy. Nad výstupmi v textovej podobe je umiestnený obrázok reprezentujúci výsledky programu v grafickej jednoznačnej podobe.

Pole „Attenuations“ umiestnené v spodnej časti reprezentuje hodnoty jednotlivých útlmov vzniknutých na prenosovej ceste samostatne. Toto pole útlmov je tvorené čiastkovými útlmami: útlm na časticiach, útlm spôsobený turbulentnosťou atmosféry, útlm spôsobený rozptylom

zvázku laserového lúča medzi vysielateľom a prijímačom, výsledný útlm spôsobený čistou atmosférou, útlm spôsobený rozptylom na časticiach menších, ako je použitá vlnová dĺžka (najmä útlm spôsobený hmlou), útlm spôsobený dažďom, útlm spôsobený suchým (ľahkým, nie mokrým) snehom, útlm spôsobený mokrým snehom a výsledný útlm spôsobený zhoršenými atmosférickými podmienkami.

Pole „Results“ umiestnené napravo od poľa „Attenuations“ reprezentuje celkové výsledky programu a zhodnotenie navrhovaného spoja. Je tvorené: hodnotu bezpečnostného okraja, t.j. hodnotu, aký veľký útlm by ešte zvládol navrhnutý prenosový systém; hodnotu Link Margin vztiahnutú k požadovanej vzdialenosti; hodnotou výsledného útlmu spôsobeného zhoršenými atmosférickými podmienkami. „Link status“ predstavuje zobrazenie vyššie popísanej skutočnosti, t.j. či systém so zadanými vlastnosťami bude alebo nebude fungovať. Tieto stavy reprezentujú na prvý pohľad jednoznačné heslá: „Link OK“ a „Link DOWN“.

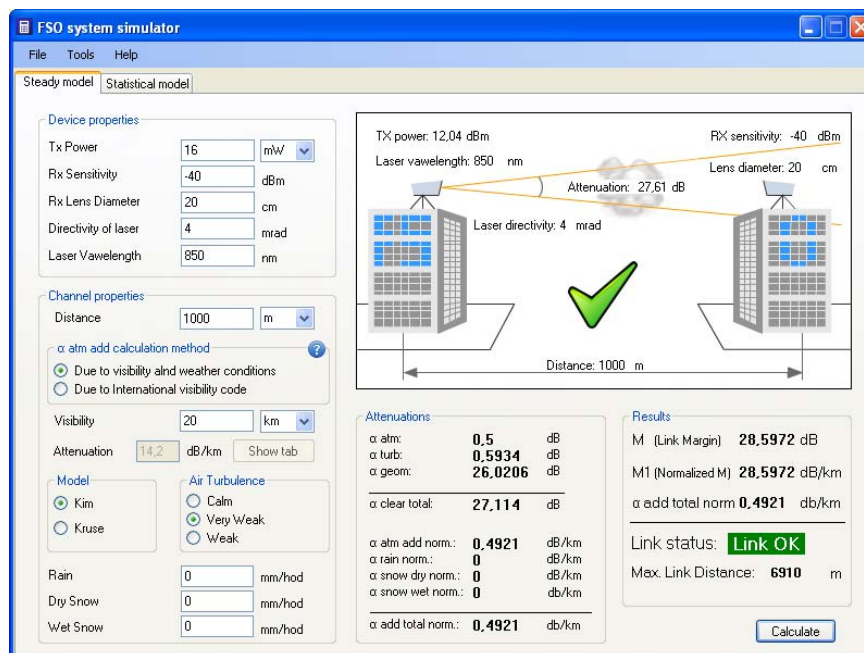
B. Experimenty s „FSO System Simulator“

Pomocou simulačného programu „FSO System Simulator“ bolo vykonaných veľa experimentov spoľahlivosti dosahu FSO linky pri zadaní rôznych vstupných podmienok. Hlavnou požiadavkou bolo zistenie, či navrhovaný FSO prenosový systém bude v prevádzke bez obmedzení a či bude funkčný.

Nech vstupné parametre skúmaného FSO systému budú:

- Stredná hodnota vysielaného výkonu lasera je 16 mW;
- Citlivosť prijímača je -40 dBm;
- Priemer prijímacej šošovky je 20 cm;
- Smerovosť laserového zväzku je 4 mrad;
- Vlnová dĺžka použitého lasera je 850 nm;
- Vzdialenosť, na ktorú chceme prenos uskutočniť je 1000 m;
- Viditeľnosť je 20 km (čistý vzduch).

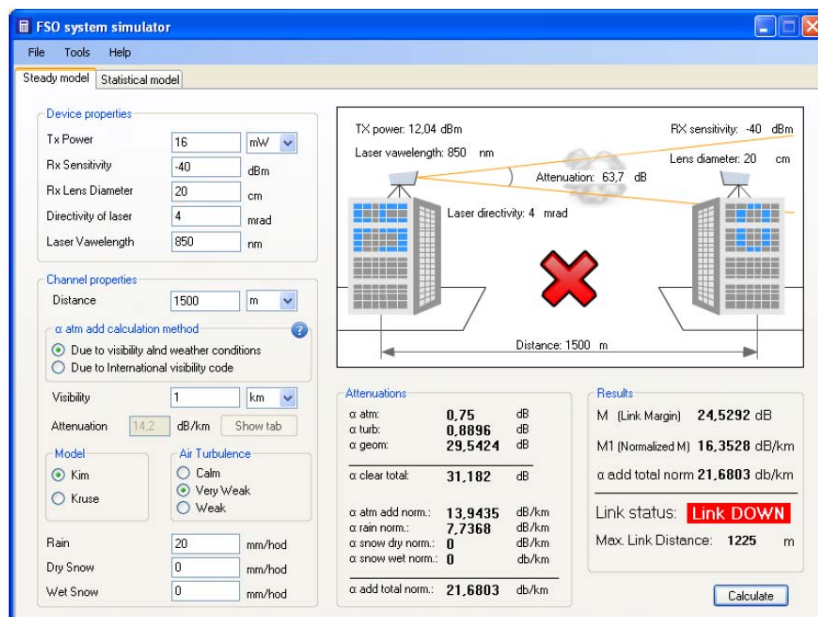
Tieto vstupné hodnoty sú zadané ako vstupné parametre programu, je zvolený Kimov model, turbulentnosť atmosféry Very Weak a predpokladáme prenos, ktorý sa uskutoční bez dažďa alebo snehu. Po spustení programu tlačidlom „Calculate“ Simulator prepočíta všetky potrebné hodnoty a výsledky zapíše do jednotlivých okien (Pole útlmov - Attenuation a Pole výsledkov - Results) obr. 3. Na prvý pohľad je zrejmé, že tento prenos bude úspešný a to vďaka grafickej reprezentácii výsledku v pravej hornej časti okna (obr. 3). Ako môžeme vidieť táto linka bude fungovať až na vzdialenosť 6910 m pri týchto vstupných podmienkach. Celkové výsledne okno programu je zobrazené na obr. 5.



Obr. 3 Stav navrhovanej linky je „Link OK“

V nasledujúcom experimente použijeme rovnaké parametre FSO systému, ale zmeníme požadovanú vzdialenosť prenosu na 1500 m a pridáme dažďové zrážky. Číže do políčka „Rain“ napíšeme hodnotu zrážok a to 20 mm/hod. Takto nastavený experiment dáme otestovať. Ako môžeme z výsledkov simulácie vidieť (obr. 4), reálny prenosový FSO systém pri takto

nastavených podmienkach prenosu zlyháva, a teda nie je schopný udržať spojenie na 100 %. Maximálna vzdialenosť, ktorú FSO systém pri takýchto podmienkach dokáže prekonať je len 1225 m.



Obr. 6 Stav navrhovanej linky je „Link DOWN“

III. ZÁVER

Optické komunikácie voľným prostredím sú vyvíjajúcou sa technológiou, ktorá sa podľa nášho názoru v blízkej budúcnosti rýchlo rozšíri od objektov potrebných spoľahlivú zálohu a vysokorýchlostný dátový prenos (banky, veľké radiace podniky), cez rôzne inštitúcie umiestnené v husto obývaných a zastavaných oblastiach až po internetových providerov, na ktorých trh tlačí s potrebou čoraz väčšej priepustnosti dátových liniek.

Pre návrh kvalitných a efektívnych FSO systémov je potrebné podrobne poznať každý element zariadení a vidieť súvislosti medzi zmenami parametrov zariadení, zmenami vlastností prenosového kanála a výsledným efektom. Simulácia optických komunikácií voľným prostredím je preto nevyhnutnou pomôckou pri navrhovaní alebo experimentovaní s takýmito zariadeniami. Počítačovým simulačným programom „FSO System Simulator“, je možné jednoduchým a intuitívnym spôsobom uskutočňovať rôzne experimenty založené na doterajších teoretických poznatkoch štúdia FSO systémov. Užívateľ z výsledkov programu vyčíta údaje o útlmoch pôsobiacich v FSO prenosovom kanáli a konečný výsledok vrátane teoretického návrhu na maximálnu dosiahnuteľnú vzdialenosť prenosu FSO systému za daných podmienok. Jednoduchou zmenou parametrov v programe dokáže užívateľ predpovedať teoretické správanie sa FSO linky počas rôznych stavov APP, vyskytujúcich sa počas všetkých ročných období (hmla, dážď, sneh).

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] R. Kvicala, V. Kvicera, M. Grabner, O. Fiser, „BER and Availability Measured on FSO Link,“ *Radioengineering*, Vol. 16, No. 3, 2007.
- [2] I. Kim, B. McArthur, E. Korevar, „Comparison of Laser Beam Propagation at 785 nm and 1550 nm in Fog and Haze for Optical Wireless Communications,“ in *Proceedings of SPIE*, Volume 4214, Boston, USA, 2001, pp. 26-37.
- [3] M. Alnaboulsi, H. Sizun, F deFornel, „Fog Attenuation for Optical and Infrared Waves“, *Journal of Optical Eng.*, Vol.43, 2004, pp. 319-329.
- [4] O. Bouchet, D. O'Brien, M. El Tabach, M. et al. „State of the Art – Optical Wireless,“ Deliverable D4.1, *ICT – Omega*, 2008.

Experimentálne pracovisko pre meranie hustoty hmly

Matúš TATARKO, Ľuboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

matus.tatarko@student.tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — V tejto publikácii je popísaná činnosť prístroja na meranie hustoty hmly, vlhkosti a teploty prostredia. Dlhodobé meranie týchto parametrov nám umožňuje vytvoriť modely prenosového prostredia pre optické komunikácie voľným prostredím (FSO - Free Space Optics). Toto zariadenie môže slúžiť aj pre meranie poveternostných podmienok počas jazdy bezpečného a ekologického automobilu budúcnosti a tým zvýšiť jeho aktívnu bezpečnosť.

Kľúčové slová — hmlový senzor, bezpečný automobil, FSO, automobil budúcnosti

I. ÚVOD

Free Space Optics (FSO) je zariadenie používajúce modulovaný viditeľný alebo infračervený lúč na dosiahnutie optickej komunikácie vzduchom. FSO systém vysiela očiam bezpečný optický lúč z jedného zariadenia do druhého použitím LED diódy alebo lasera s malým výstupným výkonom. Prijímač tohto lúča je vybavený vysoko citlivým fotodetektorom. FSO systémy fungujú na vzdialenosť niekoľkých kilometrov (priama viditeľnosť) pokiaľ nie je medzi zariadeniami žiadna prekážka a je zabezpečený dostatočný výkon optického vysielača.

Keďže sa jedná o systém, ktorého prenosové médium je voľné prostredie, tak najväčšou nevýhodou tejto technológie je silná závislosť spoľahlivosti spoja od poveternostných vplyvov a počasia. Existuje viacero rušivých vplyvov na kvalitu a dosah prenosu. Hlavným rušivým vplyvom v optických komunikáciách voľným prostredím je hustá hmla. Hmla je para, zložená z vodných kvapôčok, ktorých priemer je len niekoľko stoviek mikrometrov, ale pri používaných vlnových dĺžkach v FSO systémoch a pri kombinácii vplyvov absorpcie, rozptylu a odrazu dochádza k degradácii signálu alebo až k úplnému zabráneniu prechodu svetla cez prostredie.

II. SENZOR HMLY

Senzor hmly je zariadenie, ktoré meria údaje o hustote hmly, teplote a relatívnej vlhkosti prostredia. Sú to dôležité parametre pre následné statické a štatistické vyhodnocovanie kvality optickej komunikácie voľným prostredím [1].

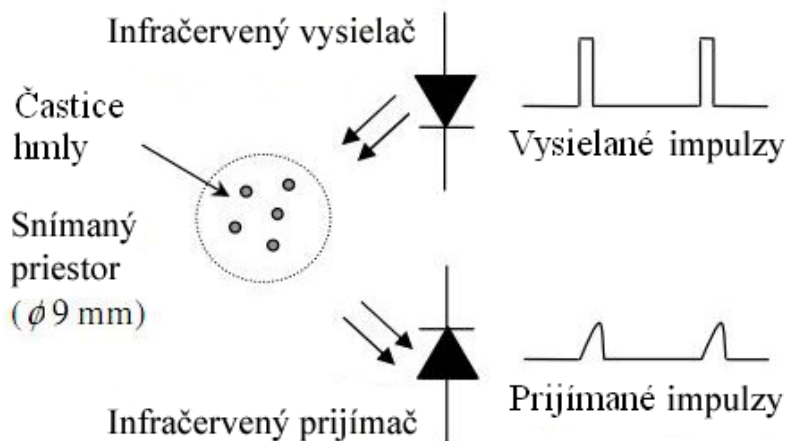
Za účelom odhadu útlmu spôsobeného hmlou sa meria obsah kvapalnej vody v jednotke objemu (g/m^3). Táto úloha vyžaduje pomerne zložitú a nákladnú meraciu prístroje. Na experimentálne účely bol vyvinutý jednoduchý a nízko nákladový prístroj, ktorý meria kvantitatívne hodnoty kvapalnej vody z ovzdušia [2].



Obr. 1 Senzor hmly – vonkajšia jednotka

A. Princíp merania hustoty hmly

Na obr. 1 znázornená vonkajšia jednotka senzora hmly osadeného na stene Lab. Optoelektronických systémov KEMT FEI TUKE. Vonkajšia jednotka pravidelne vysiela a prijíma krátke infračervené optické impulzy. Amplitúda odrazeného optického signálu súvisí s odrazivosťou prostredia v oblasti senzora, táto informácia môže byť použitá na dobrý odhad obsahu kvapalnej vody. Okrem optického senzora je prístroj vybavený aj senzormi teploty a relatívnej vlhkosti vzduchu. Tieto parametre poskytujú ďalšie dôležité informácie a majú charakteristické hodnoty pri hmlistom počasi. Snímač je stavaný do dažďa a je tienený pred priamym slnečným svetlom, ale zároveň umožňuje voľný pohyb kvapôčkam hmly do a zo snímača [2].



Obr. 2 Princíp merania hustoty hmly optickým senzorom

B. Technické prostriedky senzora hmly

Zariadenie (senzor hmly) pozostáva z vonkajšej a vnútornej jednotky (obr. 3, 4).



Obr. 3 Vonkajšia jednotka

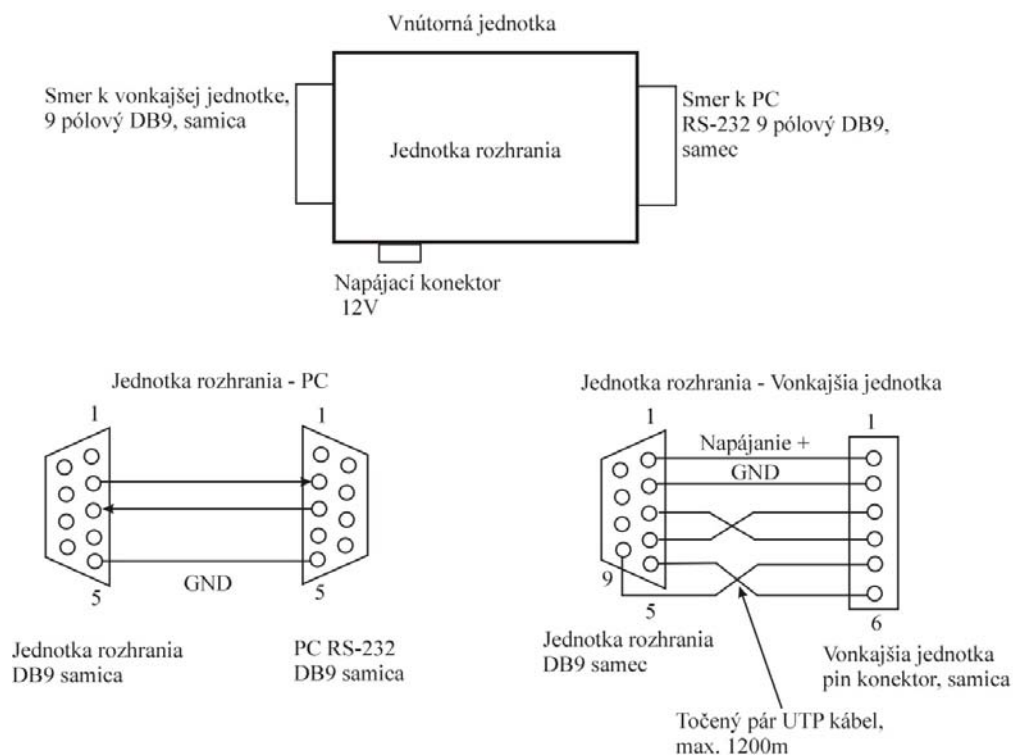


Obr. 4 Vnútna jednotka

Mikroprocesor vonkajšej jednotky má sériové komunikačné rozhranie (RS-422) a je stavaný na meranie obsahu kvapalnej vody vo vzduchu, teploty a relatívnej vlhkosti. Vnútna jednotka systému (obr. 4) prijíma informácie z vonkajšej jednotky (obr. 3) pomocou rozhrania RS-422, prekonvertuje ich na rozhranie RS-232 a zariadenie je napojené k štandardnému počítaču typu PC pomocou sériovej linky. Vnútna jednotka napája aj vonkajšiu jednotku, a preto vonkajšia jednotka nepotrebuje samostatný zdroj napájania.

Komunikácia medzi vonkajšou a vnútornou jednotkou je pomocou rozhrania RS-422 s nasledujúcimi parametrami:

- 38400 kbit/s, nesúbežne, 1 stop bit, žiadna hardvérová kontrola toku;
- Maximálna dĺžka kábla je približne 1200 m [1].



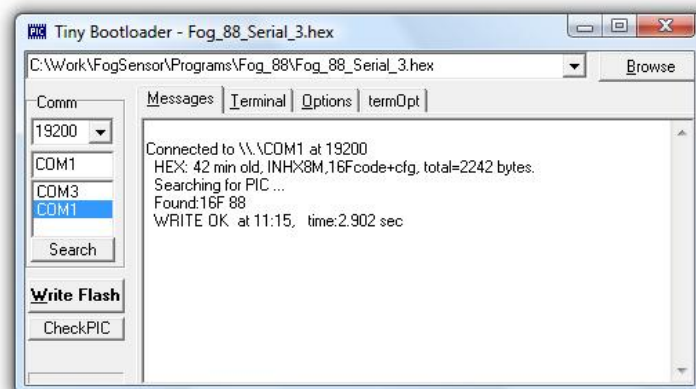
Obr. 5 Schémy zapojenia káblov pre vnútornú jednotku

C. Programové prostriedky senzora hmly

C.1 Vonkajšia jednotka

Vonkajšia jednotka obsahuje mikroprocesor, ktorý zberá namerané dáta a komunikuje s vnútornou jednotkou. Mikroprogramové vybavenie je dostupné na nasledovnej www stránke a je voľne šíriteľné : (<http://www.etc.ugal.ro/cchiculita/software/picbootloader.htm>) [1].

Bootloader je program, ktorý je súčasťou mikroprocesora a komunikuje s počítačom obyčajne cez sériové rozhranie. Bootloader prijme užívateľský program z počítača a zapíše ho do flash pamäte a bootloader program následne spustí. Bootloader je možné použiť iba s tými mikroprocesormi, ktoré môžu robiť zápis do svojej flash pamäte pomocou softvéru. Samotný Bootloader musí byť zapísaný do flash pamäte pomocou externého programovacieho zariadenia [3].



Obr. 6 Softvérové okno programu „Tiny Bootloader“

Návod pre Bootloader (obr. 6):

1. Spustíte Tiny Bootloader
2. Zvoľte HEX súbor na odsunutie
3. Zvoľte Comm 19200 kbps
4. Zapnite vonkajšiu časť a počkajte približne 3 sec.
5. Kliknite na tlačidlo “Write Flash”

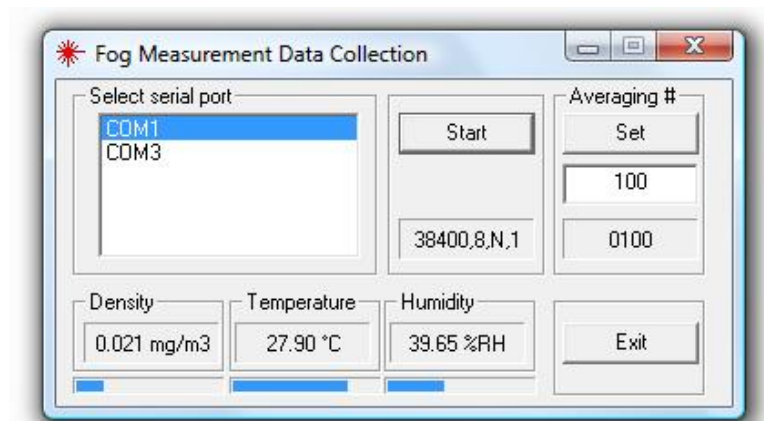
C.2 Vnútoraná jednotka

Na prijatie a uloženie nameraných dát z vonkajšej jednotky do počítača s operačným softvérom MS Windows XP alebo Vista je potrebný komunikačný port RS-232. Ak počítač má sériový port RS-232, tak sa v programe zvolí COM1, ak počítač tento sériový port nemá, tak sa použije redukcia z RS-232 na USB 2.0 a zvolí sa COM3 (obr. 7).

Požadované nastavenia MS Windows pre počítač PC zberajúci dáta sú:

- Nastavenie časovej zóny na GMT+0;
- Nastavenie 24 hodinového časového formátu;
- Stále pripojenie počítača k internetu a nastavenie automatickej časovej synchronizácie;
- Môžeme pridať do Startup menu "serial.exe com1", aby sa po reštartovaní programu začal autoštart.

Serial.exe je program na zberanie nameraných dát, ktorý zberá a ukladá dáta do súboru.



Obr. 7 Hlavné okno programu „Serial.exe“

Namerané údaje sú vzorkované vo vonkajšej jednotke každých 10 ms, následne aritmeticky spriemernené hodnotou "Averaging#" a posielané do vnútornej jednotky každú sekundu. Rozsah hodnôt pre hodnotu „Averaging#" je od 1 do 1000. V prípade uvedenom na obr. 7, je hodnota "Averaging#" nastavená na 100 [1].

III. EXPERIMENT

Program na zberanie dát uloží každú sekundu jednu vzorku do textového súboru. Na ilustráciu je zobrazených prvých 15 nameraných hodnôt z 3. decembra 2010 (336_Fog_Humi_Temp_Fri_Dec_03_2010.txt) (Tab. 1) [1]:

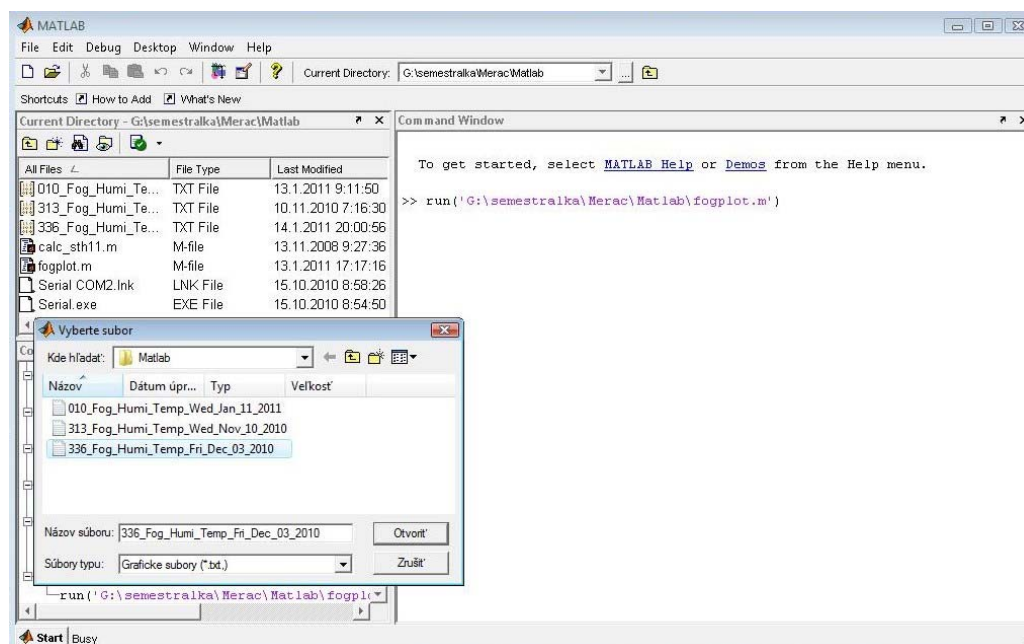
Tabuľka 1
TEXTOVÝ SÚBOR S NAMERANÝMI DÁTAMI

Density	Temperature	Humidity	Averaging	Relative Time
0113	4252	3127	0100	0
0113	4252	3127	0100	1
0113	4252	3127	0100	2
0115	4252	3127	0100	3
0115	4252	3128	0100	4
0117	4252	3127	0100	5
0113	4252	3127	0100	6
0113	4254	3127	0100	7
0114	4252	3127	0100	8
0113	4253	3126	0100	9
0111	4252	3126	0100	10
0113	4252	3126	0100	11
0113	4253	3125	0100	12
0114	4253	3126	0100	13
0114	4253	3125	0100	14
0115	4253	3125	0100	15

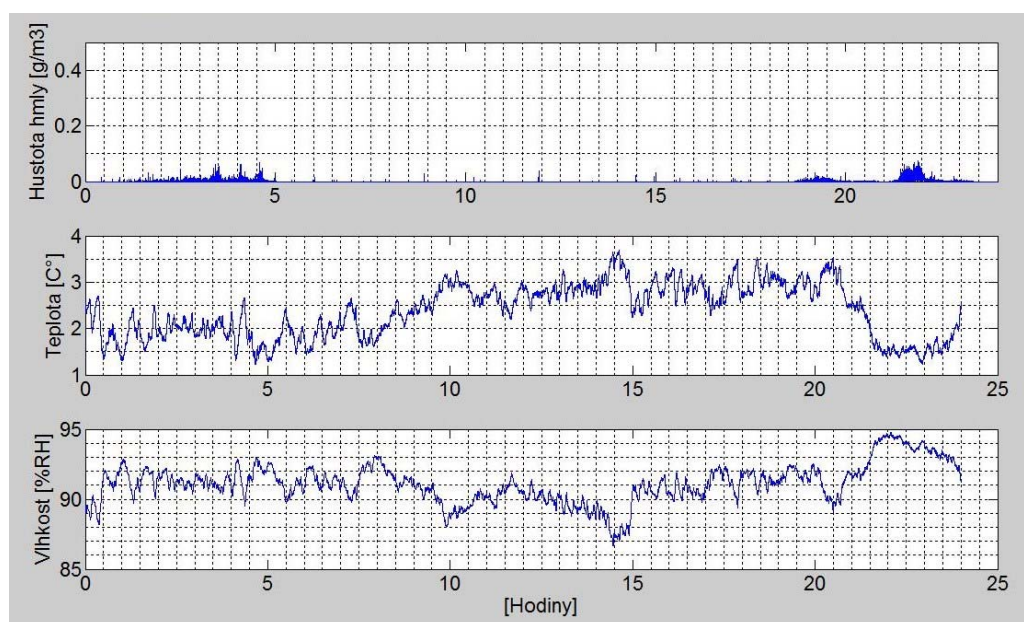
Uložené údaje možno ďalej spracovať v programe Matlab, ktorý z nameraných hodnôt vykreslí grafy. Hodnota “Relative Time” je počet sekúnd v aktuálnom dni, začína stále s nulou. “Averaging” predstavuje priemerňujúcu hodnotu. Aktuálny súbor je zatvorený a nový súbor je otvorený každý deň o 24:00. Ak je program prerušený a neskôr opäť spustený, tak zápis dát pokračuje v tom istom súbore, pokiaľ nedôjde k zmene dňa. Program sám pomenúva vytvorené textové súbory nasledovne:

Číslo aktuálneho dňa v roku, začína 0_Fog_Humi_Temp_deň_mesiac_rok.txt.
336_Fog_Humi_Temp_Fri_Dec_03_2010.txt

Na obr. 8 je znázornené prostredie programu vytvoreného v Matlab-e, ktorý pri spustení požaduje výber príslušného textového súboru. Tento textový súbor bol vytvorený programom Serial.exe. Po označení požadovaného textového súboru, v tomto prípade „336_Fog_Humi_Temp_Fri_Dec_03_2010.txt“, program následne vykreslí 3 priebehy. Sú to priebehy hustoty hmly, teploty a vlhkosti za celý príslušný deň (obr. 9).



Obr. 8 Prostredie v „Matlab-e“



Obr. 9 Vykreslené priebehy hustoty hmly, teploty a vlhkosti za celý príslušný deň v prostredí „Matlab“

IV. ZÁVER

Pre návrh kvalitných a efektívnych FSO systémov je potrebné podrobne poznať každý element zariadení a vidieť súvislosti medzi zmenami parametrov zariadení, zmenami vlastností prenosového kanála a výsledným efektom. Simulácia optických komunikácií voľným prostredím je preto nevyhnutnou pomôckou pri navrhovaní alebo experimentovaní s takýmito zariadeniami. Súčasťou takéhoto systému na KEMT FEI TUKE je aj senzor hmly, ktorý je schopný 24 hodín denne a 365 dní v roku monitorovať a zaznamenávať: hustotu hmly, teplotu a vlhkosť vonkajšieho prostredia, čo napomáha k efektívnemu a správne analyzovaniu spoľahlivosti FSO systémov na prenos ľubovoľných signálov.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] László CSURGAI-HORVÁTH, *Fog sensor manual*. Hungary, Budapest, 2010.
- [2] László CSURGAI-HORVÁTH, János BITÓ, *Fog attenuation on V band terrestrial radio and a low cost measurement setup*. Hungary, Budapest, pp 5-6.
- [3] <http://www.etc.ugal.ro/cchiculita/software/picbootloader.htm>.

Experimentálne pracovisko s optickým korelátorom

Tomáš HARASTHY, Jan TURÁN, Ľuboš OVSEŇÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

tomas.harasthy@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — V tejto práci je stručne popísaný hardvér Cambridge korelátor a rozmiestnenie častí v optickom systéme. Práca ďalej obsahuje stručný popis SLM a Cmos senzora, ich pripojenie k PC a správne nastavenie pre prácu s PC.

Kľúčové slová — optický korelátor, Cambridge korelátor, SLM, CMOS senzor

I. ÚVOD

V súčasnosti má dopyt o spracovanie dvojrozmerných polí dát, resp. obrazov stále stúpajúci charakter. Práve Cambridge korelátor (CC) a softvér Fourier Optic Experimenter (FOE) umožňuje používateľom prakticky preskúmať skvelé možnosti Fourierovej optiky a poskytuje možnosti ako by bolo vhodné riešiť rôzne problémy pri spracovaní údajov v komerčných, vedeckých a výskumných oblastiach.

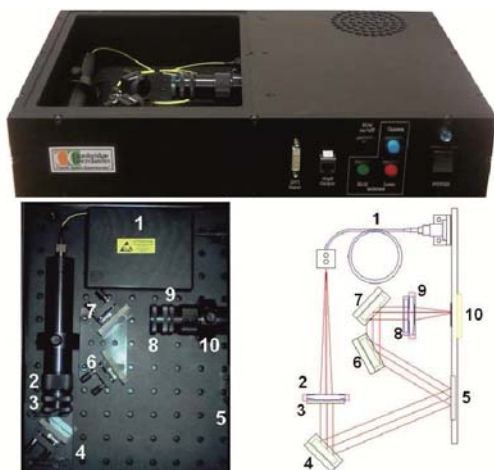
V tomto článku bude stručne popísaná konfigurácia FOE a Cambridge korelátor a jeho spúšťanie z externého počítača. V druhej kapitole bude kladený hlavný dôraz na hardvér Cambridge korelátor, jeho hlavné časti, ich funkciu a popis činnosti. Posledná kapitola sa bude zaoberať pripojením korelátor k PC.

II. OPIS SYSTÉMU

A. Optický systém Cambridge korelátor

Optický systém Cambridge korelátor (CC) je založený na výkonnom a kompaktnom jadre Fourier Transform Engine©, postavenom na princípe difrakcie Optickej Fourierovej Transformácie. Jeho revolučný „W“ dizajn umožňuje využiť plný potenciál jeho elektrooptických častí, a tak minimalizovať čas spracovania dát [1].

Hlavné časti CC je možné vidieť odstránením vrchnej časti, tak ako je to znázornené na obr. 1.



1. Laserová dióda a jednovidové vlákno
2. Kolimačné šošovky
3. Polarizátor
4. Rovinné zrkadlo
5. Priestorový modulátor svetla (Spatial Light Modulator - SLM)
6. Rovinné zrkadlo
7. Rovinné zrkadlo
8. Analyzátor
9. Transformačné šošovky
10. CMOS senzor(kamera)

Obr. 1 Jadro Cambridge korelátoru [1]

V prednej časti ohniskovej roviny systému šošoviek, ktoré zabezpečujú Fourierovu transformáciu, je umiestnený priestorový modulátor svetla (SLM) z tekutého kryštálu (5). SLM je pripojený k PC pomocou DVI rozhrania. V zadnej časti ohniskovej roviny je umiestnený CMOS senzor (10). Obraz zobrazený na (SLM) je osvetlený laserovým lúčom s malým výkonom, ktorý je do systému privedený z lasera použitím optického vlákna. Systém šošoviek tvorí optickú Fourierovu transformáciu v zadnej časti ohniskovej roviny a výstupná svetelná intenzita je zachytená kamerou.

Kolimačné aj transformačné šošovky sú umiestnené na presných pozíciách. Pre citlivejšie nastavenie je ich možné doladiť.

Tri rovinné zrkadlá sú rovnako umiestnené na presne určených pozíciách a rovnako je možné ich doladiť pomocou mikropohybu (mikroskrutiek).

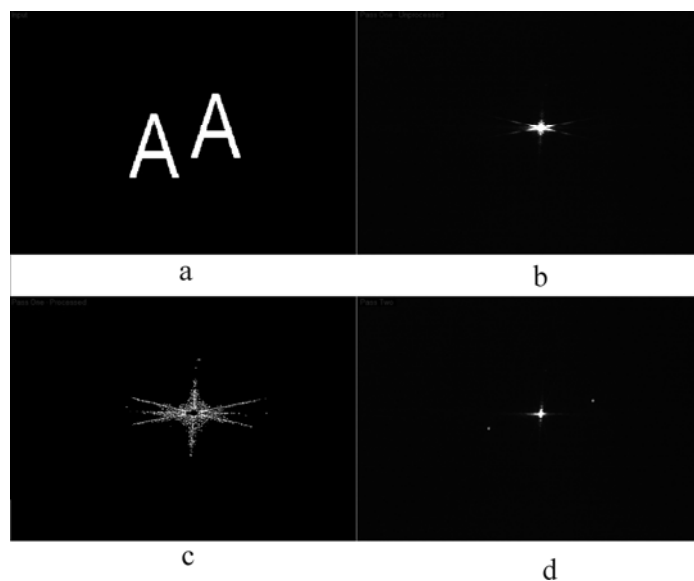
Polarizátor a analyzátor je taktiež možné nastavovať za účelom získania čo najlepších výsledkov. Otáčajú sa okolo svojej osi, a tak sa nastavuje požadovaná fáza a amplitúda optického výstupu [1].

B. Funkčnosť Cambridge korelátora

Existujú dva základne typy korelátorov a to korelátory s využitím Match filtra a korelátory založené na základe Join transformácie. Cambridge korelátor patrí do skupiny korelátorov s využitím Join transformácie. Sú označené ako JTC (Join Transform Corelator) korelátory.

Pri JTC korelátoroch sú porovnávané dva obrazy zlúčené do jedného. Obrazy sú vo fáze usporiadané vedľa seba a zobrazené na SLM (obr. 2a). SLM je potom osvetlený kolimačným koherentným, kruhovo polarizovaným svetlom z lasera, prechádzajúceho optickým vláknom a kolimačnými šošovkami. Fourierova transformácia vstupu je vytvorená na zadnej ohniskovej rovine transformujúcej šošovky. Intenzita rozdelenia tejto transformácie je označovaná ako Join Power Spectrum (JPS) a je zachytená CMOS senzorom (kamerou) obr. 2b. JPS je v PC alebo pomocou špeciálnej elektroniky binárne alebo prahovo transformované (obr. 2c).

V ďalšom kroku sa realizuje ďalšia Fourierovej transformácia. Vstupným obrazom pre túto transformáciu je novovzniknuté binárne JPS. Intenzita rozdelenia tejto druhej Fourierovej transformácie je tiež zobrazená na SLM. Intenzita rozdelenia už teraz obsahuje požadované svetlé body označujúce koreláciu, ktoré sa javia ako pár symetrických bodov (obr. 2d) [1,2,3].



Obr. 2 Join Transform korelácia [1]

III. PRIPOJENIE KORELÁTORA K PC

Cambridge korelátor obsahuje dva základné komponenty, ktoré je potrebné pripojiť k PC. Jedným z nich je priestorový modulátor svetla (SLM), ktorý zobrazuje vykonávané Fourierove transformácie. Ďalšou dôležitou časťou korelátora, ktorá nevyhnutne potrebujeme pripojenie k PC je kamera (CMOS senzor). Ostatné komponenty korelátora pracujú bez potreby pripojenia ku PC [2].

A. Priestorový modulátor svetla SLM

SLM použitý v Cambridge korelátore je na báze tekutého kryštálu (VAN - Vertically Aligned Nematic). Tento SLM je pomocou svojej vlastnej základovej dosky pripojený k PC cez DVI port. SLM aj so základovou doskou je zobrazený na obr. 3.

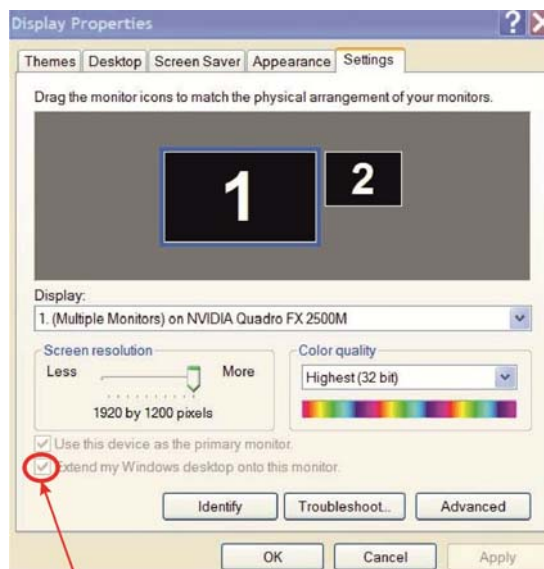


Obr. 3 SLM [2]

Základová doska má tiež vypínač vyvedený na prednú stranu korelátoru, ktorý umožňuje resetovanie a aj vypínanie SLM.

Po pripojení SLM k PC je potrebné vykonať niekoľko jednoduchých nastavení grafickej karty v PC.

Pred použitím je potrebné označiť a aktivovať DVI port, ako „druhý“ displej pre SLM. Vykoná sa to v menu pre nastavenie vlastností grafiky PC. Potom je treba vybrať „druhý“ displej označený číslom 2 a kliknúť na „Rozšíriť plochu Windows na tento monitor“ (obr. 4). Taktiež je nutné manuálne nastaviť rozlíšenie druhého displeja na rozlíšenie SLM, ktoré je 1024 x 768 obrazových prvkov (o.p.). V tab. 1 sú uvedené niektoré parametre SLM [2].



Rozšírenie plochy

Obr. 4 Nastavenie SLM v systéme „Windows“ [2]

Tabuľka 1
 PARAMETRE SLM [2]

Parameter	Hodnota
Rozlíšenie	1280 x 720 o.p.
Bitová hĺbka	8 bitov
Adresovanie vstupných rámcov	60 Hz (DVI adresovací port)
Veľkosť o.p.	9,5 mikronový odstup, 0,55“ uhlopriečka aktívnej plochy
Tekutý kryštál	Vertikálne usporiadaná nematická (VAN) bunka
Odozva	10ms ± 5ms
Obnovenie rámca	120 Hz

B. CMOS senzor - kamera

Ďalším zariadením, ktoré obsahuje Cambridge korelátor a ktoré musí byť nevyhnutne pripojené k PC je CMOS kamera. Je použitá Leutron PicSight kamera s Gigabitovým Ethernetovým pripojením. Najdôležitejšou podmienkou, aby bolo možné kameru používať, je nastavenie PC LAN (TCP/IP). Parametre TCP/IP musia byť nastavené nasledovne:

- IP adresa: 192.168.1.2
- Maska siete: 255.255.255.0

Ďalej je potrebné skontrolovať firewall nastavenia a povoliť komunikáciu PC so sieťou 192.168.1.0/255.255.255.0

V Tab. 2 sú popísané podrobné parametre kamery.

Tabuľka 2
PARAMETRE KAMERY [2]

Parameter	Hodnota
Rozhranie	Gigabit Ethernet
Maximálna snímková rýchlosť	Do 200 Hz
Rozlíšenie	VGA, 640x480 o.p.
Výstup	8/12 bit

IV. ZÁVER

Cieľom tejto práce bolo popísať hardvérových častí Cambridge korelátoru, a teda získanie základných znalostí pre prácu s ním. Po pripojení korelátoru k PC a nastavení podľa vyššie uvedených postupov je možné inštalovať riadiaci softvér dodaný spolu s korelátorom. Ďalším postupom práce bude testovanie korelátoru pomocou jednoduchých binárnych obrazov, a teda nastavenie polarizátora, analyzátoru a zrkadiel.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] Cambridge Correlators
Dostupné na internete: <<http://www.cambridgecorrelators.com>>, január 2011.
- [2] Fourier Optics Experimenter, User Guide, (1.0.46), 27/5/10.
- [3] J. W. Goodman, *Introduction to Fourier Optics*. Colorado: Roberts & Company, 2005.

Teoretické základy práce s optickým korelátorom

Tomáš HARASTHY, Ľuboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

tomáš.harasthy@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — V tejto práci je stručne popísaný princíp optickej Fourierovej transformácie a jej využitie v optických korelátoroch. Práca ďalej obsahuje stručný opis funkciu a prehľad optických korelátorov. V závere je experimentálne overenie Cambridge korelátoru.

Kľúčové slová — fourierová transformácia, korelácia, optický korelátor, Cambridge korelátor

I. ÚVOD

Rýchle spracovanie signálov, resp. informácií a ich analýza je téma, ktorá bude vždy aktuálna. Čoraz viac sa však vstupnými informáciami stávajú rôzne obrazové informácie. Využitie optických princípov spracovania umožňuje zvyšovať výkon ich spracovania z hľadiska objemu dát. Jeden zo systémov, ktorý využíva tento optický princíp, je optický korelátor. Vďaka optickej Fourierovej transformácii dokáže spracovať dvojrozmerný obraz v reálnom čase. V práci je popísaný princíp Optickej Fourierovej transformácie, jej využitie v Cambridge korelátoch a overenie korelátoru jednoduchým experimentom. Experiment spočíva v nájdení zhody medzi porovnávanými obrazmi.

II. FOURIEROVA TRANSFORMÁCIA V OPTIKE

A. Fourierova Transformácia

Fourierova transformácia (FT) patrí medzi základné stavebné bloky a prvky pri spracovaní signálu a jeho frekvenčných analýz. Jednoducho povedané, umožňuje rozdelenie signálu do jeho základných frekvenčných častí. Digitálne spracovanie signálu sa vďaka jednorozmernej Fourierovej transformácii a jej rýchlemu algoritmu stali jednou z najdôležitejších častí elektronických procesorov na spracovanie signálu.

$$F(u) = FT[f(x)] = \int_{-\infty}^{\infty} f(x) \exp(-i2\pi ux) dx \quad (1)$$

kde x je priestorová/časová premenná a u je priestorová/časová frekvenčná premenná [1].

Rýchla Fourierova transformácia (FFT), ktorá podstatne znižovala dobu spracovania signálu, však už nie je postačujúca pre dvojrozmerné polia dát. Toto dvojrozmerné spracovanie údajov použitím klasickej FFT je veľmi náročné na dobu spracovania, ktorá sa rapídne zvyšuje zvýšením rozlíšenia spomínaného dvojrozmerného poľa dát.

B. Teória optickej Fourierovej transformácie

Optická Fourierova transformácia (OFT) je analógiou pôvodnej FT a FFT, ktoré sa zvyčajne využívajú v softvérových aplikáciách. Využitím všeobecného zápisu FT, pre dvojrozmerné pole dát platí:

$$F(u, v) = FT[f(x, y)] = \iint_{-\infty}^{\infty} f(x, y) \exp[-i2\pi(ux + vy)] dx dy \quad (2)$$

kde x, y sú priestorové/časové premenné a u, v sú priestorové/časové frekvenčné premenné [1].

C. Základný princíp optickej Fourierovej transformácie

Podľa teórie Fourierovej transformácie je možné dvojrozmerný signál (teda aj vizuálny obraz) vyjadriť ako sumu postupností sínusoid.



Obr. 1 Sínusový vzor [4]

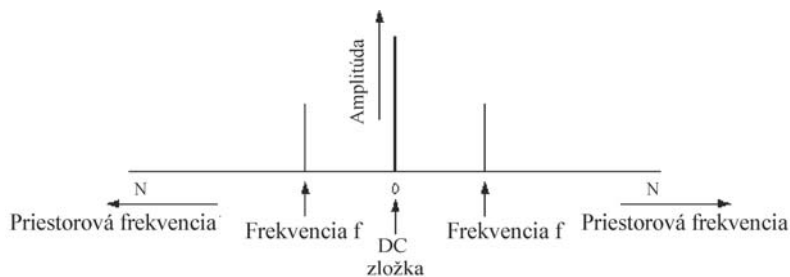
Na obr. 1 je znázornený príklad sínusového striedania jasu cez obraz, ktorý dokáže dekodovať informáciu o:

- Priestorovej frekvencii – t. j. o „frekvencii cez priestor“, pričom na obr. 2 sú porovnané nižšia a vyššia priestorová frekvencia.
- Hodnote – určuje kontrast alebo rozdiel medzi najtmavšími a najsvetlejšími špičkami v obraze, pričom táto hodnota môže byť kladná alebo záporná.
- Fáze – ktorá reprezentuje relatívny posun vln a ak nastáva nejaký posun oproti začiatku tak o koľko a či vpravo alebo vľavo [3,4].



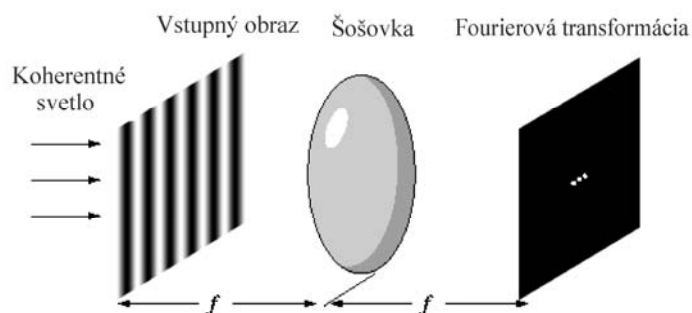
Obr. 2 Porovnanie priestorovej frekvencie [4]

OFT transformuje všetky priestorové frekvencie v danom obraze od nulových frekvencií až po najvyššie „Nyquistove frekvencie“, ktoré súvisia s rozlíšením obrazu. Na obr. 3 je pre jednoduchosť zakreslené spektrum sínusoid z obr. 1,



Obr. 3 Spektrum periodického signálu [4]

kde stredná špička predstavuje DC zložku a teda priemerný jas celého obrazu. Po oboch stranách sú dve špičky a to v bode, ktorý zodpovedá priestorovej frekvencii vstupnej sínusoidy. Výška špičky odpovedá amplitúde alebo kontrastu tohto sínusového signálu. Každý obraz musí mať kladnú hodnotu priemerného jasu a teda aj každý obraz má kladnú DC zložku [1,4].



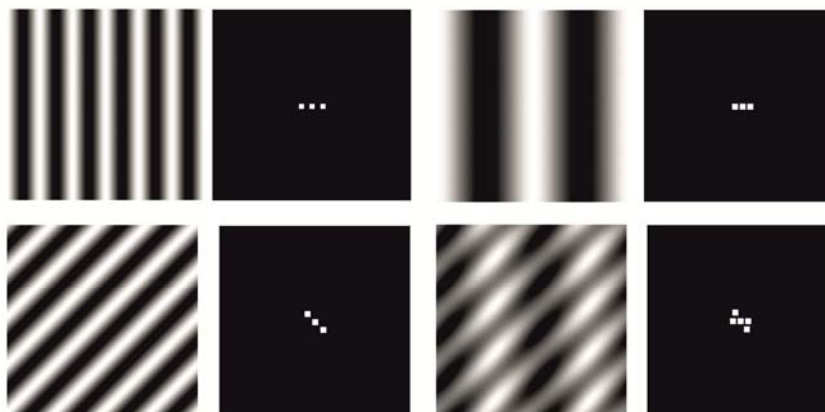
Obr. 4 Optická Fourierova transformácia [4]

D. Optická Fourierova transformácia

Hlavnou časťou zariadenia, ktoré by dokázalo produkovať OFT v reálnom čase sú šošovky. Na obr. 4 je znázornený princíp Optickej Fourierovej transformácie pomocou šošovky. V strede je umiestnená šošovka a v jej ohniskovej vzdialenosti sa nachádza vstupný obraz, ktorý je osvetlený koherentným svetlom. V druhom ohnisku šošovky je tienidlo, na ktorom sa zobrazí Fourierova transformácia vstupného obrazu.

Keďže pre jednoduchosť je za vstupný obraz vybraná sínusoida, preto jej spektrom budú iba tri jasové body. V strede sa nachádza jasová zložka, ktorá je obkolesená dvomi špičkami, ktoré budú meniť svoju vzdialenosť vzhľadom na stred podľa priestorovej frekvencie vstupnej sínusoidy.

Na obr. 5 sú znázornené mriežky sínusoid s rôznou priestorovou frekvenciou, prípadne inak modifikované a ich prislúchajúce Fourierove transformácie [1,3].



Obr. 5 Sínusoidy a ich transformácie [4]

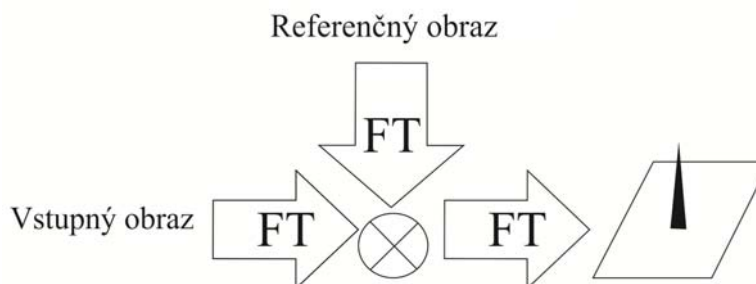
III. OPTICKÝ KORELÁTOR

Fourierove transformačné vlastnosti šošoviek sa stali základom pre funkciu optických korelátorov. Optický korelátor je zariadenie, ktoré porovnáva dva dvojrozmerné signály a výstupom je informácia o podobnosti daných obrazov, prípadne aj o relatívnej pozícii nájdennej podobnosti. Optický výstup pozostáva z tzv. korelačných špičiek, ktoré reprezentujú výraznú zhodu medzi porovnávanými obrazmi. Intenzita týchto bodov hovorí o miere podobnosti medzi danými obrazmi a umiestnenie týchto špičiek o relatívnom umiestnení nájdennej zhody.

Existujú dva základné typy optických korelátorov. Prvý typ je založený na prispôbenom filtri (Match filter) a druhým typom sú korelátor s využitím Spojenej (Joint) transformácie [2].

A. Korelátor s využitím prispôbeného filtra

Tento typ je založený na Vunderlugtovom filtri. Vyžíva samostatnú transformáciu vstupného a referenčného obrazu ako to je znázornené na obr. 7.

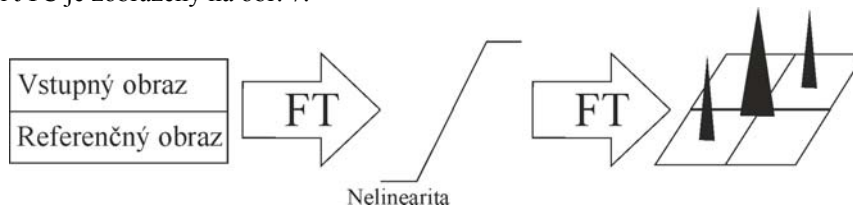


Obr. 6 Korelátor založený na Match filtri [2]

Transformácia referenčného obrazu je vykonaná off-line v elektronike využitím FFT za účelom vytvorenia filtra. Transformácia vstupného obrazu je už vykonaná opticky pomocou šošovky a potom opticky násobená s referenčným filtrom. Výsledok násobenia je následne opäť opticky Fourierovo transformovaný, výsledkom čoho je korelácia vstupného a referenčného obrazu [2].

B. Korelátor s využitím spojenej transformácie

Joint transform korelátor (JTC) sa od predošlého typu líši najmä spôsobom, akým sú definované vstupné zložky. Pri JTC je vstupný a referenčný obraz spojený do jedného a potom sú spolu transformované. Tu sa využíva už len OFT a teda nie je potrebný žiaden filter. Spôsob činnosti JTC je zobrazený na obr. 7.

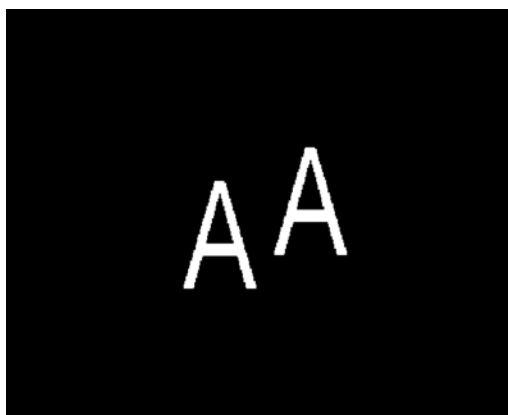


Obr. 7 Korelátor založený na Joint transformácii [2]

Nelinearita zobrazená na obr. 7 predstavuje šošovku kamery, ktorá zachytáva rozdelenie intenzity transformácie, tzv. „Joint Power Spectrum“ (JPS). JPS je taktiež transformované pomocou OFT. Výsledkom je súmerný pár korelačných špičiek pri nájdení zhody vo vstupných obrazoch. Výstup obsahuje taktiež jednosmernú DC zložku umiestnenú v strede medzi korelačnými špičkami [2].

IV. EXPERIMENTÁLNE OVERENIE CAMBRIDGE KORELÁTORA

Cambridge korelátor je optický korelátor založený na Joint transformácii. Vstupný a referenčný obraz je spojený do jedného obrazu a ďalej sa už považuje za jeden. Tento obraz vstupuje do optického systému a jeho výstupom je dvojica špičiek okolo DC zložky, ak bola nájdená zhoda. Na obr. 8 je znázornený vstup korelátoru [2].

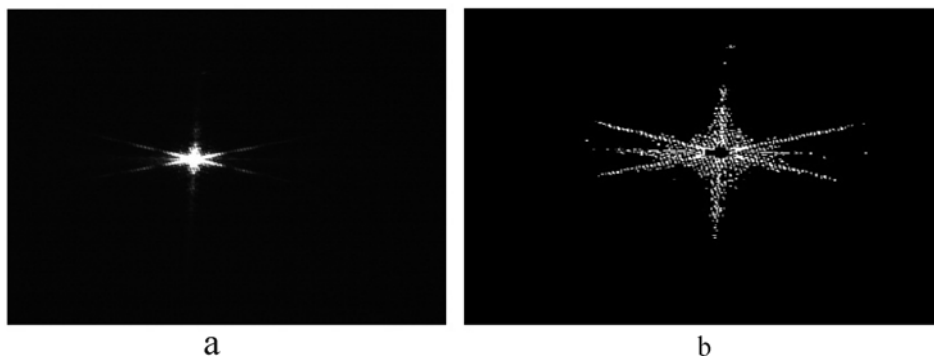


Obr. 8 Vstup korelátoru

Tento obraz je zobrazený na displeji v korelátore. Ako displej je použitý priestorový modulátor svetla (Spatial Light Modulator, SLM). Využitím OFT je tento obraz transformovaný.

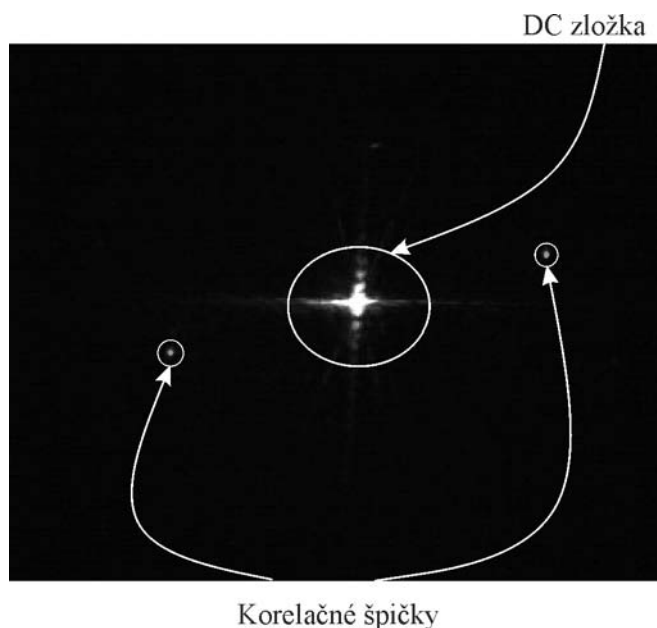
Jeho transformácia je zobrazená na obr. 9 [2]:

- a) bez predspracovania
- b) s predspracovaním (metóda najbližšieho suseda)



Obr. 9 Spektrum vstupného obrazu

Toto spektrum (Joint Power Spectrum) je ďalej zobrazené na spomínanom SLM a zosnímané kamerou. Následne je spektrum znovu transformované, výsledkom čoho je už dvojica špičiek, ktorá zodpovedá zhode nájdenej na vstupnom obraze. Výstup korelátoru je znázornený na obr. 10.



Obr. 10 Výstup korelátoru, detekovaná zhoda[2]

Veľkosť DC šumu a intenzita korelačných špičiek silno závisí od predspracovania vstupného obrazu (referenčný + porovnávaný). Pri binárnom obraze sú korelačné špičky jasné a zreteľné, no pri použití reálnych scén je výsledok silne závislý od zvoleného predspracovania vstupného obrazu.

V. ZÁVER

Výsledkom experimentu je získaná zhoda medzi porovnávanými obrazmi. Táto zhoda bola detekovaná pri porovnávaní dvoch identických obrazov na čiernom pozadí (písmeno A). Vzhľadom nato, že boli porovnávané rovnaké binárne obrazy, výsledné korelačné špičky sú zreteľné a jasné. Ďalším postupom bude aplikácia korelátoru do reálnych scén. Na to, aby bolo možné využívať reálne scény ako vstup korelátoru, bude potrebné určiť vhodné predspracovanie, ktoré by čo najlepšie odstránilo nepotrebné obrazové informácie (pozadie scény) a dalo do popredia informáciu o „potrebnom“ skúmanom objekte.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] J. W. Goodman, *Introduction to Fourier Optics*. Colorado: Roberts & Company, 2005.
- [2] Cambridge Correlators
Dostupné na internete: <<http://www.cambridgecorrelators.com>>, január 2011.
- [3] Fourier Optics Guide
Dostupné na internete: < <http://www.fourieroptics.org.uk/>>, január 2011.
- [4] L. Steven, *An Intuitive Explanation of Fourier Theory*.
Dostupné na internete :< <http://sharp.bu.edu/~slehar/fourier/fourier.html>>, január 2011.

Experimenty s transformačným systémom na rozpoznávanie dopravných značiek

Tomáš HARASTHY, Ján TURÁN, Ľuboš OVSENÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

tomáš.harasthy@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — V tejto práci je stručne opísaná architektúra videoasistenčného systému vodiča pre invariantné rozpoznávanie dopravných značiek založené na Trace transformácii. Článok stručne popisuje program realizujúci rozpoznávanie značiek a popisuje experimenty realizované týmto systémom a ich výsledky.

Kľúčové slová — videoasistenčný systém vodiča, Trace transformácia, farebná segmentácia, tvarová detekcia

I. ÚVOD

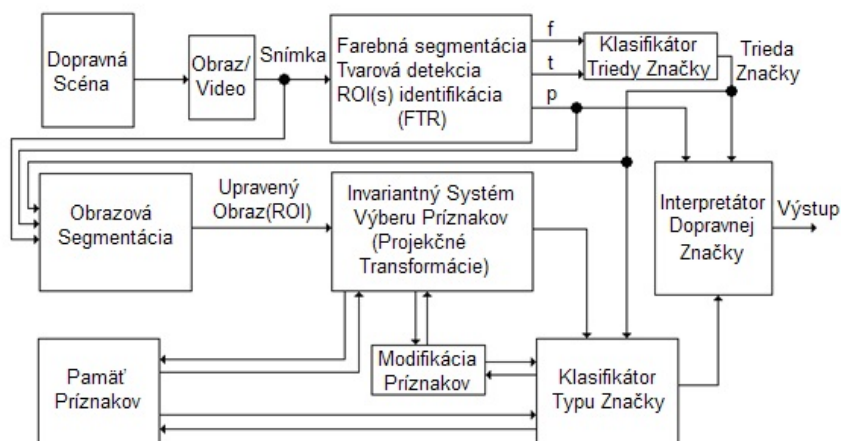
Elektronikou nabité súčasné automobily umožňujú zmenšenie fyzickej a psychickej záťaže vodiča, a tak zvyšujú nielen komfort ale hlavne bezpečnosť jazdy. Jednou z hlavných príčin dopravných nehôd je nepozornosť vodiča alebo nerešpektovanie dopravných značiek. Mnohokrát sa stáva, že vodič je natoľko zaneprázdnený situáciou pred vozidlom, že si nevšimne alebo prehliadne dopravnú značku. Preto sa čoraz viac do popredia dostávajú systémy, ktoré by dokázali rozpoznať dopravné značky a bezprostredne na to zobrazíť získanú informáciu v zornom poli vodiča, pričom by vodič nemusel oči ani na okamih odtrhnúť od vozovky a dopravnej situácie pred automobilom.

Tento článok opisuje videoasistenčný systém vodiča schopný rozpoznať dopravnú značku s využitím Trace transformácie, zobrazíť ju, určiť jej triedu a nebezpečnosť situácie, a tak ju dať do povedomia vodiča.

II. POPIS TEÓRIE SYSTÉMU

Realizovaný videoasistenčný systém vodiča (obr. 1) využíva detailné preskúmanie vizuálnej informácie pomocou farebnej kamery, ktorej výhodou oproti čiernej kamere je, že okrem detekcie tvaru značky môžeme využiť aj jej farebnú segmentáciu [2].

A. Architektúra systému



Obr. 1 Architektúra systému na rozpoznávanie dopravných značiek [2]

Systém je možné rozdeliť na dva podsystémy, tzv. „rýchly“ a „pomalý“. Rýchly podsystém pozostáva iba z bloku „FTR“ a z bloku detekcie triedy dopravnej značky „Klasifikátor Triedy Značky“. Pomalý obsahuje bloky „Obrazovej Segmentácie“, „Invariantného Systému Výberu Príznakov“, „Pamäte Príznakov“, „Modifikácie Príznakov“ a bloku určenia typu dopravnej značky „Klasifikátora Typu Značky“ [2].

B. Funkcia systému

Z dopravnej scény zachytenej farebnou kamerou je vytvorený dátový tok, z ktorého sa využívajú statické snímky pre nasledujúce bloky systému. Statické snímky vstupujú do FTR bloku, kde sú vybraté tri základné príznaky, t.j. farba, tvar a pozícia oblasti záujmu (ROI). Zo snímky sú vyextrahované tri farby a to červená, modrá a žltá. Výstupom pre každú farbu je jej bitmapa, v ktorej sa budú hľadať možné tvary značiek. Ak sa v ROI nachádza niektorý z tvarov dopravnej značky a vieme jej farbu, potom „Klasifikátor Triedy Značky“ určí triedu značku (Výstražné značky (červené trojuholníky); Koniec zákazov (čierno-biele kruhy); Informačné značky (modro/zelené obdĺžniky); Príkazové značky (modré kruhy); Zákazové značky (červené kruhy); Značka hlavnej cesty (žltý kosoštvorec); Značka STOP (červený vyplnený šesťuholník); Doplnkové značky (obdĺžniky s čierno-bielymi obrazmi); Značka daj prednosť v jazde (obrátený červený trojuholník)) [1,3].

Trieda dopravnej značky spoločne s pozíciou dopravnej značky na obraze vstupuje do bloku „Obrazovej Segmentácie“, kde sú nájdené a určené všetky ROI a tieto potom vstupujú do bloku „Invariantného Systému Výberu Príznakov“, pričom invariantné príznaky sú také príznaky, ktoré majú rovnaké hodnoty v prípadoch, keď je napríklad objekt posunutý, otočený, zmenšený alebo má nejakú podobnú modifikáciu. Tento blok využíva Trace transformáciu a v spolupráci s blokmi „Pamäť Príznakov“ a „Modifikácia Príznakov“ poskytuje informácie pre „Klasifikátor Typu Značky“. Spolu s „Klasifikátorom Triedy Značky“ je možné rozpoznať dopravnú značku na vstupnom obraze a interpretovať ju v bloku „Interpretátor Dopravnej Značky“ [3].

III. EXPERIMENTÁLNE OVERENIE SYSTÉMU

A. Experimenty s FTR blokom

Ako bolo spomenuté vyššie, spracovanie statického obrazu v FTR bloku spočíva v extrahovaní základných príznakov. Najprv sú zo snímky vyextrahované červená, modrá a žltá farba. Každá farba tak vytvára jednu bitmapu. V týchto bitmapách sa budú hľadať možné tvary dopravných značiek. Ak je v niektorej bitmape detekovaný tvar dopravnej značky, vieme určiť do ktorej kategórie značka patrí. Výstupom FTR bloku sú teda farba (f), tvar (t), a pozícia (p). Testovaný program používa na hľadanie značiek RGB aj HSV farebný model a podľa toho sa aj líšia operácie vykonávané v FTR bloku [2].

Napríklad v tab. 1 sú uvedené výsledky detekcie značky: „Daj prednosť v jazde“ pre RGB aj HSV farebný priestor. Podobne v tab. 2 pre značku: „Hlavná cesta“ a v tab. 3 je vyjadrená celková úspešnosť detekcie ROI pre testované dopravné značky. Bolo testovaných 14 dopravných scén, na obr. 2 je uvedený príklad dopravnej scény so značkou „Daj prednosť v jazde“ (a), a „Hlavná cesta“ (b), ktoré boli použité pri testovaní systému.

Tabuľka 1
 CELKOVÁ ÚSPEŠNOSŤ DETEKcie ROI PRE ZNAČKU „DAJ PREDNOSŤ V JAZDE“

Dopravná scéna	Počet značiek v obrázku	Nájdené ROI v HSV	Nájdené ROI v RGB
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	0	0
14	1	0	0
Úspešnosť (%)		85,7	85,7

Tabuľka 2
 CELKOVÁ ÚSPEŠNOSŤ DETEKcie ROI PRE ZNAČKU „HLAVNÁ CESTA“

Dopravná scéna	Počet značiek v obrázku	Nájdené ROI v HSV	Nájdené ROI v RGB
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	0
9	1	1	0
10	1	0	0
Úspešnosť (%)		90	70

Tabuľka 3
 CELKOVÁ ÚSPEŠNOSŤ DETEKcie ROI PRE JEDNOTLIVÉ ZNAČKY

Značka	HSV (%)	RGB (%)
Daj prednosť v jazde	85,75	85,71
Maximálny rýchlosť 40	100	100
Kruhový objazd	100	100
Pozor kruhový objazd	100	100
Pozor deti	100	76,92
Zákaz stáť	100	100
Zákaz v oboch smeroch	100	93,75
Zákaz vjazdu	100	100
Zákaz zastavenia	91,67	83,33
Prechod pre chodcov	80	80
Hlavná cesta	90	70
Prikázaný smer	92,31	92,31
Celková úspešnosť (%)	95	90,2



Obr. 2 Dopravné scény

B. Experimentálne overenie celého systému

Celý systém pre rozpoznávanie značiek bol otestovaný trinástimi najčastejšie sa vyskytujúcimi dopravnými značkami. Aby sme získali čo najreálnejšie výsledky, testované obrazy boli získané za rôznych osvetľujúcich podmienok, pri rôznych otočeniach značiek a pri rôznej vzdialenosti. Na obr. 3 sú zobrazené dopravné značky, ktoré boli vybraté pre testovanie rozpoznávacieho systému.

V tab. 4 sú uvedené výsledky testov, z ktorých vyplýva, že úspešnejšie je vyhľadávanie pomocou HSV farebného modelu. Celková úspešnosť pri použití HSV modelu je 93,6 %, oproti 88,9 % úspešnosti pri použití RGB modelu.



Obr. 3 Testované dopravné značky

Tabuľka 4
 CELKOVÁ ÚSPEŠNOSŤ DETEKCIE ROI PRE ZNAČKU PRE JEDNOTLIVÉ ZNAČKY

Značky	Počet dopravných značiek	Rozpoznané v HSV	Rozpoznané v RGB	Rozpoznané v HSV (%)	Rozpoznané v RGB (%)
Žlté	23	22	14	95,65	85,71
Modré	98	85	85	86,73	86,63
Červené	223	215	207	96,41	92,83
Celkovo	344	322	306	93,6	88,9

Z tab. 4 je zrejmé, že systém mal najväčšie problémy pri žltej farbe v RGB farebnom priestore. Naopak najlepšie výsledky boli pri HSV farebnom priestore pri značkách žltej farby. Úplne rozpoznanie značiek nenastalo ani v jednom prípade. Vyskytli sa tri najväčšie problémy, ktoré znemožňovali detekciu a rozpoznanie dopravných značiek. Prvý problém, ktorý platí najmä pre RGB priestor, je vysoká citlivosť na jasovú zložku obrazu. Ďalším bolo zasahovanie iných objektov do obrazu. Tento problém bol citlivý pre obidve farebné priestory. Dopravné značky, ktoré boli prekryté napríklad konármi stromov neboli rozpoznané. Tretím problémom boli rôzne vplyvy externého žiarenia, napr. slnka.

IV. ZÁVER

V tomto príspevku je v skratke popísaný nami navrhnutý a zrealizovaný systém pre rozpoznanie dopravných značiek. Systém bol experimentálne overený na rôznych typoch dopravných značiek. Z experimentov vyplýva, že systém dokáže rozpoznať väčšinu dopravných značiek pri reálnych dopravných scénach. Ďalšie pokračovanie v tomto výskume bude smerovať do úpravy vstupnej informácie, t.j. za vstup použiť videosekvenciu prípadne reálne snímané video. Ďalším vylepšením bude využitie optických procesorov (optického korelátor) pre zníženie času potrebného pre rozpoznanie značiek.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, T. Koehler, „A System for Traffic Sign Detection, Tracking and Recognition Using Color, Shape, and Motion Information“, *IEE Intelligent Vehicles Symposium*, June 2005, pp. 255–260.
- [2] M. Fífik, J. Turán, L. Ovseník, „Experiments with a Transform based Traffic Sign Recognition System“, *IWSSIP 2010*, Rio de Janeiro, Brazil, June 17-19, 2010, pp. 227-230.
- [3] M. Fífik, „Traffic Sign Detection with use of Invariant Transform“, *Carpathian Journal of Electrical Engineering*, Vol.1, No.1, 2008, pp. 19-21.

Projektovanie OVKS s využitím programového balíka „OptSim“

Luboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — V tejto práci je stručne opísané intuitívne modelovacie a simulačné prostredie programovacieho balíka „OptSim“, podporujúce návrh a určenie výkonnosti fyzickej vrstvy (prenosovej časti) optických komunikačných systémov. Obsiahla knižnica modelov komponentov a simulačné algoritmy poskytujú najlepší kompromis medzi presnosťou a rýchlosťou pre skrátenie času potrebného pre získanie správnych výsledkov návrhu. Dalším efektom je zníženie nákladov na realizáciu komplexných projektov a návrhov OVKS. Software je možné použiť nielen pre optimalizáciu optických trás (disperzné mapy, nastavenia výkonu, zisku, polohy zosilňovačov,...), ale aj pre analýzu vplyvu lineárnych a nelineárnych javov na výkonnosť systému (PMD, SPM, FWM, ...).

Kľúčové slová — projektovanie OVKS, programový balík OptSim, optické vláknové komunikačné systémy, OVKS

I. ÚVOD

Programový balík „OptSim“ je pokročilý simulačný program, určený pre profesionálny výskum v oblasti rôznych typov multiplexov (WDM, DWDM, TDM), káblovej televízie (CATV), optických sietí LAN, paralelných optických zberníc a ďalších optických systémov v oblasti telekomunikácií, ako aj dátových komunikácií pracujúcich s jednovidovými optickými vláknami. Tento program môže byť použitý pre návrh optických komunikačných systémov a ich následnú simuláciu pre zistenie výkonov, ktoré sa menia v závislosti od parametrov použitých komponentov.

„OptSim“ je navrhnutý tak, aby skombinoval čo najväčšiu presnosť a schopnosť návrhu, s jednoduchosťou použitia na obidvoch operačných platformách, t.j. Windows a Unix. „OptSim“ obsahuje rozsiahlu knižnicu komponentov (Obr. 1 ❶), v ktorej sa nachádzajú najčastejšie používané prvky pri návrhu optoelektronických systémov. Táto knižnica je neustále rozširovaná vďaka prebiehajúcemu výskumu a spolupráci so svetovo uznávanými špecialistami na simuláciu a analýzu optických vláknových komunikačných systémov (OVKS).

Okrem rozsiahlej knižnice komponentov, „OptSim“ zahŕňa aj sofistikované, ale zároveň ľahko použiteľné grafické rozhranie (GUI – grafický užívateľský interfejs) na efektívnu analýzu výsledkov simulácie. GUI ponúka hierarchické, objektovo orientované CAD prostredie pre návrh systémových zapojení a schém.

II. POPIS PROGRAMOVACIEHO BALÍKA „OPTSIM“

A. Simulačné módy

Programovací balík „OptSim“ podporuje dva vzájomne sa dopĺňajúce prístupy simulácie, tzv. simulačné módy.

Mód blokovej simulácie vykonáva simulácie, v ktorých je dátový signál odovzdávaný medzi jednotlivými komponentmi, predstavujúcimi simulovaný čas v jednom bloku dát. Výhodou tohto prístupu je, že komponenty a algoritmy môžu ľahko pracovať s celým signálom, transformovať ho vzad a vpred medzi časovou a frekvenčnou oblasťou, a tak pôsobiť na údaje v oblastiach výhodných pre simuláciu algoritmu.

Mód vzorkovacej simulácie vykonáva simulácie, v ktorých je dátový signál odovzdávaný medzi komponentmi predstavujúcimi jednu vzorku alebo časový krok v čase. Na rozdiel od módu blokovej simulácie, kde jeden komponent môže odovzdávať dáta na iný komponent len raz v priebehu simulácie, ktorá pokrýva celý simulovaný čas v jednom bloku dát, v móde vzorkovacej simulácie prejde nová vzorka dát do iného komponentu v každom časovom kroku

simulácie. Jednou z výhod tohto prístupu je skutočnosť, že simulácia sa môže vykonávať v neobmedzenom simulačnom čase a môže zahŕňať neobmedzené dĺžky prenášaných sekvencií. V tomto režime sa spracovanie signálu vykonáva úplne v časovej oblasti. Režim vzorkovacej simulácie tiež ponúka dva typy simulácií: *Metódu zobrazenia spektra*, ktorá vykonáva rýchle simulácie optického spektra, optického zosilňovania, ako aj optického filtrovania a *Simuláciu s premenlivou šírkou pásma*, ktorá vykonáva simuláciu priebehu pri úplnej časovej základni s možnosťou zahrnúť všetky efekty s alebo bez šumu v lineárnej alebo nelineárnej simulácii.

Cez flexibilné užívateľské možnosti, ktoré ponúka „OptSim“, môžu užívatelia vytvárať modely, ktoré fungujú buď ako dátové signály vzoriek v čase, alebo ako veľké bloky v časovej alebo frekvenčnej oblasti, pre mód vzorkovej, resp. mód blokovej simulácie.

B. Výsledky analýzy a ich následné spracovanie

Výsledky simulácií, ktoré sú produkované programom „OptSim“ zahŕňajú mapovanie priebehov signálov a diagramy oka na akomkoľvek mieste v rámci modelovaného OVKS. Sú dostupné aj ďalšie výsledky simulácie, ako je spektrum signálov, napájanie a mapy disperzie, atď. Výsledky simulácií môžu byť taktiež vykreslené vedľa seba, čo umožňuje ich porovnávanie so zosnímanými hodnotami parametrov. „OptSim“ vykonáva analýzy vplyvu šumu, presluchov, skreslenia a zmeny parametrov použitých komponentov pomocou metódy Monte Carlo. Na následné spracovanie výsledkov simulácií sú k dispozícii analytické nástroje WinPlot™ a RPlot™ (vykresľovanie výsledkov simulácie), rovnako ako VirLab™ (spracovanie výsledkov simulácie). Tieto nástroje uľahčujú interaktívnu prácu s výsledkami simulácie a umožňujú prezentovať výsledky analýzy požadovaným spôsobom.

C. Rozšírenie programového balíka „OptSim“

„OptSim“ taktiež poskytuje užitočné funkcie pre vytváranie užívateľských modelov v oboch simulačných módoch (blokovom aj vzorkovacom). Užívateľské modely môžu byť napísané v programovacích jazykoch C++ alebo FORTRAN, a potom nahraté do simulačného nástroja využívajúceho rozhranie pre programovanie aplikácií (API), ktoré poskytuje „OptSim“.

Okrem toho podporuje „OptSim“ aj rozhranie s programom MATLAB, ktoré umožňuje používateľom zapisovať svoje užívateľské modely v tomto prostredí pre plnú spoluprácu s programom „OptSim“, a tiež možnosť využitia MATLABu pre spracovanie výsledkov simulácie. „OptSim“ poskytuje ďalšie rôzne rozhrania na pripojenie rôznych prístrojov, ako sú napr. optický spektrálny analyzátor, či optický vektorový analyzátor.

Má aj SPICE rozhranie, vrátane SPICE modelu generátora, čím vyplňuje medzeru medzi elektronickými obvodmi a simuláciami optických komunikačných systémov. Táto unikátna funkcia umožňuje návrh obvodov pomocou lasera a modelovanie systémov jednoduchšie a presnejšie než kedykoľvek predtým.

D. GUI – CAD prostredie

Ako už bolo spomínané, „OptSim“ poskytuje hierarchicky, objektovo orientovanú topológiu schém pre jednoduchú definíciu systému. Optické prvky komunikačného systému, ako je pseudonáhodný binárny generátor sekvencií, modulátor, multiplexery, optické zosilňovače a útlmové filtre, môžu byť ľahko vybrané z palety panela nástrojov (Obr. 1 ②) alebo stromu, graficky pridané do schémy a prepojené pomocou myši na tvorbu požadovanej topológie.

Každý komponent reprezentovaný ikonou v schéme má svoju vlastnú sadu parametrov, ktoré sú prístupné jednoduchým kliknutím pravého tlačidla myši. Parametre môžu obsahovať číselné hodnoty (napr. laser vlnovú dĺžku, optické vlákno priemer), alebo výber medzi rôznymi nastaveniami (ako je napr. typ filtra).

Schémy je možné prispôbiť si zapnutím alebo vypnutím mriežky (Obr. 1 ③), zobrazenie alebo skrytie mena komponentu pri jeho ikone, zobrazenie vybraných parametrov komponentov na schéme, či zmenu rozmiestnenia portov na ikone.

Taktiež sú podporované aritmetické výrazy a užívateľské premenné. Každý numerický parameter komponentu v schéme môže byť špecifikovaný ako aritmetický výraz zahŕňajúci premenné nadefinované užívateľom.

Parametre komponentu môžu byť testované ešte pred spustením simulácie celého systému, čo umožňuje jednoduchšie meranie danej vlastnosti skúmaného komponentu.

Komponenty môžu byť zoskupené a použité na vytvorenie združených komponentov (Compound Components) tiež známych ako superbloky, pre opakované použitie alebo zníženie „neporiadku“ v schéme, rovnako ako pre použitie v mnohých iných schémach.

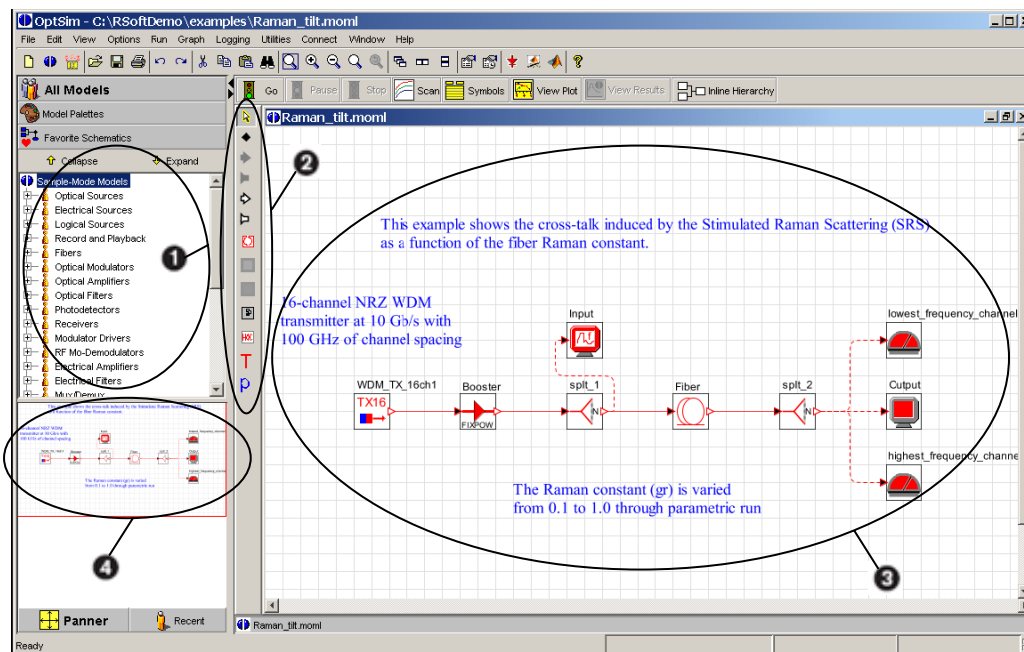
Nastavenia parametrov komponentov možno ukladať do súborov alebo načítať zo súborov priamo z editovacieho okna komponentov. To podporuje vytváranie a používanie knižníc komponentov, kde rôzne súbory nastavení parametrov komponentov môžu reprezentovať rôzne fyzické komponenty.

Nový prispôbitelný prieskumník umožňuje užívateľom vytvoriť si ich vlastné palety modelov, použiť stromové zobrazenie pre výber nedávno použitých modelov, vytváranie a

využívanie knižníc združených komponentov, ako aj definovať zbierku „obľúbených“ schém pre ľahký prístup.

Schému možno zväčšovať a zmenšovať pomocou zoomu neobmedzene a stopa je uvedená v ľavom dolnom rohu (Obr. 1 4) pre ľahšiu identifikáciu a kontrolu, ktorá časť celej schémy je viditeľná v editačnom okne.

Animačný režim simulácie je k dispozícii spolu so stavom simulácie (simulation status) a oknom pre uloženie výsledkov simulácie



Obr. 1 GUI programového balíka „OptSim“

E. Simulácia a analýza

Algoritmus plánovania optimalizuje účinnosť simulácie na simuláciu iba tých komponentov, ktoré počas skenovania zmenili hodnotu svojej premennej a ak je to možné, použije už vypočítané hodnoty. Táto funkcia umožňuje časovo efektívnu analýzu parametrov, ktoré používajú už naskenované premenné v celom systéme a umožňuje parametrické štúdie a návrh optimalizácie.

Taktiež sú podporované štatistické rozdiely parametrov komponentov, pričom číselné parametre komponentov môžu mať svoje rozdelenia pravdepodobnosti a smerodajná odchýlka môže byť samostatne definovaná. Spustenie viacnásobnej simulácie môže byť použité na určenie rozsahu očakávaných výsledkov zapojenia. Tieto informácie môžu byť použité pre projektantov na návrh parametrov, vykonanie analýzy starnutia spojov, štúdia vplyvu prostredia na výkon zariadenia a jeho spoľahlivosť. Tiež sa to používa na štúdium paralelných optických spojov, kde zmeny parametrov komponentov medzi kanálmi môžu viesť k zníženiu výkonu zariadenia.

Opakovanie slučiek umožňuje častiam topológie byť upravené tak, že opakovanie n-krát sa v simulácii javí, ako keby tam bolo n komponentov reprezentovaných ikonou a v rámci slučiek sú tieto ikony poprepájané. Opakovacie slučky môžu byť vnorené a môžu mať viac vstupných a výstupných signálov. Tie sú obzvlášť užitočné pre definovanie rozpätia vlákien a zosilňovačov v systémoch s veľkou dĺžkou.

Externe namerané alebo vypočítané priebehy signálov môžu byť použité na riadenie simulácií v programovom balíku „OptSim“. To poskytuje flexibilitu pri konfigurácii simulácií a umožňuje programu „OptSim“ byť v spojení s experimentálnymi výsledkami, ako aj s ďalšími programami na simuláciu a analýzu. Signály môžu byť uložené v priebehu simulácie a môžu byť použité ako riadiace signály pre budúce simulácie, alebo načítané zo súborov na riadenie aktuálnej simulácie.

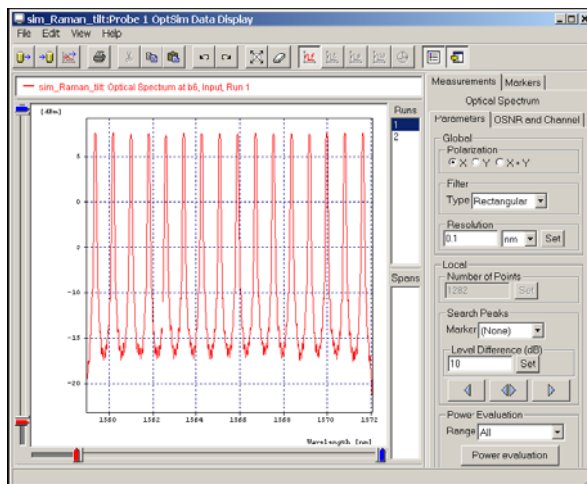
Užívatelia tohto programového balíka môžu vytvárať vlastné modely komponentov optoelektronických prvkov v systémoch GratingMOD™, LaserMOD™, BeamPROP™, alebo iných a následne ich môžu začleniť do prostredia „OptSim“, pričom tieto užívateľské modely sú rovnako sofistikované a efektívne, ako vstavané modely. Táto mnoho úrovňová simulácia poskytuje najlepšiu presnosť analýzy pri kombinácii vysokej kvality a zložitosti návrhu.

„OptSim“ umožňuje integráciu programového balíka „ModeSYS“, ktorý je v priemysle jediným mnovovidovým simulačným systémom pre analýzu mnohovidových OVKS, ktorý pri simulácii berie do úvahy ako priestorové, tak aj časové účinky.

III. EXPERIMENTÁLNE OVERENIE SYSTÉMU

A. Mód vzorkovacej simulácie - Účinky stimulovaného Ramanovho rozptylu na presluchy medzi kanálmi vo WDM systémoch

Zapojenie pre analyzovanie účinkov stimulovaného Ramanovho rozptylu (SRS) na presluchy medzi kanálmi vo WDM systémoch je na schéme Obr. 1 [3]. 16-kanálový systém WDM je pripojený na 20 km dlhé jednovidové optické vlákno (OV). Všetky kanály sú NRZ modulované pri prenosovej rýchlosti 10 Gb/s. Odstup kanálov je 100 GHz s plnou šírkou pásma pre všetky kanály 12 nm. Celkový príkon do OV je 20 dBm (pri 8 dBm/kanál). Optické spektrum na vstupe OV je znázornené na Obr. 2.



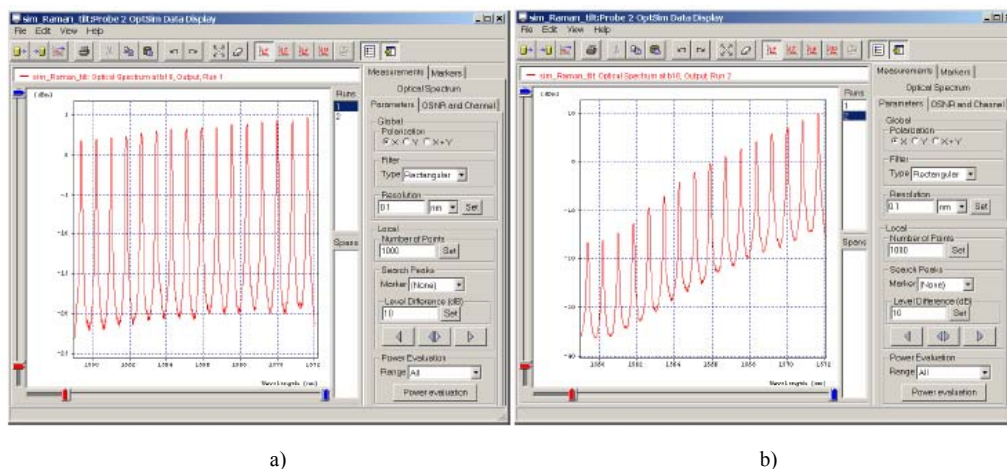
Obr. 2 Optické spektrum na vstupe optického vlákna

Štandardná maximálna hodnota Ramanovho zosilnenia g_R je asi 1×10^{-13} m/W pri vlnovej dĺžke 1 μ m a táto hodnota s mení lineárne s frekvenciou čerpania, takže pre 1550 nm to bude asi 6×10^{-13} m/W.

Aby sme videli vplyv presluchov pri rôznych hodnotách Ramanovho zosilnenia g_R , bol zavedený koeficient navýšenia $Const$, pre ktorý platí *standard*

$$(g_r)_{new} = (Const) \times (g_r)_{standard} \quad (1)$$

a študovali sme prípady s veľkosťou konštánt $Const = 0,1$ a $Const = 1,0$.



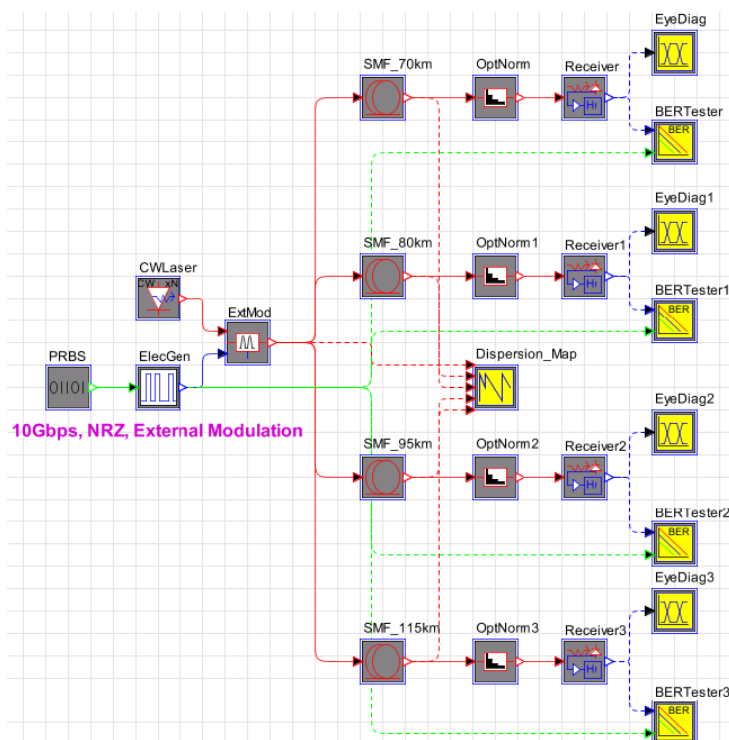
Obr. 3 Optické spektrum na výstupe optického vlákna pre konštanty: a) 0,1; b) 1,0

Optické výkonové spektrum prijímaného signálu (Obr. 3) ukazuje spektrálny posun (naklápanie) vyvolané SRS. Naklápanie sa zvyšuje s rastúcou konštantou (t.j. s rastom Ramanovho zisku). Očakávali sme, že s 10 krát vyšším koeficient zosilnenia bude aj 10 krát väčší náklon, no ako môžeme vidieť z výsledkov simulácie, náklon ide z ~ 3 dB pre $Const = 0,1$ (Obr. 3a) na ~ 28 dB pre $Const = 1,0$ (Obr. 3b).

Na záver sme dokázali, že vplyv SRS na presluchy medzi WDM kanálmi spôsobuje sila - nelineárny rozptyl medzi vyššími a nižšími frekvenciami kanálov, čo spôsobuje náklon optického spektra na výstupe optického vlákna.

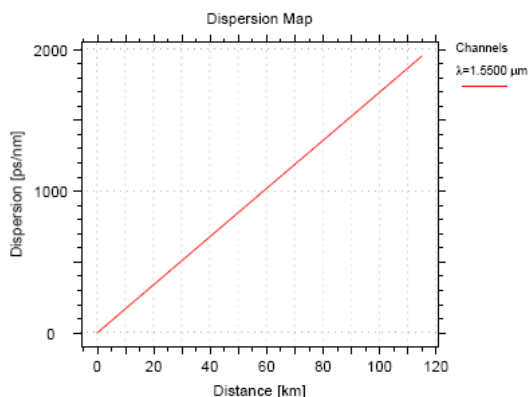
B. Múd blokovej simulácie – Vplyvy chromatickej disperzie

V tomto príklade budeme skúmať účinky chromatickej disperzie na jednovlákňové optické vlákna rôznej dĺžky (70km, 80km, 95km a 115 km). Topológia jednovlákňovej simulácie účinkov disperzie je na Obr. 4. Podľa dĺžky použitého SM OV sa hromadí rôzne množstvo chromatickej disperzie, ako je to uvedené na rozptylovej mape Obr. 4. Vzhľadom k tomu, že naše hlavné zameranie je analyzovať vplyv chromatickej disperzie, preto sú ostatné vlastnosti, ako tlmenie vlákna, jeho nelinearita, atď. ignorované.



Obr. 4 Zapojenie optickej linky pre štúdium chromatickej disperzie

Obr. 5 ukazuje, že čím je dĺžka OV väčšia, tým je celková disperzia vyššia, čo vedie k rozšíreniu impulzov (medzisymbolovej interferencii) a tým ku zníženiu, t.j. zhoršeniu chybovosti prenosu (BER – Bit Error Rate) (Obr. 6).

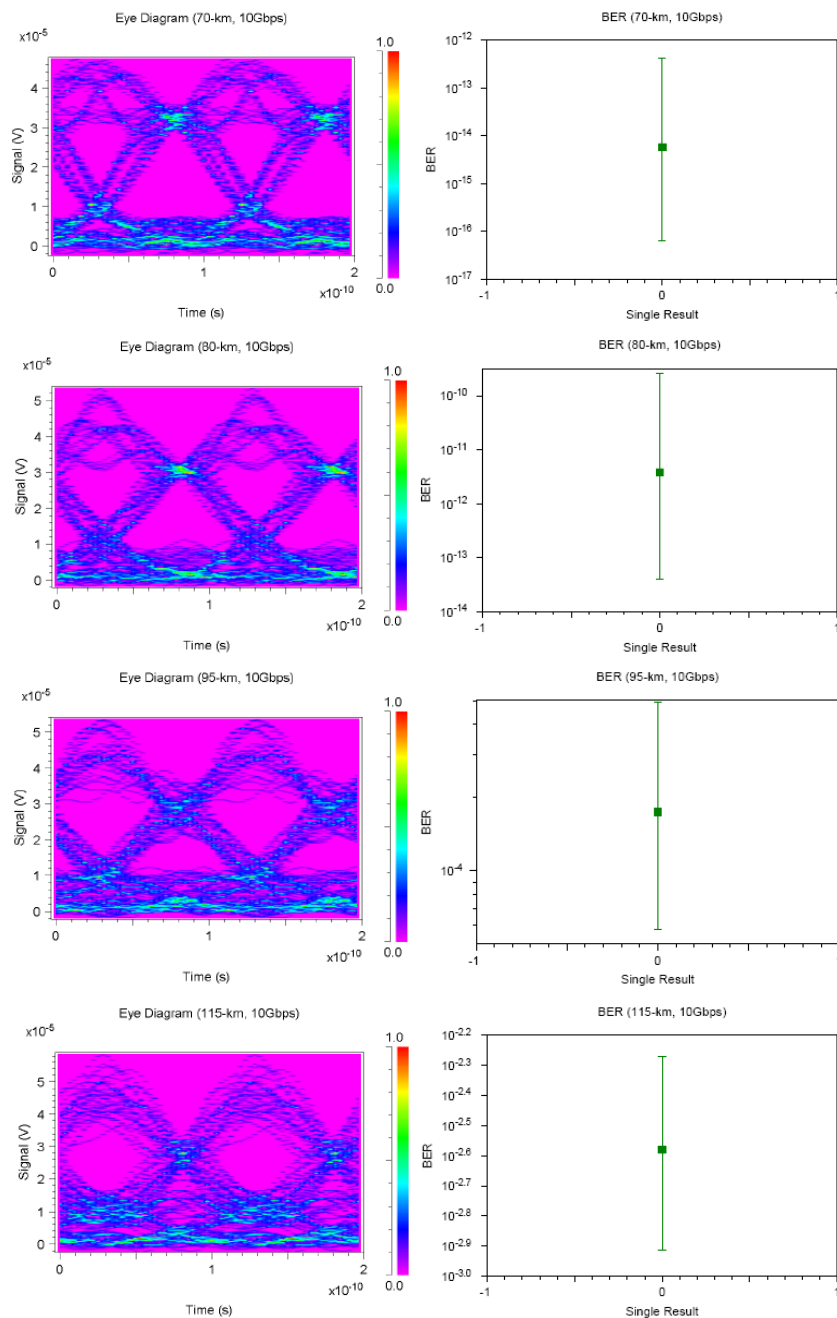


Obr. 5 Optické spektrum na výstupe optického vlákna pre konštanty: a) 0,1; b) 1,0

IV. ZÁVER

Programovací systém „OptSim“ je intuitívne modelovacie a simulačné prostredie podporujúce návrh a stanovenie výkonnosti fyzickej vrstvy (prenosovej časti) OVKS. Vďaka ľahkému ovládaniu GUI a analytickým nástrojom umožňuje rýchle získanie vierohodných

výsledkov. Obsiahla knižnica modelov komponentov a simulačné algoritmy poskytujú najlepší kompromis medzi presnosťou a rýchlosťou analýzy optického systému. Softvér je možné použiť nielen pre optimalizáciu optických trás (disperzné mapy, nastavenia výkonu, zisku, poloha zosilňovačov,...), ale aj na analýzu vplyvu lineárnych i nelineárnych javov na výkonnosť systému (PMD, SPM, FWM,...).



Obr. 6 Diagramy oka a zodpovedajúce hodnoty BER pre rôzne dĺžky optického vlákna

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] <http://www.rsoftdesign.com/products.php?sub=System+and+Network&itm=OptSim>
- [2] http://www.safibra.cz/download/Rsoft_Optsim.pdf
- [3] http://www.emo.org.tr/ekler/3b7e214bca4bf66_ek.pdf

Projektovanie OVKS s využitím programového balíka „ModeSYS“

Ján TURÁN, Ľuboš OVSENÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — V tejto práci je stručne opísané intuitívne modelovacie a simulačné prostredie programovacieho balíka „ModeSYS“, podporujúce návrh, analýzu a simuláciu mnohovidových optických komunikačných systémov. Obsiahla knižnica modelov komponentov a simulačné algoritmy poskytujú najlepší kompromis medzi presnosťou a rýchlosťou pre skrátenie času potrebného pre získanie správnych výsledkov návrhu.

Kľúčové slová — projektovanie OVKS, programový balík ModeSYS, optické vláknové komunikačné systémy, OptSim

I. ÚVOD

Programovací systém „ModeSYS“ bol vyvinutý pre potreby návrhu, analýzy a simulácie mnohovidových optických vláknových komunikačných systémov (OVKS). Tento softvér umožňuje presnú analýzu elektromagnetického poľa od jeho vygenerovania v laseri, cez jeho naviazanie do optických komponentov, či do mnohovidových (MM – Multi Mode) optickými vláknami (OV), až po jeho detekciu na optickom prijímači. Program dovoľuje užívateľovi vyhodnotiť ako časové, tak aj priestorové atribúty šírenia optického signálu [1,4].

Program „ModeSYS“ je jedinečným CAD nástrojom ponúkajúci progresívne schopnosti, ktoré sú nedostupné kdekoľvek inde. Ideálne sa hodí na počítačmi podporované návrhy mnohovidových OVKS, ale nič ho neobmedzuje ani pri simulácii a návrhoch:

- Gigabit-ový Ethernet, napr. 1GbE, 10GbE;
- 10GBASE - systémy, napr. SX, LX-4 a LRM;
- Serial/WDM systémov;
- FTTx/PON systémov;
- Plne optických sietí;
- Optických rozdeľovačov, prepínačov a odbočníc;
- FSO (Free Space Optics) systémov.

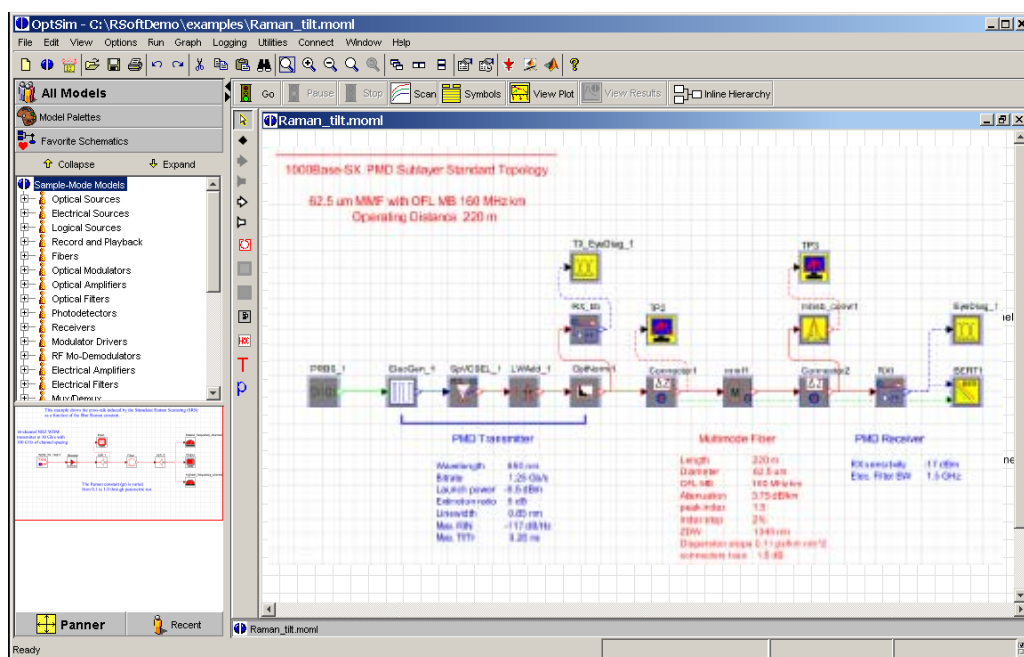
„ModeSYS“ je samostatný nástroj, ktorý môže byť plne integrovaný s programom „OptSim“. Spoločne tvoria jedinečný nástroj pre simuláciu jednovidových („OptSim“) a zároveň mnohovidových („ModeSYS“) optických systémov.

„OptSim“ je navrhnutý tak, aby skombinoval čo najväčšiu presnosť a schopnosť návrhu, s kompatibilitou a jednoduchosťou použitia na oboch operačných platformách, t.j. Windows (2000/XP/Vista (32- a 64-bit)) a Unix (Linux (32- a 64-bit)). Obsahuje rozsiahlu knižnicu komponentov, v ktorej sa nachádzajú najčastejšie používané prvky pri návrhu optoelektronických systémov. Táto knižnica je neustále rozširovaná vďaka prebiehajúcemu výskumu a spolupráci so svetovo uznávanými špecialistami na simuláciu a analýzu optických vláknových komunikačných systémov (OVKS). Okrem rozsiahlej knižnice komponentov, softvér zahŕňa aj sofistikované, ale zároveň ľahko použiteľné grafické rozhranie (GUI – grafický užívateľský interfejs) na efektívnu analýzu výsledkov simulácie. GUI ponúka hierarchické, objektovo orientované CAD prostredie pre návrh systémových zapojení a schém.

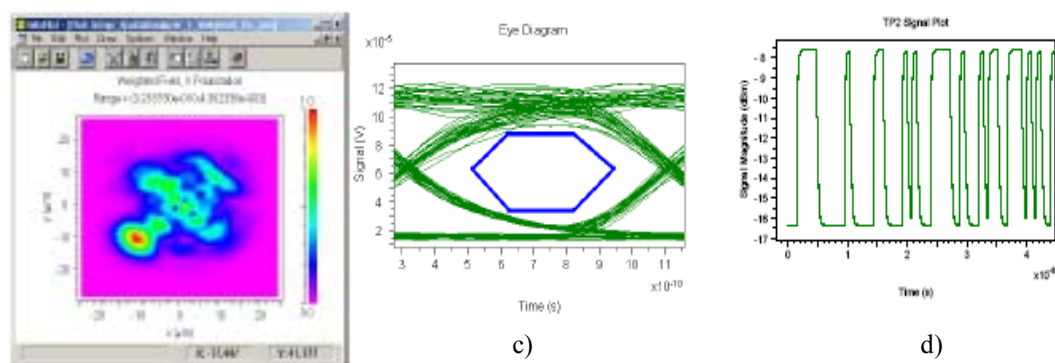
Funkcie programovacieho balíka „ModeSYS“ [2,3]:

- Model MM vlákna zahŕňa riešenie Helmholtzovej rovnice, pričom podporuje ľubovoľný index lomu a poruchy indexu lomu OV;
- Rozsiahly súbor nástrojov pre meranie umožňuje užívateľovi analyzovať kľúčové vlastnosti MM optických komunikačných systémov a prvkov, ako sú:
 - index lomu a profil indexu lomu vlákna;
 - väzba vlákno/laser, vlákno/vlákno, vlákno/fotodetektor;

- multiplexovanie optických signálov;
- módomá a chromatická disperzia;
- efektívna šírka pásma;
- časová simulácia optického signálu;
- spektrálna (priestorová) simulácia optického signálu;
- tvorba diagramov „oka“;
- chybovosť prenosu (BER – Bit Error Rate), atď.
- Model simulovania spôsobu šírenia v MM vlákne podporuje meranie disperzie a tlmenia vlákna (DMA - Differential Mode Attenuation);
- Komplexný priestorový model podporuje deterministické a štatistické analyzovanie parametrov optických komponentov (spojok, konektorov, atď.);
- Model výkonovej bilancie systému;
- Aplikačný programový inerfejs (API - Application Programming Interface) pre programovacie jazyky, ako napr. C/C++ či Fortran, pre vývoj vlastných užívateľských modelov.



a)



c)

d)

b)

Obr. 1 GUI programového balíka „ModeSYS“ pre a) 10GBASE-LX4 systém s b) rozložením podľa optického signálu na výstupe OV, c) diagramom oka s maskou oka d) priebehom signálu na výstupe vysielača

II. POPIS PROGRAMOVACIEHO BALÍKA „MODESYS“

A. Simulačné techniky

V mnohovidových komunikačných systémoch je vidová disperzia kľúčovým limitujúcim faktorom pre priepustnosť systému. Rôzne druhy vidov vo vlákne majú rôzne skupinové rýchlosti, čo vedie ku spomínanej vidovej disperzii, ktorej dôsledkom je medzisymbolová interferencia, t.j. rozširovanie impulzov na prijímači. Budenie jednotlivých vidov prenášaných

MM optickým vláknom má veľký vplyv na celkový výkon prenosového systému a je silne závislá od počiatočných podmienok výstupného výkonu optického signálu lasera, ktorý sa naväzuje do prenosového OV.

Programovací balík „ModeSYS“ môže byť použitý na simuláciu podmienok naviazania optického výkonu do OV s ľubovoľným indexom lomu, pričom výsledkom tejto simulácie bude stanovenie či prislúchajúca dvojica zdroj/vlákno spĺňa požiadavky telekomunikačných štandardov na prenos optických signálov bez zníženia kvality prenosu (BER) pod stanovenú hranicu. Okrem toho sa pre daný pár dajú určiť aj ďalšie parametre, ako napr. analýza diferenciálneho oneskorenia jednotlivých vidov pre daný typ vlákna s príslušným indexom lomu, či výpočet efektívnej prenosovej šírky pásma (EMB – Effective Modal Bandwidth) tohto vlákna.

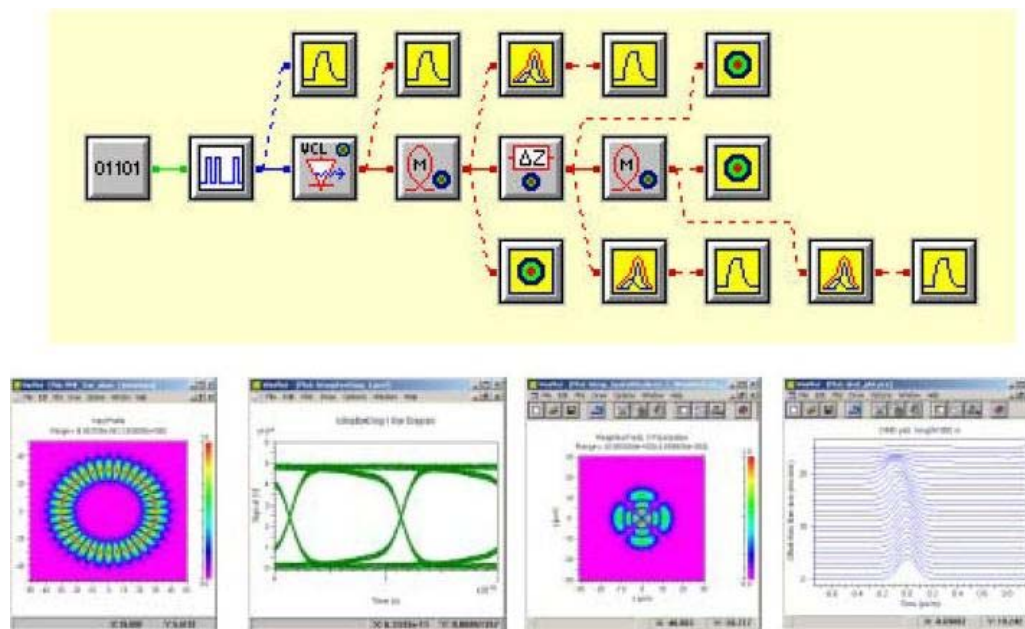
Softvér „ModeSYS“ pracuje v tzv. *móde blokovej simulácie*, čo znamená že vykonáva simulácie, v ktorých je dátový signál odovzdávaný medzi jednotlivými komponentmi, predstavujúcimi simulovaný čas v jednom bloku dát. Výhodou tohto prístupu je, že komponenty a algoritmy môžu ľahko pracovať s celým signálom, pričom ho môžu transformovať medzi časovou a frekvenčnou oblasťou, a tak pôsobiť na údaje v oblastiach výhodných pre simuláciu algoritmu.

Signál je teda pri tejto simulácii reprezentovaný ako časovými priebehmi, tak aj spektrálnymi závislosťami. Tento prístup podporuje simuláciu a analýzu nielen jednonábových mnohovidových systémov, ale aj multikanálových systémov využívajúcich multiplexovacie techniky, ako sú napr. WDM (Wavelength Division Multiplexing), CWDM (Coarse Wavelength Division Multiplexing), a pod.

Cez flexibilné užívateľské možnosti, ktoré ponúka „ModeSYS“, môžu užívatelia vytvárať modely, ktoré fungujú buď ako dátové signály vzoriek v čase, alebo ako veľké bloky v časovej alebo frekvenčnej oblasti.

B. Výsledky analýzy a ich následné spracovanie

Výsledky simulácií, ktoré sú produkované programom „ModeSYS“ zahŕňajú mapovanie priebehov signálov a diagramy oka na akomkoľvek mieste v rámci modelovaného OVKS. Sú dostupné aj ďalšie výsledky simulácie, ako je spektrum signálov, napájanie a mapy disperzie, atď. Výsledky simulácií môžu byť taktiež vykreslené vedľa seba, čo umožňuje ich porovnávanie so zosnímanými hodnotami parametrov. Na následné spracovanie výsledkov simulácií sú k dispozícii analytické nástroje, ktoré uľahčujú interaktívnu prácu s výsledkami simulácie a umožňujú prezentovať výsledky analýzy požadovaným spôsobom.



Obr. 2 Ukážka DMD (Differential Mode Delay) analýzy v programovacom balíku „ModeSYS“

C. Rozšírenie programového balíka „ModeSYS“

Pre vlastný vývoj modelu užívateľa „ModeSYS“ poskytuje rozhranie pre nástroje tretích strán, tj. užívateľské modely môžu byť napísané v programovacích jazykoch C++ alebo FORTRAN, a potom nahraté do simulačného nástroja využívajúceho rozhranie pre programovanie aplikácií (API).

Okrem toho podporuje „ModeSYS“ aj rozhranie s programom MATLAB, ktoré umožňuje používateľom zapisovať svoje užívateľské modely v tomto prostredí pre plnú spoluprácu s programom „ModeSYS“, a tiež možnosť využitia MATLABu pre spracovanie výsledkov simulácie.

D. GUI – CAD prostredie

„ModeSYS“ poskytuje hierarchicky, objektovo orientovanú topológiu schém pre jednoduchú definíciu systému., tzn. že všetky požadované komponenty a prvky simulovaného OVKS môžu byť ľahko vybrané z palety panela nástrojov, graficky pridané do schémy a prepojené pomocou myši na tvorbu požadovanej topológie.

Každý komponent reprezentovaný ikonou v schéme má svoju vlastnú sadu parametrov, ktoré sú prístupné jednoduchým kliknutím pravého tlačidla myši. Parametre môžu obsahovať číselné hodnoty (napr. laser vlnovú dĺžku, optické vlákno priemer), alebo výber medzi rôznymi nastaveniami (ako je napr. typ filtra).

Schémy je možné prispôbiť si zapnutím alebo vypnutím mriežky, zobrazenie alebo skrytie mena komponentu pri jeho ikone, zobrazenie vybraných parametrov komponentov na schéme, či zmenu rozmiestnenia portov na ikone. Schému možno zväčšovať a zmenšovať pomocou zoomu neobmedzene a stopa je uvedená v ľavom dolnom rohu pre ľahšiu identifikáciu a kontrolu, ktorá časť celej schémy je viditeľná v editačnom okne.

Komponenty môžu byť zoskupené a použité na vytvorenie združených komponentov (Compound Components) tiež známych ako superbloky, pre opakované použitie alebo zníženie „neporiadku“ v schéme, rovnako ako pre použitie v mnohých iných schémach.

Nastavenia parametrov komponentov možno ukladať do súborov alebo načítať zo súborov priamo z editovacieho okna komponentov. To podporuje vytváranie a používanie knižníc komponentov, kde rôzne súbory nastavení parametrov komponentov môžu reprezentovať rôzne fyzické komponenty.

E. Simulácia a analýza

Užívatelia tohto programového balíka môžu vytvárať vlastné modely mnohovídiových komponentov optoelektronických prvkov v systémoch GratingMOD™, LaserMOD™, BeamPROP™, alebo iných a následne ich môžu začleniť do prostredia „ModeSYS“, pričom tieto užívateľské modely sú rovnako sofistikované a efektívne, ako vstavané modely. Táto mnohého úrovňová simulácia poskytuje najlepšiu presnosť analýzy pri kombinácii vysokej kvality a zložitosti návrhu.

III. ZÁVER

Programovací systém „ModeSYS“ podporuje návrh a simuláciu mnohovídiových optických komunikačných systémov, s primárnym zameraním na aplikácie pre dátovú komunikáciu. Softvér „ModeSYS“ umožňuje používateľom hodnotiť časové a priestorové charakteristiky optického šírenia signálu. Môže byť použitý ako samostatný simulačný a analytický nástroj, alebo v kombinácii s programovým balíkom „OptSim“ môže slúžiť na vytvorenie komplexného návrhového aparátu pre skúmanie vlastností pracovného režimu jednovídiových a mnohovídiových optických komunikačných systémov.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] <http://www.micro.feld.cvut.cz/home/X34NFS/prednasky/R%20Soft%20n%A0vrhov%EC%20software%20.pdf>
- [2] http://www.safibra.cz/download/Rsoft_modesys.pdf
- [3] <http://www.foservice.com/SOFTWARE/modesys.htm>
- [4] http://www.allopticdesigns.com/productos/simulacion/modesys/ModeSYS_f.pdf

Videodohľadový systém pre detekciu opustenia batožiny

Anna KAŽIMÍROVÁ KOLESÁROVÁ, Ľuboš OVSENÍK, Ján TURÁN

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

anna.kolesarova@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — Viac ako kedykoľvek predtým je dôležité udržiavať bezpečnosť občanov, verejnej infraštruktúry alebo budov. S rastúcim množstvom bezpečnostných videí je veľmi dôležité, aby videodohľadový systém podporoval bezpečnosť sledovaných osôb a ich činností. V tomto príspevku je popísaný nový systém video dohľadu, ktorý sa zaoberá detekciou opustenia batožiny.

Kľúčové slová — videodohľadové systémy, bezpečnosť ľudí a ich majetku, opustenie batožiny

I. ÚVOD

Videodohľadové systémy sú veľmi rozšírené a bežné v mnohých prostrediach. Videodohľad bol kľúčovou úlohou v zaistení bezpečnosti na letiskách, bankách, kasínach a nápravných zariadeniach. V poslednom čase sa vládne agentúry, školy a obchodné spoločnosti obracajú na videodohľad ako na prostriedok, ktorý umožňuje zvýšiť verejnú bezpečnosť. S rozšírením lacných kamier a dostupnosti vysoko rýchlostných širokopásmových bezdrôtových sietí a rozmiestnenia obrovského počtu kamier sa videodohľad stal ekonomicky, ale aj technicky prijateľným pre širokú škálu užívateľov [10].

Mnoho otázok vedcov zostáva nezodpovedaných a to napríklad, či sa skutočne môžeme spoliehať na videodohľad ako na efektívny nástroj kriminálnej prevencie, vyriešenia kriminality, a ochrany pred kriminalitou [6].

Mnoho súčasných vedcov považuje za základ videodohľadových algoritmov analýzu videa, iných médií a automatickú detekciu významných udalostí [5].

Aplikácia vizuálnych dohľadov zahŕňa monitorovanie pohybu áut a osôb, sledovanie ľudskej činnosti pri detekcii neobvyklej aktivity, počítanie chodcov a pod. Typická dohľadová aplikácia pozostáva z troch blokov, a to z detekcie pohybu, sledovania objektov a vysoko úrovňovej pohybovej analýzy.

Multimediálne systémy môžu poskytovať dohľad na veľmi rozsiahlych oblastiach, a tak zabezpečiť viditeľnosť objektov aj na rozsiahlych plochách. Techniky, ktoré realizujú odovzdávanie dát medzi kamerami, v konfiguráciách so zosťrihanými alebo rozdelenými oknami, sa stávajú stále dôležitejšími. Zaujímavé javy, ktoré boli identifikované ako pohybujúce sa objekty a osoby, musia byť koordinované vo viacoknovom systéme a javy, ktoré vyvolávajú špeciálny záujem, musia byť sledované na celej scéne [7, 9].

Zopár videodohľadových produktov, ako napríklad systém na kontrolu strážnej služby, elektronické sledovanie dochádzky, GPS zariadenia, satelitné monitorovanie vozidiel, kamerové sledovacie systémy (CCTV) a iné, sú veľmi dobre dostupné na trhu pre verejnú ale aj domácu bezpečnosť. Tieto produkty monitorujú domácnosti, kancelárie alebo iné objekty, o ktoré je záujem. Tieto zachytávajú pohyb pomocou web kamier alebo videokamier [8]. V prípade web kamier sa vizuálne dáta ukladajú do komprimovaných alebo nekomprimovaných videosekvencií a systém spúšťa varovné alarmy, napríklad poslaním e-mailu alebo SMS-ky.

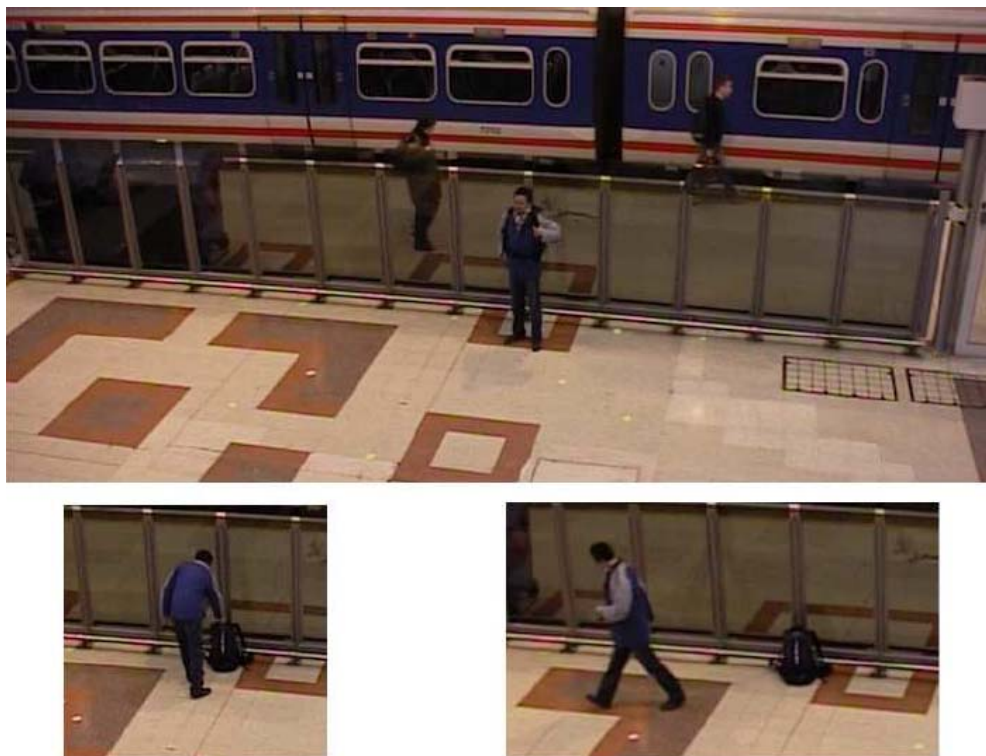
Nevyhnutnosť práce s komplexnými scénami charakterizovanými vysokou premenlivosťou vyžaduje použitie špecifických a sofistikovaných algoritmov pri video zbere, kalibrácii kamier, filtrovaní šumu a detekcii pohybu.

II. DETEKCIA OPUSTENIA BATOŽINY

Za posledné roky počet video dohľadových kamier dramaticky stúpa. Typickým účelom týchto kamier je podporovať udržanie bezpečia na verejných priestranstvách ako je metro,

centrum mesta, školy, nemocnice, finančne inštitúcie a športové priestranstvá. So zvyšujúcim množstvom kamier stúpa požiadavka pre automatické metódy pre inteligentnú analýzu.

Typickým bezpečnostným problémom je detekovanie opustenej batožiny bez dozoru, napr. na frekventovaných staniciach hromadnej prepravy osôb, tj. vlakovej stanici, stanici metra, atď. Pri takomto scenári, keď je detekovaná batožina bez dozoru a opustená na viac ako 30 sekúnd sa automaticky spustí alarm. Je to výzva pre automatické systémy a požaduje to dve základné schopnosti, t.j. schopnosť spoľahlivo detekovať batožinu a schopnosť spoľahlivo označiť vlastníka danej batožiny, ktorý ju tam zanechal bez dozoru. Obr. 1 predstavuje opustenie batožiny.



Obr. 1 Opustenie batožiny

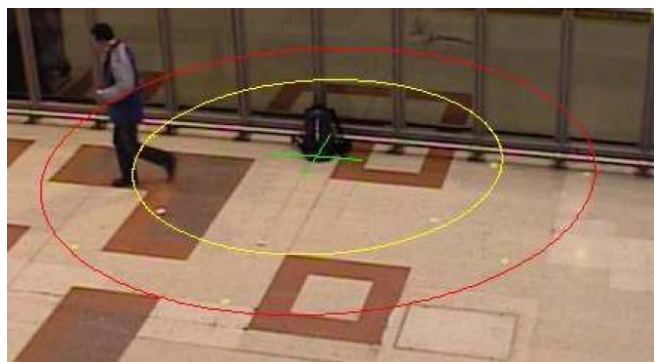
Môžeme na to využiť viac metód, ktoré navzájom spolupracujú, pretože táto detekcia a klasifikácia je veľmi zložitá. V prvom kroku aplikujeme model pravdepodobnostného sledovania. Model sledovania využíva dynamickú Bayesianovú sieť pre spoločnú reprezentáciu čísel ľudí v scéne a ich lokalizáciu. Automaticky odhadneme číslo objektu v scéne a jeho pozíciu odhadom základnej konfigurácie trans-rozmerného reťazca vzoriek MCMC (Markov Cain Monte Carlo). V ďalšom kroku sú výsledky z kroku sledovania odovzdané procesu detekcie batožiny, ktorá používa objektovú identifikáciu a lokalizáciu zo sledovača pre vyriešenie problému opustenej batožiny. Proces najprv vyhledá potenciálne objekty, ktoré by mohli byť batožinou a vyhodnotí pravdepodobnosť či je to batožina. Overí kandidátov na batožinu a vyhledá sekvencie vlastníka danej batožiny. Nakoniec batožina a vlastník budú identifikovaný a skontroluje sa, či kritériá pre zapnutie alarmu boli splnené.

Proces opustenej batožiny používa výstup z modelu sledovania a segmentáciu pozadia pre každý snímok. Identifikácia batožiny a rozhodnutie či a kedy boli opustené. Výstup z modelu sledovania pozostáva z čísla objektu ich identifikácie a lokalizácie.

Proces detekcie opustenej batožiny závisí od troch predpokladov ich vlastností:

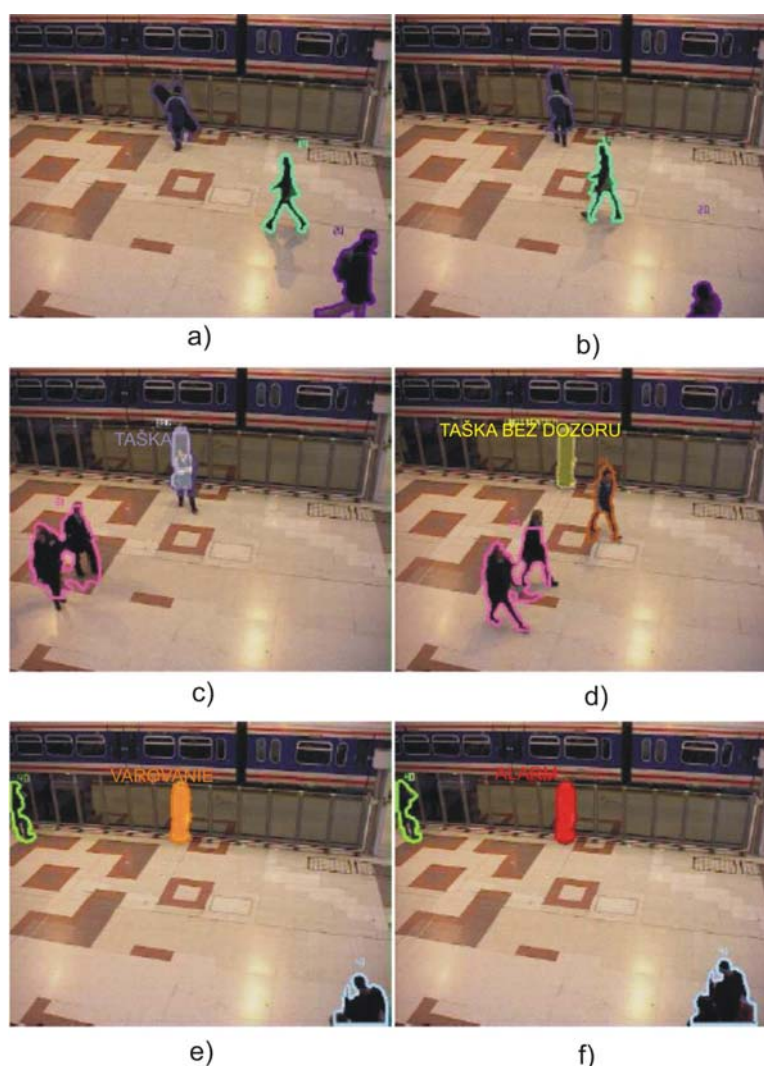
- opustená batožina sa pravdepodobne nehýbe,
- opustená batožina je pravdepodobne menšia ako ľudia,
- opustená batožina musí mať vlastníka.

Batožina ktorú vlastní osoba vstupujúca na scénu je pod dozorom pokiaľ je batožina vo fyzickom kontakte z osobou, alebo najďalej v rozmedzí dvoch metrov od batožiny (Obr. 2), t.j. žltý (vnútorný) kruh. Ak sa vlastník vzdiali od batožiny na viac ako tri metre, t.j. červený (vonkajší) kruh, tak systém čaká 30 sekúnd. Keď sa po ich uplynutí vlastník nevráti do blízkosti batožiny, tak systém automaticky vyhlási alarm.



Obr. 2 Zóny opustenia batožiny žltý kruh (vnútorný-2m) a červený kruh (vonkajší-3m)

Na Obr. 3 vidíme, ako na snímkach 3a) a 3b) osoba ukladá batožinu ku stene (batožina sa skladá z lyžiarskeho vybavenia). Na snímke 3c) bola správne zdetekovaná taška, ale osoba ktorá ju vlastní je v zóne dvoch metrov, takže ju ešte neopustila. Snímka 3d) nás už upozorňuje, že vlastník batožiny ju opustil, pretože je vo väčšej vzdialenosti ako sú tri metre. Snímka 3e) upozorňuje, že prešlo 30 sekúnd od doby čo jej vlastník opustil dvojmetrovú vzdialenosť od batožiny. Preto sa aktivovalo upozornenie. Na poslednom snímku 3f) môžeme vidieť ako sa aktivoval alarm, pretože prešlo 30 sekúnd od kedy majiteľ batožiny opustil priestor vo vzdialenosti väčšej ako je tri metre od batožiny [11].



Obr. 3 Detekcia opustenej batožiny na vlakovej stanici

III. ZÁVER

Videodohľadové systémy významne prispievajú ku kontrole a vyhodnocovaniu situácií tak, že systém transformuje videodohľad z nástrojov, ktoré zbierajú dáta na informáciu pre inteligentné dolovacie systémy. Video analýza vykonaná v reálnom čase dodáva dohľadovým systémom schopnosťou zareagovať na určitú aktivitu okamžite, a tak poskytnúť dôležitú informáciu [4].

Aj napriek dôležitosti témy a intenzívnemu výskumu platí, že detekcia pozadia zostáva zložitým problémom v aplikáciách pri rôznych situáciách, ako sú meniace sa osvetlenie, vlniace sa stromy, voda, otáčanie sa listov, pohybujúce sa tieň, odrazy, zakrývanie, výnimočné zmeny na pozadí, veľká premávka a pod. Zjednodušene, modely statických pozadí nevyriešia tento problém. Niektoré sú výpočtovo veľmi zdĺhavé a preto nemôžu byť použité v aplikáciách, ktoré musia byť vykonané v reálnom čase [1,2].

Problém vzdialených dohľadov má v posledných rokoch stúpajúcu tendenciu a to hlavne v súvislosti s monitorovaním infraštruktúry pre aplikácie v doprave, s bezpečnosťou pri kontrole kvality v priemysle a s vylepšením verejnej bezpečnosti z hľadiska kriminality. Vývoj dohľadových systémov vyžaduje expertízy z rôznych vedných oblastí, ktoré obsahujú znalosti spracovania signálov a obrazov, počítačového videnia, komunikácií, vytvorenia siete rozpoznávania vzorov a vývoja senzorov [3].

Úlohou nášho systému je zabrániť nepovolenému ponechaniu podozrivej batožiny v stráženom priestore. V takejto podozrivej batožine či balíku, by sa mohla nachádzať bomba, zbraň, drogy, atď. Ďalšou úlohou môže byť kontrola, či niekto kradne batožinu, balíček alebo iné hodnotné veci. Náš systém zvyšuje bezpečnosť ľudí v školách, na staniciach, letiskách, a pod.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] V. Beran, A. Herout, I. Řezníček, "Video-Based Bicycle Detection in Underground Scenarios", in: *Proceedings of WSCG'09*, Plzeň, Czech Republic, 2009, 4 pp.
- [2] A. Bergeron, "Optical Correlator for Industrial Application, Quality Control and Target Tracking", *Sensor Review*, Vol. 20, No. 4, 2000.
- [3] M. Bhargava, CH. Chen, M. S. Ryoo, J. K. Aggarwa, "Detection of Object Abandonment Using Temporal Logic", Springer Berlin, January 2009, pp. 271-281.
- [4] T. Boult, et al. "Into the Woods: Visual Surveillance of Noncooperative and Camouflaged Targets in Complex Outdoor Settings", in *Proceeding of the IEEE*, Vol. 89, No. 10, Oct. 2001.
- [5] R. Cucchiara, "Multimedia Surveillance systems", *Proc. ACM Workshop on Video Surveillance and Sensor Networks*, 2005, pp. 3-10.
- [6] Z. Dimitrijevič, G. Wu, E. Chang, "A Multi-Sensor Fusion and Mining System", *Proceeding of the 2nd Usenix FAST*, March 2003.
- [7] A. Elgammal, D. Harwood, L. A. Davis, "Non-Parametric Model for Background Subtraction", in: *Proc. ICCV'99*, 1999.
- [8] T. Horprasert, D. Harwood, L. A. Davis, "A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection", in: *Proc. ICCV'99 Frame Rate Workshop*, 1999.
- [9] H. Haritaoglu, "Hartwood and Devis, W4: Real Time Surveillance of People and their Activities", *IEEE Trans. Pattern Anal. Machine Intell*, Vol. 22, No. 8, August 2000, pp. 809-830.
- [10] R. Rangaswami, Z. Dimitrijevic, K. Kakligian, E. Chang, Y. F. Wang, "The SfinX Video surveillance system", 2003.
- [11] K. Smith, P. Quelhas, and D. Gatica-Perez, "Detecting Abandoned Luggage Items in a Public Space", *IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL)*, Switzerland, IEEE 2007.

Architektúra videodohľadového systému na detekciu opustenej batožiny

Anna KAŽIMÍROVÁ KOLESÁROVÁ, Ján TURÁN, Ľuboš OVSENÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

anna.kolesarova@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — Udalosti v posledných rokoch nás nútia zvyšovať našu bezpečnosť, ako aj bezpečnosť nášho majetku. Preto základným cieľom aplikácií videodohľadových systémov je detekovať, sledovať a klasifikovať potenciálne nebezpečné ciele. V tomto článku je popísaný návrh architektúry videodohľadového systému na monitorovanie a detekovanie opustenej batožiny.

Kľúčové slová — videodohľadové systémy, bezpečnosť ľudí a ich majetku, opustenie batožiny, optický korelátor

I. ÚVOD

Udalosti v posledných rokoch nás nútia zvyšovať našu bezpečnosť, ale aj bezpečnosť nášho majetku. Jedným z prostriedkov na toto zlepšenie sú videodohľadové systémy, ktoré využívajú poznatky z rôznych vedných oblastí (napr. z oblasti počítačového videnia, spracovania signálov, neurónových sietí, optiky a pod.), väčšinou si pod dnešnými videodohľadovými systémami väčšina z nás predstavuje rozsiahle analógové, alebo digitálne videorekordéry. V skutočnosti predstavujú sofistikované digitálne systémy na spracovanie videa a obrazu.

Dôležitosť videodohľadovej techniky významne vzrastá od posledných teroristických útokov. Ochrana a bezpečnosť sa stáva kritickou v mnohých verejných priestoroch (školy, galérie, divadlá, vlakové stanice, autobusové stanice, centrá miest, obchodné centrá, letiská a pod.) a preto je obzvlášť potrebné umožniť videodohľadovým operátorom vzdialené monitorovanie aktivít v týchto prostrediach.

Poznáme mnoho aplikácií videodohľadových systémov, kde digitálna analýza obrazu a videa prináša ich lepšiu výkonnosť. Príkladom aplikácií týchto systémov je aj detekcia „nežiaducich“ osôb, detekcia nezvyčajného chovania sa ľudí, či detekcia opustenej batožiny a pod.

II. NÁVRH SYSTÉMU NA DETEKCIU OPUSTENEJ BATOŽINY

Tento systém (Obr. 1) spája nízko úrovňové sledovanie a vysoko úrovňovú klasifikáciu pomocou Bayesovho klasifikátora pre vyriešenie udalosti rozpoznávacích problémov. V prvom rade modul sledovania kombinuje výsledky z modulu sledovania zmien a sledovania ľudskej postavy. Výsledky týchto modulov sledovania nám poskytujú informácie o trajektoriách a tiež základné informácie o tom, či sa jedná o človeka alebo o batožinu. Detektor zmien na báze Kalmanovho filtra nám poskytne informácie o všetkých nových objektoch, ktoré sa objavili na scéne (ľudia, predmety) a detektor ľudskej postavy detekuje iba osoby. Ak porovnáme tieto dve informácie, tak môžeme jednoznačne určiť či v sledovanej oblasti niečo osoby zanechali.

Modul klasifikácie používa aj ďalšie užitočné vlastnosti, ktoré získava z modulov sledovania ako sú rýchlosť, oblasť a vzdialenosť medzi objektami. Tieto vlastnosti sú použité ako príznaky. Udalosti sú reprezentované ako hypotézy rozpoznávané pomocou Bayesovho klasifikátora. Udalosti pozostávajú z jednoznačných definícií:

- **Kontextové pravidlo:** osoba ktorá práve vstúpila do scény a vlastní batožinu. Táto batožina je pod jej dozorom pokiaľ je batožina vo fyzickom kontakte s osobou.
- **Priestorové pravidlo:** batožina je bez dozoru ak jej vlastník je od nej vzdialený na viac ako 3 metre.
- **Časové pravidlo:** ak batožina je bez dozoru vlastníka viac ako 30 sekúnd je aktivovaný alarm.



Obr. 1 Bloková schéma navrhnutého systému na detekciu opustenej batožiny

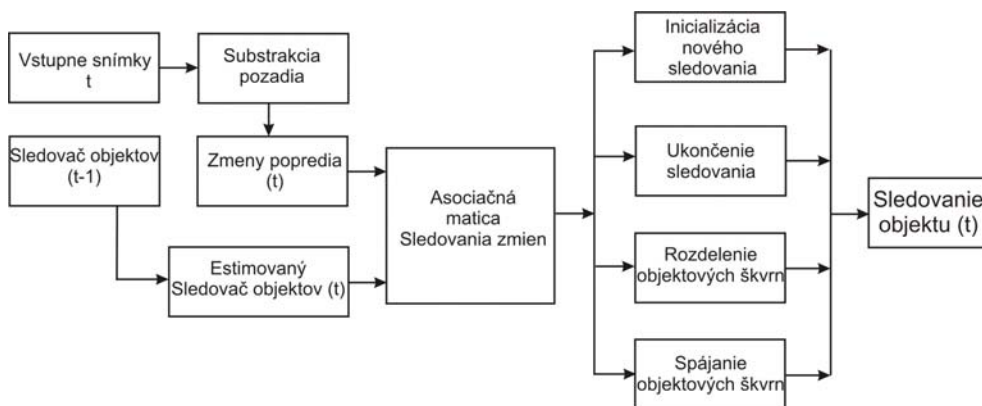
Optický korelátor slúži na rozpoznávanie tvári. Využíva výsledky z Reading People Trackera hlavne detektora hlavy. Je to doplnok k detekcii opustenej batožiny. Môžeme v ňom využívať rôzne databázy, ktoré sú napríklad napojené na políciu a tak detekovať či sa v danej oblasti pohybujú nebezpečné osoby.

A. Modul Sledovania

Pre sledovanie boli navrhnuté dva komponenty, ktorých výsledky sa neskôr spracujú. Prvým komponentom je Kalmanov filter založený na sledovaní zmien, ktorý poskytuje informácie o objavení sa objektov v scéne a ich ceste, vrátane ľudí. Ďalší komponent je Reading People Tracker, čo je programový balík pre detekciu ľudských postáv na základe ich fyziologických vlastností.

A.1 Kalmanov filter založený na sledovaní zmien

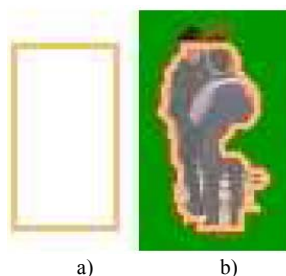
Na Obr. 2 je znázornená bloková schéma sledovania zmien na báze Kalmanovho filtra. Segmentácia je robená na základe substrakcie pozadia, kde sa porovnáva popredie s pozadím a tak sme schopní detekovať zmeny.



Obr. 2 Bloková schéma sledovanie zmien na báze Kalmanovho filtra

Každý snímok „t“ a škvrny (blob), respektíve zmeny ktoré nastali po segmentácii označíme ako B_t^i . Pozostávajú z nasledujúcich informácií:

- veľkosť a lokalizácia, pre jednoduchosť používame štvorhran, ktorý ohraničuje detekovanú zmenu, resp. škvrnu Obr. 3a),
- štvorhran, ktorý ohraničuje zmenu je farebný.



Obr. 3 Pracovný príklad dohľadového algoritmu

Na Obr. 3b) je výsledok po segmentácii určitej oblasti. Pozadie je zelené a zmena je sivá a jej hrany sú zvýraznené. Najideálnejšie je, ak jedna škvrna korešponduje s jedným reálnym objektom a naopak. Ale musíme zvažovať situácie, keď jeden objekt sa spojí alebo rozdelí do niekoľkých škvŕn. Pre škvrnu použijeme označenie B_t^i a pre objekt O_t^i , poukážeme na spojenie hodnôt založenej na prekrývaní medzi ohraničujúcim štvorhranom B_t^i a O_t^i .

$$M(B_t^i, O_t^i) = \begin{cases} 1, & \text{ak } \frac{Oblast'(B_t^i \cap \hat{O}_t^j)}{\min(Oblast'(B_t^i), Oblast'(\hat{O}_t^j))} \geq \tau \\ 0, & \text{Inak} \end{cases} \quad (1)$$

Kde $\hat{O}_t^j$ je predikovaný ohraničujúci štvorhran O_t^j z predchádzajúcej snímky použitím Kalmanovho filtra a $B_t^i \cap \hat{O}_t^j$ je prienik dvoch ohraničujúcich štvorhranov a τ je prahová hodnota medzi 0 a 1. Ak $M(B_t^i, O_t^i) = 1$, potom dané B_t^i a O_t^i považujeme za spojené.

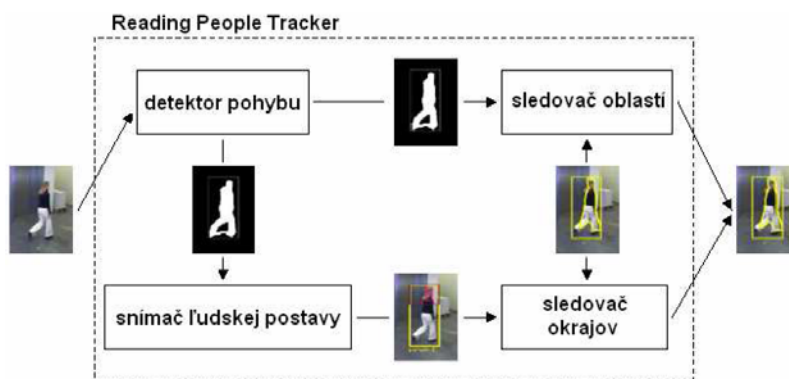
Vo väčšine prípadov sledovač zmien vykazuje veľmi dobré výsledky v sledovaní objektov a lokalizácii batožiny. Ale ak sa na scéne objavuje od začiatku sledovania viac ľudí, ktorý sa javia pre algoritmus ako jeden objekt, tak dochádza k chybám identifikácie [1].

A.2 Reading People Tracker

Je to programový balík navrhnutý na sledovanie ľudí pre pozorovacie systémy. Pozostáva zo štyroch modulov, ktoré medzi sebou spolupracujú (Obr. 4):

- Detektor pohybu,
- Detektor hlavy,
- Sledovač oblastí,
- Sledovač okrajov.

Výsledky sú stabilnejšie a spoľahlivejšie, keď sú použité tieto štyri moduly namiesto jedného, pretože nedostatky a obmedzenia jedného modulu sú kompenzované ostatnými modulmi. Systém používa dvojrozmerný model ľudskej postavy a možné variácie sú vybrané z rozsiahleho množstva preddefinovaných tvarov. Pozadie je zadané bez ľudí a obraz je priebežne upravovaný mediánovým filtrom. Sledovač oblastí ma za úlohy detekovať všetky pohyblivé oblasti, ktoré získal z výstupu snímača pohybu. Sledovač oblastí tiež sleduje, či dané oblasti sa spájajú alebo rozdeľujú. Snímač ľudskej postavy sa hlavne zaoberá detekovaním ľudskej hlavy. Je to výhodou pre sledovač okrajov, pretože je ľahšie identifikovať človeka podľa polohy jeho hlavy [2].



Obr. 4 Moduly Reading People Tracker

Detektor pohybu ma za úlohu vypočítať binárny obraz pohybu resp. zmeny v aktuálnej snímke oproti referenčnej snímke, resp. pozadiu. Výsledok z detektora pohybu vchádza do snímača ľudskej postavy a sledovača oblastí. Sledovač oblastí sleduje oblasti v čase a detektor ľudskej postavy sleduje možné polohy hlavy. Sledované oblasti v čase a možné polohy hlavy sú vstupmi pre sledovač okrajov. Výsledkom spolupráce týchto modulov je detekovanie ľudských postáv vo videosekvenciách [2].

B. Modul Rozpoznávania

Pre modul rozpoznávanie opustenej batožiny je použitý Bayesov klasifikátor, ktorý vychádza z Bayesovej vety o podmienenej pravdepodobnosti. Pre túto špecifickú úlohu detekcie opustenej batožiny sú formované nasledujúce udalosti [1].

B.1 Zdetekovanie batožiny: tri udalosti

- E1: Batožina sa neobjavila,
- E2: Batožina sa objavila,
- E3: Vzdialenosť medzi osobou a batožinou je menšia ako nastavená prahová hodnota, napr. 1,5m.

Udalosti E1 a E2 spolu určujú časové obmedzenie, t.j. batožinu nosí osoba pred tým, ako je batožina detekovaná, takže osobitné sledovanie ešte nezačalo.

Udalosť E3 je priestorové obmedzenie, t.j. vlastník je nablízku, keď sme detekovali batožinu. Toto obmedzenie eliminuje nepodstatné osoby, ktoré prechádzajú v blízkosti keď sme batožinu detekovali. Najbližšia osoba pri batožine je považovaná za jej vlastníka.

B.2 Opustená batožina: tri udalosti

- E1: Vzdialenosť medzi osobou a batožinou je menšia ako prahová hodnota, napr. 3m,
- E2: Vzdialenosť medzi osobou a batožinou je väčšia ako prahová hodnota, napr. 3m,
- E3: Čas od kedy vlastník opustil batožinu a nachádza sa vo väčšej vzdialenosti ako prahová hodnota. Časový rozsah je väčší ako 30 sekúnd odkedy vlastník prekročil prahovú hodnotu vzdialenosti.

Udalosti E1 a E2 spoločne zabezpečujú priestorové pravidlo a udalosť E3 zabezpečuje časové pravidlo. Ak je výsledok pravdepodobnosti najvyšší, čiže vlastník opustil batožinu, je vo väčšej vzdialenosti ako prahová hodnota a uplynulo 30 sekúnd od toho čo vlastník prekročil prahovú hodnotu a nevrátil sa, tak systém aktivuje alarm.

B.3 FARCO (Fast Face-Recognition Optical Correlator)

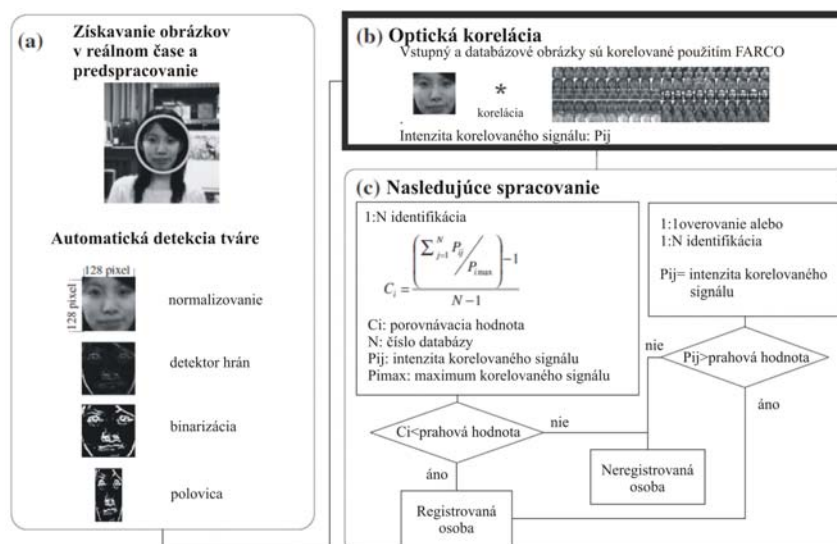
Systémy na rozpoznávanie tváre sa používajú v širokom spektre bezpečnostných systémoch, ako tých na monitorovanie používateľov kreditných kariet, identifikovanie jednotlivcov z videodohľadových kamier a monitorovanie cestujúcich pri imigračnej kontrole, atď. Rozpoznávaním tváre sa vývojári zaoberajú už od roku 1970. Ešte aj dnes sa musíme vysporiadať z niektorými podmienkami. Hlavne je to počet obrázkov, ktorý môžeme uložiť, tu sme limitovaný dostupným aktuálnym systémom a rozpoznávací rýchlosť musí byť zlepšená hlavne preto, lebo tváre môžu byť zachytené z rôznych uhlov a s premenlivými svetelnými podmienkami.

Na rozdiel od digitálneho rozpoznávania, optický operačný proces nám podobnosť dvoch 2D obrázkov určí okamžite. Používa Fourierovu transformáciu realizovanú na báze šošoviek. V roku 1960 boli vyvinuté dva druhy korelátorov: Vanderlugov korelátor a Joint Transform Correlator (JTC). Niektoré dostupné korelátory sú kombináciou týchto dvoch. FARCO korelátor, ktorý je určený pre rýchle rozpoznávanie tváre je založený na báze Vanderlugovho korelátoru. Má tri rozličné konfigurácie, ktoré sa líšia rýchlosťou rozpoznávania. Kombináciou vysoko rýchlostných digitálnych systémov spracovania obrazu a 4-kanalového spracovania môže dosahovať rýchlosť rozpoznávania tváre až 4000 tváre za sekundu [10].

Skupina okolo FARCO korelátoru navrhla ďalší systém, ktorý je super vysoko rýchlostný FARCO (S-FARCO) integruje v sebe optickú korelačnú technológiu použitú vo FARCO a súosi holografický pamäťový systém. Umožňuje optickú koreláciu bez potreby dekodovania informácií v databáze, tým sa veľmi zníži čas spracovania informácií. Predpokladaná opticko korelačná rýchlosť je okolo 3μs/snímok. Korelačná rýchlosť je viac ako 370 000 snímkov/s.

Algoritmus práce S-FARCO systému je na Obr. 5. V tomto systéme je použitý osobný počítač PC (CPU 3GHz a RAM 2GB) na predbežné a nasledujúce procesy pre zlepšenie pomeru S/N a robustnosti systému. Najskôr sú snímky tváre zachytené digitálnou video kamerou. Detekuje sa tvár a vyberie sa oblasť obrazu o veľkosti 128×128 o.p, kde sa nachádza tvár. Pre tie obrázky, ktoré sú snímané v uhle je použitá afinná transformácia k nastaveniu obrazu a následná optimalizácia obrazu je fixovaná na pozíciu očí. Pre zvýraznenie hrán sa používa sobelov detektor.

Korelačný signál je klasifikovaný úrovňou prahovej hodnoty. V praktických aplikáciách sa prahová hodnota mení v závislosti od bezpečnostnej úrovne [3].



Obr. 5 Algoritmus rozpoznávania tváre FARCO

III. ZÁVER

Videodohľadové systémy významne prispievajú ku kontrole a vyhodnocovaniu situácií tak, že systém transformuje videodohľad z nástrojov, ktoré zbierajú dáta na informáciu pre inteligentné údaje dolovacie systémy. Video analýza vykonaná v reálnom čase dodáva videodohľadovým systémom schopnosťou zareagovať na určitú aktivitu okamžite, a tak poskytnúť dôležitú informáciu. Aj napriek dôležitosti témy a intenzívnemu výskumu platí, že detekcia pozadia zostáva zložitým problémom pri rôznych situáciách, ako sú meniace sa osvetlenie, pohybujúce sa objekt v scéne, tieň, odrazy, veľká premávka atď. Niektoré takéto aplikácie sú výpočtovo veľmi náročné, preto si vyžadujú veľmi výkonný hardvér alebo sofistikovanejší softvér, aby sa mohli použiť v aplikáciách, ktoré pracujú v reálnom čase.

Ľudstvo stále vytvára lepšie a rafinovanejšie možnosti videodohľadu. Takmer každý mesiac sa na trhu objavujú stále menšie, kompaktnějšíe a výkonnejšie videodohľadové systémy.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] L. Fengjun, S. Xuefeng, W. Bo, S.K. Vivek, N. Ramakant, "Left-Luggage Detection using Bayesian Inference", In: *Institute for robotics and Intelligent Systems IEEE 2008*, University of Southern California, Los Angeles, 2008.
- [2] T. L. Bouffant, N. T. Siebel, S. Cook, S. Maybank, "Reading People Tracker Reference Manual". Computation Vision Group, Department of Science, The University of Reading, United Kingdom.
- [3] E. Watanabe, S. Ishikawa, k. Kodate "A Highly accurate face recognition system using filtering correlation". In: *Optical review*, Vol. 14, No. 5, 2007, pp. 255-259.

Experimentálne pracovisko pre zváranie optických vlákien

¹Luboš OVSENÍK, ²Ján TURÁN, ³Ludovít HINTOŠ

^{1,2}Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

³Slovak Telekom, a.s., Poštová 19, Košice 042 10, Slovenská republika

lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstrakt — V tejto práci je stručne opísané experimentálne pracovisko pre zváranie rôznych typov (SM, MM, DSF, NZD, EDF, BIF/UBIF (Bend insensitive fibre)) optických vlákien. Sú tu popísané pracovné postupy a celý cyklus zvárania pomocou zväračky a lámačky optických vlákien Fitel S178.

Kľúčové slová — spájanie optických vlákien, zváranie optických vlákien, zväračka Fitel S178

I. ÚVOD

Optické vláknové komunikačné systémy, tak ako aj iné komunikačné systémy si vo všeobecnosti vyžadujú vhodné prvky na spájanie, rozvetkovanie a ukončovanie prenosového média, t.j. optických vlákien. Keďže väčšina výrobcov zakončuje zdroje a detektory optického signálu pre komerčných užívateľov krátkym úsekom vhodného typu optického vlákna (tzv. pigtail), pri výstavbe dlhých úsekov optických trás prichádza v praxi do úvahy predovšetkým spojenie typu vlákno-vlákno. Tieto spoje možno vo všeobecnosti rozdeliť na dva druhy.

A. Optické vláknové spojky

Sú to nerozoberateľné druhy spojenia optických vlákien, určené na vytvorenie stálych a pevných spojení typu vlákno-vlákno. Môžu byť vytvorené mechanickým spojením alebo zváraním, ktoré možno špecifikovať podľa zdroja použitého tepla.

Výhody zvárania:

- Nízka cena na 1 spoj;
- Relatívne jednoduchá činnosť;
- Možná voľba metódy (ceny);
- Vložený útlm $0,01 \pm 0,15$ dB;
- Útlm odrazu >70 dB;
- Akceptovaná štandardná metóda v telekomunikáciach.

Nevýhody zvárania:

- Vyžaduje skúsenosť;
- Zmenšenie pevnosti vlákna;
- Počiatočné náklady sú vysoké.

B. Optické vláknové konektory

Sú to ľahko ručne viacnásobne rozoberateľné druhy spojenia optických vlákien. Optické konektory sa klasifikujú podľa počtu súčasne prepájaných (konektorovaných) optických vlákien. Obyčajne sa používajú jednovláknové konektory. Mnohovláknové konektory sa používajú na prepojenie najmä plochých (pásikových) optických káblov.

Tento článok opisuje pracovný postup a celý cyklus zvárania (vytvárania optických vláknových spojok) pomocou zväračky a lámačky optických vlákien typu Fitel S178.

II. POPIS ZVÁRAČKY A LÁMAČKY FITEL S178

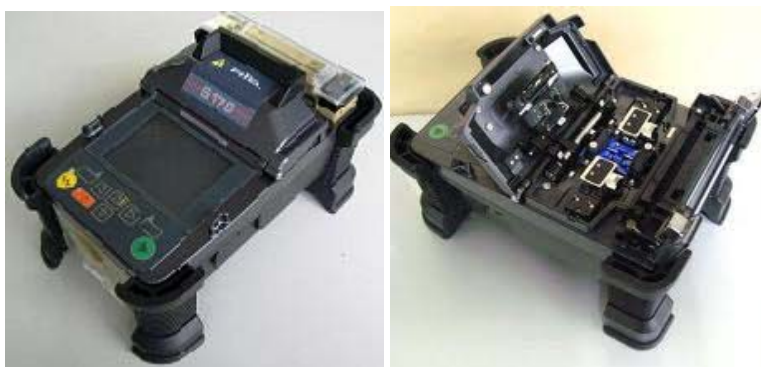
Automatická zväračka a lámačka fy Furukawa Fitel S178A (Obr. 1) s vystredením na jadro vlákna je nástupcom najpredávanejšieho modelu S177A. Nový model je vybavený odolnejším a ľahším telom a umožňuje rýchlejšie urobenie zvaru a zapečenie zmršťovacej ochrany zvarov.

Vďaka svojim vlastnostiam je S178A ideálnym prístrojom pre všetky typy aplikácií - pre siete FTTx, lokálne siete aj siete transportné. Jej hlavné výhody sú [1,2]:

- Vysoká mechanická odolnosť (bez poškodenia odolá pádu na betónovú podlahu);
- Vodeodolnosť (ipx2);
- Odolnosť proti prachu (ip5x);
- Podstatné zmenšenie tela a nižšie hmotnosti zväračky (1,9 kg);
- Rýchlejšie vykonanie zvaru (7 sekúnd);
- Rýchlejšie zapečenie zmršťovacej ochrany; 25 sekúnd pre ochranu 40 a 60mm;
- Nízka spotreba umožňuje urobiť bez nabíjania až 200 zvarov vrátane zapečenia;
- Konštrukcia bez využitia zrkadiel (minimalizácia údržby);
- Maximálne zväčšenie vlákien zaisťuje presnejšiu analýzu a zváranie optických vlákien;
- Automatické ukladanie snímok posledných 100 vykonaných zvarov;
- Možnosť použitia navarovných konektorov soc (splice on connector):
 - Technológia zvárania v konektore je špeciálna technika konektorovania, ktorá vznikla v spolupráci spoločnosti Seikoh Giken (dodávateľ optických konektorov) a spoločnosti Furukawa FITEL (dodávateľ optických zväračiek).

Základné technické údaje zväračky [1-5]:

- Použiteľné optické vlákna typu vlákna SM, MM, DSF, NZD, EDF, BIF/UBIF;
- Priemerný vložený útlm zvarov je pre SM: 0,02 dB; MM: 0,01 dB; DSF: 0,04 dB; NZD: 0,04 dB;
- Čas zvárania je 7s (semiautomatický mód);
- Čas zapekania 25 s pre 40 a 60 mm ochranu zvaru (Pre-heat mód);
- Automatická detekcia vlákien typu SM: SM, DSF, NZD, BIF/UBIF, MM: MM;
- Pamäť na 2000 zvarov;
- Zväčšenie vlákna 304x, 608x;
- Výdrž batérie na 200 zvarov/zapečení;
- 3,5" farebný LCD monitor.



Obr. 1 Zväračka a lámačka optických vlákien od fy Furukawa S178A

III. POSTUP SPÁJANIA OPTICKÝCH VLÁKIEN POMOCOU ZVÁRAČKY FITEL S178

A. Základné nástroje používané pri zváraní optických vlákien:

- Sťahovač plášťov;
- Čistiaci prostriedok (IPA, acetón);
- Lámačka vlákien;
- Zväračka vlákien.

B. Základné požiadavky na prípravu vlákna pred zváraním:

- Vlákna musia byť očistené od všetkých ochrán až na sklo (sťahovač plášťov);
- Musia byť dôkladne očistené a musia mať ploché konce (čistiaci prostriedok + lámačka);
- Musia byť presne zosúosené jadrami (zväračka vlákien).

C. Podľa techniky zosúosenia optických vlákien delíme zvaracie zariadenia na tri typy:

„V“ drážka

Pomocou veľmi presnej drážky a s využitím plášťa vlákna dochádza k zosúoseniu vlákien. Táto metóda je vhodná pre presne koncentrické jadro a plášť vlákna.

LIDS (Local Light Injection and Detection System)

Zosúosenie vlákien sa robí pomocou svetelného lúča naviazaného na jednej strane spoja a detekovaného na strane druhej. Jedno z vlákien sa posúva v dvoch osiach, kým sa nedosiahne maximum detekovaného signálu. Na naviazanie a detekciu testovacieho signálu sa využíva miesto ohybu (vlákna sú navinuté na dve rôzne podložky). Problematické sú vlákna s netransparentným plášťom (tu je táto metóda nepoužiteľná).

PAS (Profile Alignment System)

Zosúosenie vlákien sa robí využitím infračervenej (IR) kamery. Využívajú sa rôzne optické vlastnosti jadra a plášťa (index lomu). Ak je vlákno priečne osvetlené čiastočne kolimovaným zväzkom optického signálu, pomocou mikroskopu rozoznať jadro vlákna a za ním jeho tieň, čo sa pomocou zabudovaného videosystému (IR kamery) používa na zosúosenie jadier spájaných vlákien. Plášť sa javí ako valcová šošovka čiastočne fokusujúca zväzok lúčov a aj malý rozdiel indexov lomu (0,3%) je dostatočný pre vznik sekundárneho šošovkového efektu.

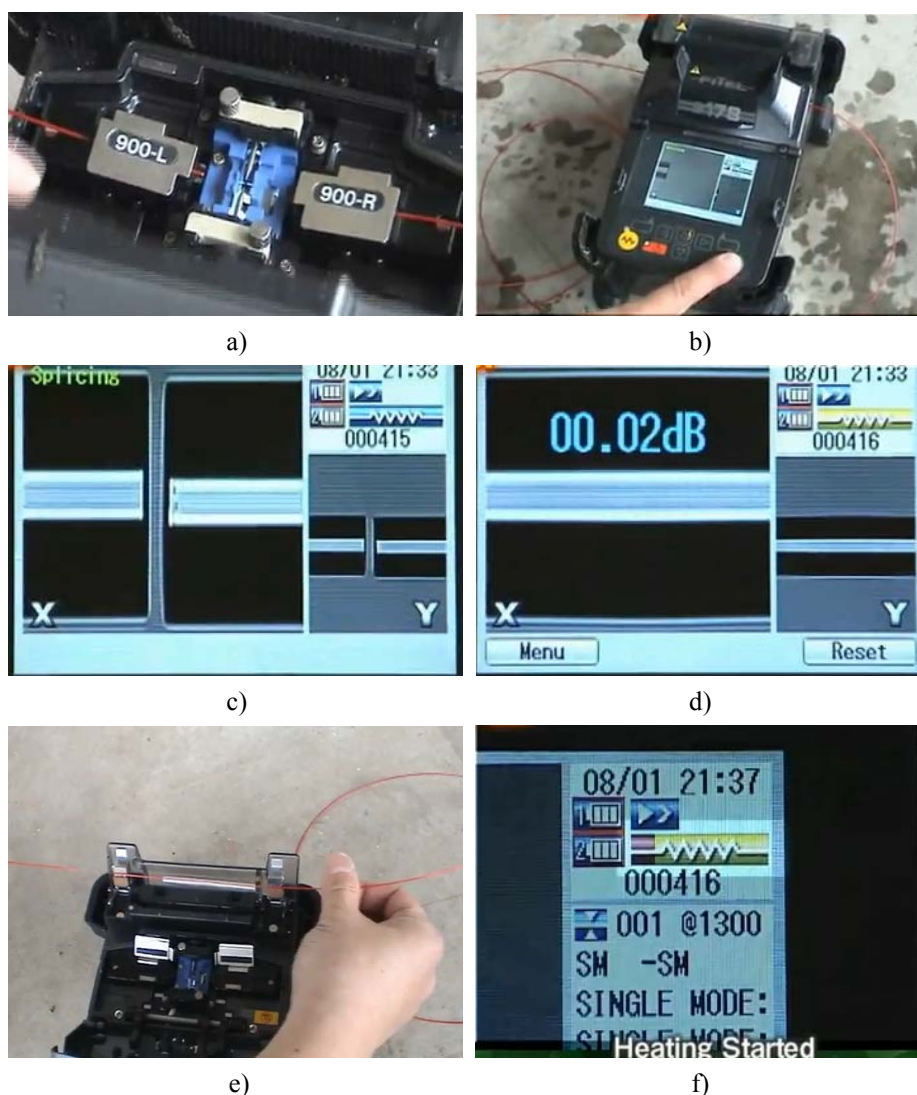
Porovnanie:

- „V“ drážka - malý prístroj, nízka cena, spoliehanie sa na kvalitu vlákna.
- LIDS - relatívne nákladný, iba pre 250 nm vlákna, poloautomatický.
- PAS - plnoautomatický, pre všetky typy vlákien, nákladný.

D. Proces zvárania:

Vlákna sú po príprave vložené do zváracieho zariadenia (Obr. 2a) a následne je spustený samotný proces zvárania (Obr. 2b), pričom vzniká pevný homogénny spoj.

Keď sú vlákna automaticky zosúosené jadrami (Obr. 2c), zapáli sa elektrický oblúk a to energiou dostatočnou na to, aby vlákna zmäkli, no ich konce sa nedeformujú. Oblúk horí presne stanovený čas (7 s.), po ktorom sa vlákna priblížia k sebe a keď sa dotknú, energia oblúka sa



Obr. 2 Proces spájania optických vlákien

zvýši tak, aby sa vlákna zliali do seba (Obr. 2d). Pre tvorbu mechanicky pevných spojov sa vlákna pri zlievaní ešte posúvajú oproti sebe (akoby sa prekrývali, prekrytie je cca 5 μm).

Zvarené vlákna sú chránené ochranným „rukávom“ (Obr. 2e), ktorý je následne zohriatím zmrštený (Obr. 2f), a tak pevne a nepriedušne obopína zvar (Obr. 3).

Týmto typom zvárania optických vlákien sa dosahuje nízky vložený útlm (podľa druhu vlákien: SM: 0,02dB; MM: 0,01dB; DSF: 0,04dB; NZD: 0,04dB) a vysoké potlačenie spätného odrazu (viac ako 70 dB).



Obr. 3 Hotový zvar optických vlákien

IV. ZÁVER

V tomto príspevku je popísaný postup zvárania optických vlákien pomocou zväračky a lámačky Fitel S178. Hoci je tento typ zväračky podstatne menší a ľahší ako iné modely, jeho dizajn (odolné kovové telo a gumové rohy) poskytuje dostatočne robustnú ochranu, čo umožňuje jej použitie nielen v laboratórnych, ale aj v náročných poľných poveternostných podmienkach (dážď, sneh, atď.) bez obmedzenia pracovného výkonu. Spolu s vylepšením riadenia spotreby umožňuje komfortne urobiť bez nabíjania až 200 zväracích cyklov (zváranie / zmršťovanie), čo je dostatočný počet pre jej využitie, ako v našich podmienkach na KEMT FEI TUKE, tak aj v komerčných podmienkach Slovak Telekom-u, a.s.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] <http://www.fiberoptichardware.com/fusion-splicers/fitel-s178-splicer.html>
- [2] <http://www.3sae.com/products/fitels178.htm>
- [3] <http://www.telonix.ca/weblit/Fitel/s178a.pdf>
- [4] <http://www.laser2000.de/index.php?id=373330>
- [5] <http://ofofof.en.made-in-china.com/offer/RojJDpSOZgrq/Sell-Fitel-S178-Fusion-Splicer.html>

Experimentálne pracovisko pre meranie na optických vláknach pomocou OTDR

¹Ján TURÁN, ²Luboš OVSENÍK, ³Ludovít HINTOŠ

^{1,2}Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

³Slovak Telekom, a.s., Poštová 19, 042 10 Košice, Slovenská republika

jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstrakt — V tejto práci je stručne opísané experimentálne pracovisko s OTDR (Optical Time Domain Reflectometry), t.j. kompaktná laboratórna platforma pre viacvrstvové multimédiové testovanie. Sú tu popísané OTDR merače EXFO FTB-200 a AXS-110 pre meranie rôznych typov optických vlákien a optických prvkov na rôznych vlnových dĺžkach.

Kľúčové slová — optická časová reflektometria, OTDR, meranie optických vlákien, EXFO FTB-200, EXFO AXS-110

I. ÚVOD

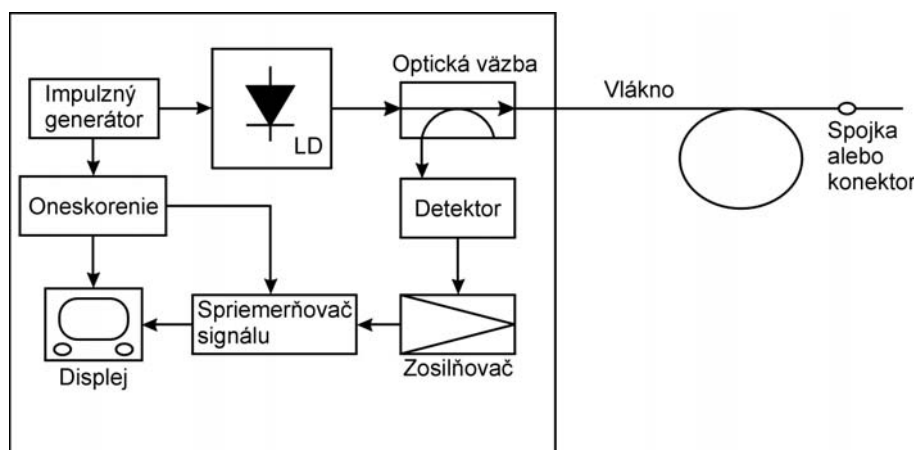
Meranie optických vlákien (OV) možno vo všeobecnosti rozdeliť do troch základných skupín:

- Meranie prenosových charakteristík (tlmenie, disperzia, šírka pásma);
- Meranie geometrických a optických charakteristík (priemer jadra a plášťa, excentricita jadra, eliptičnosť jadra a pod., numerická apertúra, profil indexu lomu, atď.);
- Meranie mechanických charakteristík (ťažná sila, polomer ohybu, tlaková sila, kritická vlnová dĺžka, atď.).

Údaje o týchto veličinách sú zvyčajne poskytované výrobcom OV, preto sa merania na OV väčšinou realizuje v priebehu výroby u výrobcu. Tieto údaje potom slúžia konštruktérom optických vláknových komunikačných systémov ako podklady pre optimálnu voľbu OV pre danú aplikáciu. No aj užívatelia musia mať možnosť uskutočniť (aj v poľných podmienkach) základné merania optických trás. Na to slúžia hlavne systémy OTDR.

II. OPTICKÁ ČASOVÁ REFLEKOMETRIA – OTDR

Je to univerzálna, v praxi často používaná, výhodná metóda merania optických vlákien (OV), optických káblov a aj celých trás optických vláknových komunikačných aj senzorových systémov.



Obr. 1 Bloková schéma optického vláknového časového reflektometra (OTDR)

Táto metóda umožňuje:

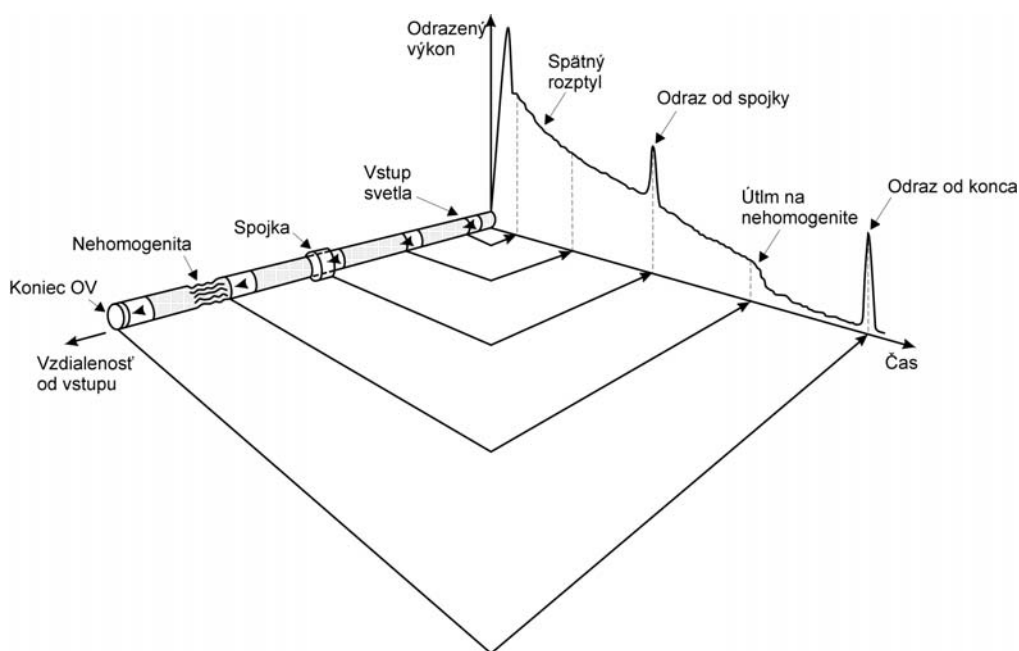
- Určiť dĺžku OV;
- Lokalizovať nehomogenity a porušenia OV;
- Meranie tlmenia OV, spojok a konektorov;
- Meranie vzdialenosti nehomogenity, resp. porušenia OV od jeho vstupu.

OTDR využíva rovnaký princíp merania ako klasická elektronická časová reflektometria (Obr. 1). Svetelné impulzy z lasera sa vedú do optického vlákna a rozptýlené svetlo, ktoré sa vracia z optického vlákna, prechádza cez optický väzobný prvok a dopadá na fotodetektor. Spätne rozptýlený optický signál možno rozdeliť na dve časti [1]:

- Svetlo rozptýlené diskretnými nehomogenitami OV (vstupné čelo OV, koniec OV, optická spojka alebo konektor, prerušenie OV a pod.);
- Svetlo, ktoré vzniká ako dôsledok distribuovaného (spojitého) rozptylu (napr. Rayleighov rozptyl, Mieho rozptyl a pod.).

Okrem toho sa časť optického signálu, šíriaceho sa OV, absorbuje. V praxi (v prípade, že svetlo má vhodnú vlnovú dĺžku) sa straty absorpciou malé v porovnaní so stratami spätným rozptylom.

V súčasnosti sú dostupné komerčne vyrábané prístroje OTDR, ktoré pracujú v oblasti vlnových dĺžok menších ako 1000 nm. Tieto prístroje emitujú do OV sériu krátkych svetelných impulzov (10 až 100 ns) s výkonom 100 až 500 mW a analyzujú odrazené svetlo. Typický odrazený signál OTDR s príslušnými nehomogenitami meraného OV je znázornený na Obr.2. Komerčné OTDR môžu v závislosti od tlmenia OV lokalizovať jeho nehomogenity do vzdialenosti 10 až 15 km s presnosťou ± 1 až ± 2 m. Nevýhodou komerčných OTDR je malý dynamický rozsah <40 dB, ktorý je možné metódou počítania fotónov zvýšiť až na 80 dB. Presnosť odčítania hodnôt tlmenia je ± 10 % [1].



Obr. 2 Nehomogenity OV a odrazený signál OTDR

III. OTDR EXFO FTB-200

FTB-200 je kompletná kompaktná platforma OTDR, vhodná na viacvrstvové testovanie aplikácií zahŕňajúcich ako optické vláknové (SM - Single Mode, MM – Multi Mode), tak aj metalické prenosové médiá. Merací prístroj FTB-200 bol navrhnutý pre supertechnologické oblasti vyžadujúce kompaktnú platformu optimalizovanú pre efektívne automatizované testy a hĺbkové analýzy riešenia problémov FTTx sietí. Je ideálny pre testovanie:

- Pri výstavbe a inštalácii sietí FTTx:
 - Meranie v oboch smeroch na rôznych (850/1300/1310/1490/1550/1625/1650 nm) vlnových dĺžkach;
 - Meranie vloženého útlmu prenosového média;
 - Meranie útlmu odrazu – nutné, pokiaľ chceme prenášať video;
 - Meranie útlmu konektorov a spojok;

- Meranie dĺžky trasy.
- Pri aktivácii a poskytovaní služieb sietí FTTx;
- Pri údržbe a riešení problémov prevádzky sietí FTTx.

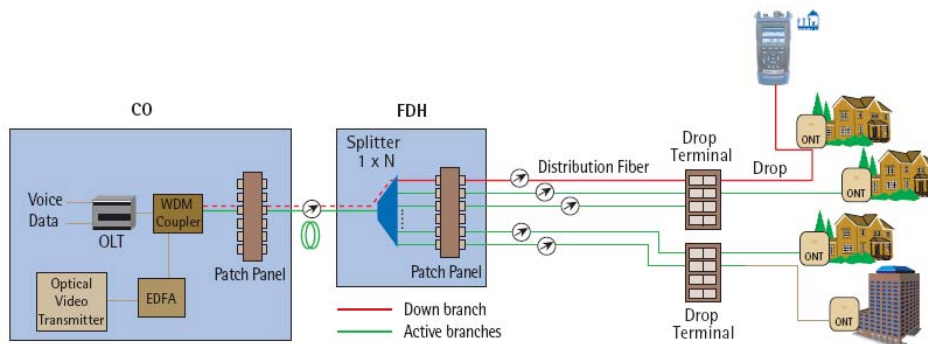


Obr. 3 OTDR EXFO FTB-200

Dá sa použiť na testovanie fyzickej, transportnej, ale aj dátovej vrstvy FTTx, tzn. že je možné vyskúšať najnovšie štandardy Ethernet a SONET/SDH/OTN sietí [2,4,6]. FTB-200 možno použiť tiež pre testovanie kvality služieb (QoS) Gbit a 10 Gbit Ethernet-ových liniek (Obr. 3).

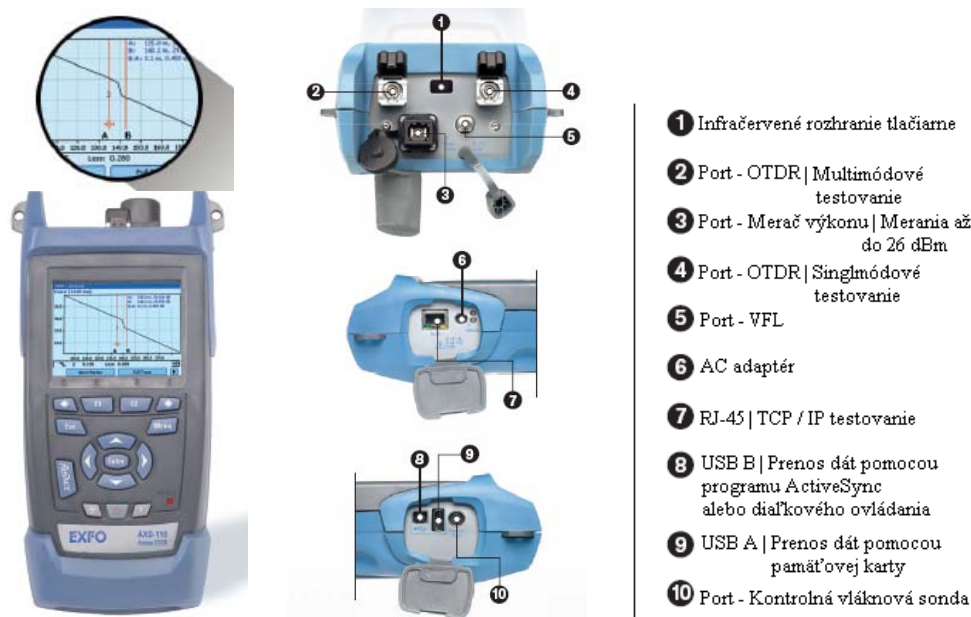
IV. OTDR EXFO AXS-110

Je výkonný ručný OTDR prístroj, určený pre detekovanie a meranie problémov spliterov až 1x32 (optické vláknové rozdeľovače/odbočnice) v sieťach FTTx. Ponúka meranie na rôznych vlnových dĺžkach (850/1300/1310/1490/1550/1625 nm) pre jednovidové (SM) aj mohovidové (MM) optické vlákna. Kombinuje špičkové technológie OTDR s funkčnosťou merača optického výkonu v rámci jednej výkonnej jednotky a je optimalizovaný pre point-to-point testovanie pasívnych optických sietí (PON) v rámci architektúry FTTx [3,5].



Obr. 4 Použitie AXS-110

Modely OTDR typu AXS-110 sú veľmi vhodné na meranie vo vnútro podnikových aplikáciách (Obr.4), pretože dokážu zaznamenať mimoriadne udalosti už v 0,8 m mŕtve zóne optickej trasy, čo umožňuje ľahko lokalizovať a charakterizovať všetky akcie medzi optickým vysielateľom a ústredím, t.j. FTTH sieťové aplikácie, kde sú udalosti zvyčajne blízko pri sebe.



Obr. 5 OTDR EXFO AXS-110

Ak sa meranie nevykonáva v laboratórnych podmienkach, nemôžeme si dovoliť strácať drahocenný čas na určovanie optimálnych nastavení OTDR. Preto sú tieto typy OTDR schopné optimalizovať všetky parametre potrebné na splnenie potrieb užívateľa a tým aj samotné meranie a kvalitu detekcie udalostí na optickej trase (nehomogenity OV, spájanie OV, atď.). Sú to prístroje schopné dlhodobej práce (8 hodín bez nabíjania) v teréne s možnosťou zapamätania si až 500 posledných meraní a umožňujú jednoduchý prenos nameraných dát pomocou USB portu (Obr. 5).

V. ZÁVER

Z reflektometrických metód je v súčasnosti najviac rozpracovaná a používaná metóda OTDR. Jej najväčšou výhodou je univerzálnosť, možnosť merania rôznych typov mnohovidových aj jednovidových OV a široká dostupnosť komerčne vyrábaných reflektometrov. Preto sa aj v našich podmienkach na KEMT FEI TUKE, tak ako aj v komerčných podmienkach Slovak Telekom-u, v hojnej miere využívajú tieto metódy na kontrolu a testovanie optických liniek.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] J. Turán, *Optoelektronika*. Košice: Harlequin, 2002, ch. 9.
- [2] <http://exfo.fiberoptic.com/FTB-200.htm>
- [3] <http://exfo.fiberoptic.com/AXS-110.htm>
- [4] <http://www.exfo.com/en/Products/Products.aspx?Id=334>
- [5] <http://www.exfo.com/en/Products/Products.aspx?Id=470>
- [6] http://www.dist-tron.com/images/FTB-200_compact_platform.pdf

DCT vizuálneho objektu s predspracovaním hraničných blokov

Iveta GLADIŠOVÁ, Ján MIHALÍK

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

iveta.gladisova@tuke.sk, jan.mihalik@tuke.sk

Abstrakt — V článku je popísaný princíp predspracovania obrazových objektov a následná transformácia obrazu po blokoch obrazových prvkov modifikovanou dvojrozmernou diskretnou kosínusovou transformáciou. Získané výsledky tohto druhu transformácie sú porovnávané s klasickou metódou transformácie pre ľubovoľný vstupný obraz. V závere je vyhodnotenie výsledkov efektivity kódovania, meranie vplyvu zanedbania určitého počtu transformačných koeficientov na kvalitu späťne transformovaných obrazov.

Kľúčové slová — diskretná kosínusová transformácia, kompresia, MPEG-4, extrapolácia, filtrácia

I. ÚVOD

Z dôvodu rýchleho rozvoja technológie sa v oblasti prenosu videa kladú čoraz väčšie nároky na jeho efektívne kódovanie. Účinnéjšie kompresné metódy sú potrebné na zlepšenie kvality služieb, ako sú ukladanie videa na prenosné médiá, digitálne vysielanie televízie, alebo pri videotelefónii. V súčasnosti, s postupným vývojom technológie prichádza vo videokomunikáciách k nástupu televízie s vysokým rozlíšením HDTV. Táto zmena si vyžaduje nasadiť účinnejšie kompresné metódy na zníženie prenosovej rýchlosti a zvýšenie kvality obrazu, aby sa dala existujúca kapacita efektívnejšie využiť. V praxi nachádza typ spracovania obrazovej informácie použitím diskretnéj kosínusovej transformácie (DCT) využitie pri moderných prístupoch kompresie videa, v ktorých sa využíva výber vizuálnych objektov z obrazovej scény. Tieto objekty sú určené na spracovanie definovanými algoritmami a na následné zakódovanie do transformovaného tvaru. DCT je jednou z najčastejších používaných transformácií v oblasti kompresie a kódovania obrazu, využívajú ju medzinárodné štandardy JPEG, H.263, MPEG-2, MPEG-4 [1] - [4].

II. VÝBER VIDEOOBJEKTOVEJ ROVINY BINÁRNÝM TVAROM A PREDSPRACOVANIE HRANIČNÝCH MAKROBLOKOV

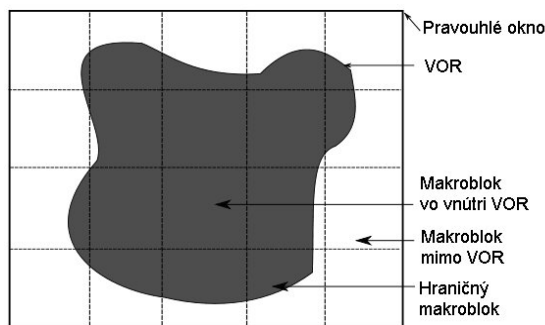
Príchodom štandardu videokompresie MPEG-4 súčasne vznikla nová metóda kompresie obrazu pomocou výberu vizuálnych objektov z obrazovej scény. Nové algoritmy, ktoré vo všeobecnosti kódujú jednotlivé videoobjekty namiesto celých obrazových rámcov, podporujú rozvoj aplikácií s obsahovou interakciou. Z vizuálnej scény sa vyberajú videoobjekty, ktoré sa môžu kódovať rôznymi technikami. V koncepcii videoobjektových rovín je každá snímka vstupnej videosekvencie segmentovaná do určitého počtu ohraničených oblastí (videoobjektových rovín), pričom táto oblasť môže pokrývať určitú časť snímky, resp. určitý fyzický objekt v scéne. Tento systém je bližší ľudskému vnímaniu obrazu a umožňuje efektívnu reprezentáciu videoobjektov, manipuláciu s nimi, editovanie v ich dátových tokoch, ako aj možnosť škálovania. Základom tohto systému je rozdelenie obsahu videa na videoobjekty [5], [6], [7].

Vizuálna scéna obsahuje viacej videoobjektov, ktoré môžu byť zoskupované do niekoľkých videosekcií. Videoobjekt je zostavený z vrstiev za účelom jeho kódovania, ktorého syntax umožňuje rekonštrukciu videoobjektu zo základnej vrstvy a pridávanie ďalších doplnkových vrstiev za účelom zlepšenia jeho kvality. Každý videoobjekt je charakterizovaný pomocou časovej a priestorovej informácie. V niektorých aplikáciách nie je potrebné rozlišovať videoobjekty kvôli zložitosti ich generovania a rozsiahlej prídavnej informácie. V tomto prípade sa kódujú pravouhlé snímky, ktoré predstavujú degenerované vyjadrenie tvaru videoobjektu.

Pre účely transformácie a spracovania je obrazová rovina rozdelená na matice obrazových prvkov (OP), na makrobloky s veľkosťou 8x8 alebo 16x16 [8]. Aplikáciou binárneho tvaru na pôvodný obraz sa určuje, ktorý makroblok je vo vnútri VOR, mimo VOR, a ktoré makrobloky sú na hranici VOR t. j. obsahujú OP vo vnútri VOR, aj OP nachádzajúce sa mimo VOR (Obr. 1) [5], [11].

Pre každý z týchto troch prípadov je nastavené iné spracovanie tohto makrobloku.

- **Makrobloky vo vnútri VOR** sa transformujú dvojrozmernou (2R) DCT.
- **Hraničné makrobloky** sa najprv predspracujú dolno-priepustnou filtráciou a extrapoláciou vzoriek a až potom sa transformujú 2R DCT.
- **Makrobloky mimo VOR** sa neprenášajú, ich ďalšie spracovanie je irelevantné.

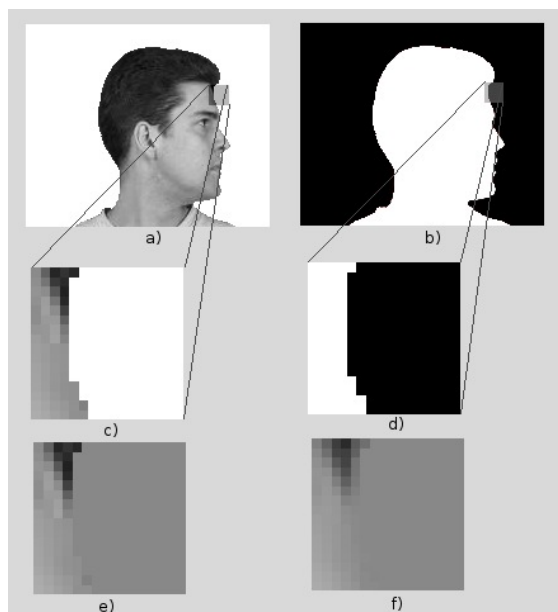


Obr. 1 Rozdelenie makroblokov podľa príslušnosti ku VOR

Na OP patriace do hraničných makroblokov sa aplikuje predspracovanie pozostávajúce z dvoch fáz:

1. **Extrapolácia:** v prvom kroku sa vypočíta priemerná hodnota OP vo vnútri VOR v danom hraničnom makrobloku a táto hodnota sa priradí všetkým prvkom mimo VOR v hraničnom makrobloku, vid' Obr. 2e.

2. **Dolno-priepustná filtrácia:** všetky prvky v hraničnom bloku nadobudnú hodnotu vypočítanú ako priemerná hodnota ich 4 susedných OP vo vertikálnom a horizontálnom smere. Ak je OP na okraji makrobloku priemerná hodnota sa počíta z menšieho počtu zodpovedajúcich susedných OP v danom makrobloku, vid' Obr. 2f.



Obr. 2 Predspracovanie hraničného bloku a) c) výber bloku zo vstupného obrazu; b) d) výber bloku z binárneho tvaru; e) extrapolácia bloku; f) DP filtrácia

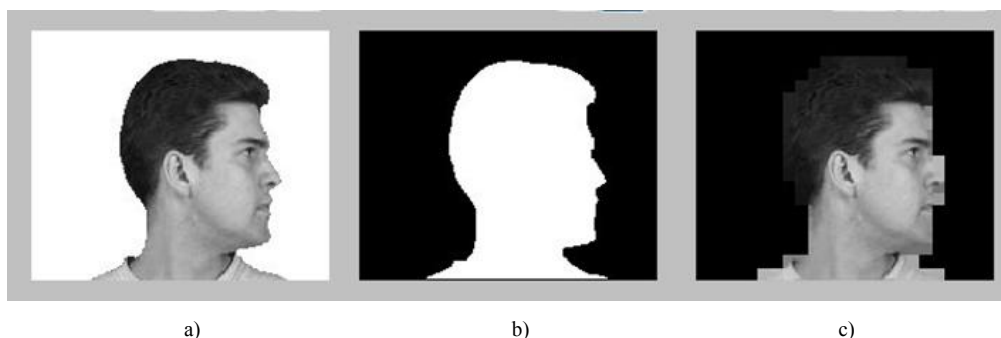
III. ZHODNOTENIE DOSIAHNUTÝCH VÝSLEDKOV SIMULÁCIE

Modifikovaná DCT [9] s predspracovaním obrazu sa simulovala navrhnutými programovými prostriedkami pre ľubovoľné vstupné obrazy, ktoré majú tieto štandardné parametre: *RGB formát obrazových dát, 8-bitové rozlíšenie jednej zložky RGB obrazových dát, riadky a stĺpce*

rastra rozlíšenia by mali byť deliteľné 16 (v opačnom prípade sa zo vstupného obrazu vyberie jeho najväčšiu možnú časť, ktorá spĺňa túto podmienku a nadbytočné riadky a stĺpce sa nespracúvajú t.j. vykoná sa orezanie obrazu). Ďalej, vstupné obrazy by mali byť uložené v jednom z týchto štandardných formátov obrazu: jpg, bmp, tiff, png, gif, pbm, ras. Ak je načítaný farebný obraz, tak sa najprv prekonvertuje na šedý obraz spriemernením zložiek RGB na hodnoty od 0 po 255 [10]. Binárny tvar je čierno-biely obraz, kde čierna farba vymedzuje oblasť mimo VOR a biela oblasť vo vnútri VOR. Ak sa načíta šedý obraz, tak všetky OP s hodnotami 0-127 určujú oblasť mimo VOR a OP s hodnotami 128-255 oblasť vo vnútri VOR. Formát uloženého obrazu by mal spĺňať podmienky ako pri vstupnom obraze.

Obraz sa transformuje po blokoch štvorcových matic, pričom je možnosť nastavenie voľby veľkosti týchto blokov na 8x8, 16x16, tomu následne zodpovedá aj veľkosť jadra transformácie. Simulačný program je určený pre dvojrozmernú transformáciu, primárne pre 2R DCT, ale môže pracovať s ľubovoľnou transformačnou maticou. Jedinou podmienkou je, aby použité transformačné jadrá boli ortonormálne. Pri kompresii dát sa využíva skutočnosť, že DCT dekoreluje signál a koncentruje informačnú energiu do ľavej hornej časti matice transformačných koeficientov (koeficienty s najnižšími hodnotami energie je možné zanedbať, filtrovať, pričom nedochádza k veľkej degradácii obrazovej informácie). Program umožňuje nastaviť pri filtrácii počet ponechaných transformačných koeficientov s vysokými hodnotami energie.

Vykonalo sa porovnanie účinnosti kompresie modifikovanej a klasickej 2R DCT na kvalitu spätne dekódovaného obrazu. Na Obr. 3 je zobrazený simulovaný vstupný obraz, jeho binárny tvar a vybraný VO s hraničnými blokmi pred spracovaním modifikovanou 2R DCT. Vyhodnocovala sa miera kvality transformácie pomocou pomeru špičkovej hodnoty signálu ku kvantizačnému šumu t.j. PSNR (Peak Signal to Noise Ratio). Pri rozdelení obrazu na bloky OP s veľkosťou 8x8 sa zisťovali hodnoty PSNR závislé od počtu použitých transformačných koeficientov (1 až 64) a pri rozdelení obrazu na bloky s veľkosťou 16x16 sa menil počet použitých transformačných koeficientov od 1 do 256.



Obr. 3 a) Vstupný obraz, b) jeho binárny tvar,
c) vybraný VO s hraničnými blokmi pred spracovaním modifikovanou 2R DCT

V Tab. 1 sú uvedené dosiahnuté hodnoty PSNR v decibeloch spätne transformovaného obrazu, pri veľkosti bloku 8x8 bez predspracovania (klasická transformácia) a s predspracovaním (modifikovaná transformácia), pričom sa menil počet použitých transformačných koeficientov pri filtrácii obrazu. Z výsledkov vyplýva, že modifikovaná DCT s predspracovaním dosahuje lepšiu PSNR pri čo najväčšej filtrácii transformačných koeficientov, napr. pri použití iba štyroch transformačných koeficientov z bloku 8x8 je PSNR o 2,3 dB lepšia, ako pri klasickej DCT bez predspracovania. Pri zvyšovaní počtu použitých transformačných koeficientov na 8, 16 a 32 je PSNR s predspracovaním lepšia približne o 1,5 dB oproti PSNR bez predspracovania. Pri počte 43 použitých transformačných koeficientov dochádza k vyrovnaniu pomeru PSNR pri oboch metódach transformácie a pri ďalšom zvyšovaní tohto počtu koeficientov je už klasická DCT lepšia oproti modifikovanej transformácii.

Podobné výsledky sa dosiahli pri spracovaní obrazu po blokoch 16x16 (Tab. 2). Aj tu je modifikovaná DCT s predspracovaním lepšia oproti klasickej DCT bez predspracovania, ak sa pri filtrácii odstráni čo najväčší počet spektrálnych koeficientov s nízkymi energiami. Pri použití 4 transformačných koeficientov je PSNR s predspracovaním o 3,1 dB lepšia, ako PSNR bez predspracovania, pri 8 použitých koeficientoch je lepšia o 2,5 dB a pri ďalšom zvyšovaní počtu koeficientov je lepšia približne o 1,5 dB. Pri použití od 150 a viac transformačných koeficientov z bloku už klasická DCT dosahuje vyššiu PSNR oproti modifikovanej DCT.

TAB. 1
 VEĽKOSŤ BLOKU 8x8

Počet použitých transf. koeficientov	4	8	16	32	40	48	56	64
PSNR [dB] bez predspracovania	25,5	29	31,7	36,2	38	41,2	45,2	infinity
PSNR [dB] s predspracovaním	27,8	30,3	33,2	37,6	38,6	39,6	39,6	39,5

TAB. 2
 Veľkosť bloku 16x16

Počet použitých transf. koeficientov	4	8	16	32	64	128	192	256
PSNR [dB] bez predspracovania	21	23,8	27	29,6	32,6	36,0	42,1	infinity
PSNR [dB] s predspracovaním	24,1	26,3	28,5	30,9	34,0	37,4	38,7	38,6

IV. ZÁVER

Článok sa venoval problematike diskretnej kosínusovej transformácii vizuálnych objektov, výberu vizuálnych objektov z obrazovej scény a možnosťou predspracovania tejto informácie procesmi dolnopriepustnej filtrácie a extrapolácie obrazových prvkov. Pre účely porovnania kvality kódovania boli vytvorené univerzálne programové prostriedky, ktorými sa sledovala miera degradácie kvality obrazu v závislosti od počtu aktívnych transformačných koeficientov pri prenose signálu. Porovnávali sa vlastnosti modifikovanej a klasickej 2R DCT na vybraný vstupný testovací obraz.

Z výsledkov merania ako aj z porovnania veľkostí hodnôt transformačných koeficientov vyplýva, že použitím modifikovanej 2R DCT s predspracovaním dochádza k oveľa väčšej dekorelácii signálu ako pri použití klasickej 2R DCT. Dôvodom je nielen pridaná dolnopriepustná filtrácia, ale hlavne extrapolácia OP zahrnutá v predspracovaní pred samotnou transformáciou. Pri porovnaní kvality kódovania obrazového signálu – závislosť PSNR od počtu použitých transformačných koeficientov má modifikovaná transformácia lepšie výsledky pri nastavenej väčšej informačnej kompresii prenosu obrazu (viac transformačných koeficientov je zanedbaných) ako klasická transformácia. Predspracovanie obrazu má vplyv na hodnoty PSNR, avšak spôsobený rozdiel ľudské oko ťažko postrehne, parameter PSNR nemusí zohľadňovať subjektívne vnímanie obrazu divákom.

V článku popisovaný modifikovaný typ dvojrozmernej diskretnej kosínusovej transformácie vizuálnych objektov má veľký potenciál využitia v moderných prístupoch spracovania videa. Dôvodom je, že sa v súčasnosti kladie väčší dôraz na kódovanie obrazu s výberom objektov z obrazovej scény a tento typ modifikovanej diskretnej kosínusovej transformácie obrazu je vhodným kandidátom pre efektívne spracovanie obrazových scén tohto druhu.

LITERATÚRA

- [1] J. Mihalík, „Štandardný videokodek MPEG-4“. Slaboproudý obzor, roč. 60, č.2, s.7-11, 2003.
- [2] J. Mihalík, „Štandardný videokodek H.263 a jeho porovnanie s H.261“. Slaboproudý obzor, roč. 60, č.1, s.1-7, 2003.
- [3] I. Richardson, „H.264 / MPEG-4 Part 10 : Transform & Quantization“, 2007.
http://www.rgu.ac.uk/files/h264_transform.pdf
- [4] MPEG Requirements Group, “MPEG-4 requirements”, Doc. ISO/IEC JTC1/SC29/WG11 N1595, Sevilla MPEG meeting, February 1997.
- [5] J. Mihalík, „Kódovanie obrazu vo videokomunikáciách“. TU-FEI, 2001.
- [6] P. Correia – F. Pereira, „Coding Video Objects with the Emerging MPEG-4 Standard“. May 1997.
- [7] ISO/MPEG N2725, „Mpeg-4 Overview“. Seoul, 1999.
- [8] I. Gladišová, J. Mihalík, J. Zavacký, „Obrysová reprezentácia a kódovanie binárneho tvaru videoobjektu“. Slaboproudý obzor, roč. 66, č.3, s.14-18, 2010.
- [9] T. Chen, H. R. Wu, and B. Qiu, “Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts”. IEEE Trans. Circuits Syst. Video Technol, vol.11, pp. 594-604, May 2001.
- [10] F. X. Coudoux, M. Gazelet, and P. Corlay, “An Adaptive Postprocessing Technique for the Reduction of Color Bleeding in DCT-Coded Images”. IEEE Trans. Circuits Syst. Video Technol, vol. 14, pp. 114-121, Jan 2004.
- [11] T. Sikora, „Trends and Perspectives in Image and Video Coding“. Proc. of the IEEE, vol.93, pp. 6-17, Jan. 2005.

CAN zbernica vo vložených aplikáciách na báze ARM procesorov STM32

Martin PETRVALSKÝ, Miloš DRUTAROVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

martin.petrvalsky@student.tuke.sk, milos.drutarovsky@tuke.sk

Abstrakt — Príspevok sa zaoberá problematikou využitia CAN zbernice ako aj protokolu CANopen vo vložených aplikáciách na báze procesorov STM32 s jadrom ARM Cortex-M3. Sú opísané základné hardvérové prostriedky, CAN zbernica, protokoly CANopen a MicroCANopen ako aj základná možnosť využitia operačného systému reálneho času vo vložených aplikáciách. Vytvorené programové prostriedky pre vývojové dosky demonštrujú využitie protokolu CANopen na vzájomnú komunikáciu viacerých vložených systémov. Ako vývojové nástroje sú využité vývojové prostredie Keil μ Vision s externým GNU prekladačom firmy CodeSourcery a vývojové dosky od firiem ST Microelectronics, Keil ako aj experimentálna doska vyvinutá na Katedre elektroniky a multimediálnych telekomunikácií Technickej univerzity v Košiciach.

Kľúčové slová — CAN zbernica, STM32, Cortex-M3, CANopen, MicroCANopen, CoCoX CoOS

I. ÚVOD

Protokoly CANopen [10] a MicroCANopen sú na trhu už niekoľko rokov podobne ako aj mikrokontroléry STM32, avšak ich kombinácia je veľmi málo prepracovaná a ak existuje, tak nie je voľne šíriteľná a je neprístupná pre výukové účely. Situácia medzi operačnými systémami reálneho času je iná. Niekoľko rokov existuje napríklad voľne šíriteľný operačný systém FreeRTOS, ktorý je kvalitne prepracovaný a využívaný v mnohých aplikáciách. Iba nedávno však vznikol nový systém s názvom CoCoX CoOS [8], ktorý je oproti všeobecne použiteľnému FreeRTOS špecializovaný a vysoko optimalizovaný pre mikroprocesory ARM s jadrom Cortex-M3.

Cieľom príspevku je najskôr v stručnosti oboznámiť čitateľa s architektúrou mikroprocesorov Cortex-M3 [3] a s radom mikrokontrolérov STM32 [2], ktoré využívajú procesor s jadrom Cortex-M3. Krátko bude vysvetlený princíp činnosti CAN zbernice. Ďalej budú podrobne objasnené základné princípy štandardov CANopen a MicroCANopen, ich vzájomný vzťah ale aj odlišnosti. Samostatná kapitola je venovaná operačnému systému reálneho času CoCoX CoOS, ktorá opisuje jeho funkcie a použitie. Následne sú stručne opísané hardvérové a softvérové prostriedky, ktoré boli využité pri vývoji softvéru a jeho testovaní. Hlavným cieľom príspevku je opis vývoja aplikácií. Prvá aplikácia demonštruje jednoduchú komunikáciu založenú na štandarde CANopen. Druhá aplikácia overuje hlavné periférie vývojovej dosky KEMT STM32 [12].

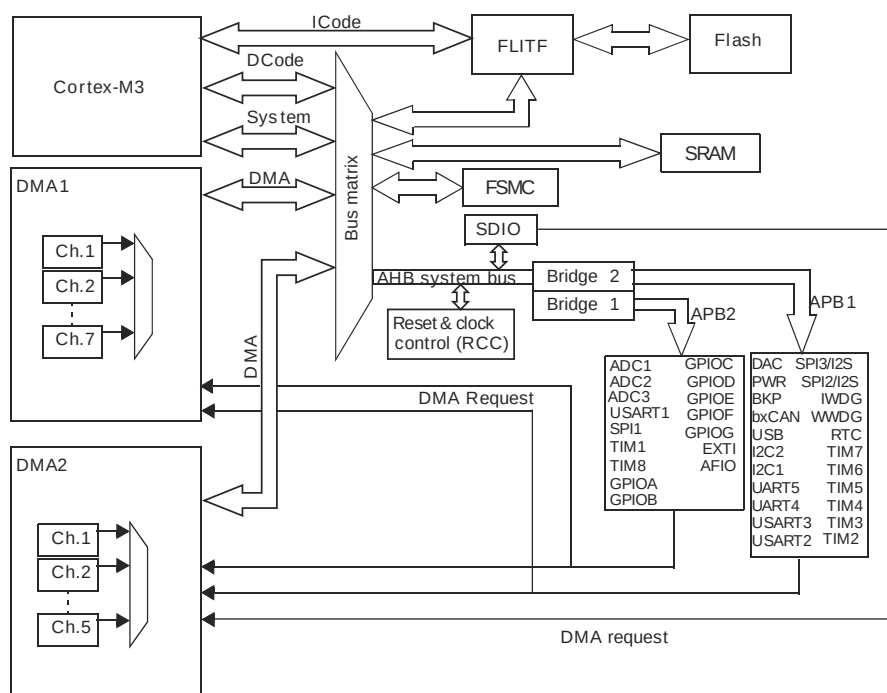
Hlavným výsledkom je jednoduchá fungujúca sieť, postavená z vývojových dosiek s mikrokontrolérmi STM32, ktorá využíva CANopen pre vzájomnú komunikáciu. Na základe tejto siete bude možné v budúcnosti vyvíjať zložitejšie aplikácie pre praktické použitie vložených systémov na báze STM32.

II. SKUPINA MIKROKONTROLÉROV STM32

Mikrokontroléry STM32 od firmy ST Microelectronics sú postavené na báze procesorov ARM Cortex-M3. Delia sa na niekoľko skupín podľa zamerania (Performance Line, Value Line, Connectivity Line a Access Line) alebo podľa použitých veľkostí interných pamätí FLASH a SRAM. FLASH pamäte sú v rozsahu od 16 kB do 1 MB a SRAM od 4 kB do 96 kB. Tieto mikrokontroléry sa vyrábajú s veľkým množstvom dnes už takmer štandardných periférií ako napríklad všeobecne použiteľné časovače, USART, USB, SPI, I2C, CAN, duálne ADC a RTC, ktoré komunikujú medzi sebou a s jadrom procesora cez systém zberníc znázornených na

Obr. 1. Každá z týchto periférií má bohaté funkcie, rôzne módy a nastavenia. STM32 obsahujú až 7 DMA kanálov, ktoré dokážu preniesť dáta z a do každej periférie používajúc 8/16 alebo 32-bitové slová.

Posledná dôležitá vlastnosť je to, že okrem výkonných módov, dokážu mikrokontroléry pracovať aj vo viacerých úsporných režimoch. Navyše umožňujú rýchle a efektívne prepínanie medzi týmito módmí, čo prispieva k šetreniu energie v reálnych aplikáciách.

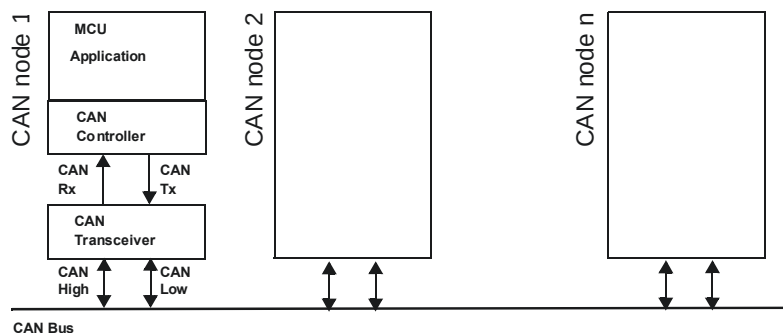


Obr. 1 Bloková schéma STM32 skupiny Performance Line [2], ktorá znázorňuje jednotlivé periférie a ich vzájomné prepojenie pomocou dátových zberníc.

III. CAN ZBERNICA

CAN zbernica bola pôvodne navrhnutá firmou Bosch pre automobilový priemysel, kde veľké množstvo senzorov často posiela krátke správy. Vyznačuje sa veľkou spoľahlivosťou, ktorá je zabezpečená CRC kódom v každej správe, ktorý musí vždy potvrdiť každý uzol. Ak jeden z uzlov zaznamená chybu v CRC súčte, tak všetky uzly správu ignorujú a tá sa znovu automaticky preposiela. Kvôli týmto výhodným vlastnostiam sa CAN zbernica rozšírila aj do iných oblastí ako sú napríklad automatizácia v priemysle, automatizácia domácností, robotika a iné.

Typické prenosové médium CAN zbernice je stočený pár vodičov, ktorého maximálna dĺžka závisí od požadovanej prenosovej rýchlosti, ktorá môže byť od 10 kbit/s pre dĺžku vodičov až 6 km po 1 Mbit/s pre komunikáciu na vzdialenosť do 40 m. Na CAN zbernici môže ľubovoľný uzol v ktoromkoľvek čase začať vysielanie, pričom konflikty vo vysielaní správ sa riešia podľa priority správ. Správe s nižším identifikátorom sa povolí vysielanie, ostatné správy musia počkať, kým nebude zbernica voľná. Tieto správy sú prijímané každým uzlom a ten následne rozhodne o tom, či ju akceptuje. Preto má väčšina CAN uzlov implementované rôzne techniky pre filtrovanie správ.



Obr. 2 Topológia CAN siete. Každý uzol na to, aby mohol komunikovať na CAN zbernici musí obsahovať mikrokontrolér s CAN perifériou. Navyše musí byť prítomný ešte budič, ktorý je priamo spojený so zbernicou, ktorá je tvorená stočenou dvojlinkou [2].

Radič CAN v procesoroch STM32 je plne vybavený CAN uzol, ktorý podporuje štandardy CAN 2.0A a 2.0B s prenosovými rýchlosťami do 1Mbit/s. Navyše podporuje protokol TTCAN, ktorý je využívaný v real-time aplikáciách. Radič je označený ako bxCAN, čo znamená, že disponuje tromi schránkami pre prijímanie správ a dvoma schránkami pre vysielanie správ. Každá z prijímacích schránok je usporiadaná ako FIFO (First In First Out, prvý dnu prvý von) zásobník pre 3 správy. Vysielacie schránky majú miesto len pre jednu správu a môžu byť zoradené buď ako FIFO alebo podľa identifikátora vysielanej správy [2].

Ďalšia dôležitá súčasť CAN radiča je filter prichádzajúcich správ, ktorý využíva 14 baniek. Každá z týchto baniek môže uložiť jeden až štyri identifikátory správ podľa toho, aký mód filtrácie sa pre danú banku použil. Môže byť použitý mód filtrácie alebo maskovania a to pre obidva formáty identifikačného čísla – základného a rozšíreného. Radič tiež zabezpečuje správu chýb na CAN zbernici. Veľkou výhodou je, že táto správa, filtrovanie správ ale aj ich prijímanie a vysielanie sa deje automaticky, bez zaťažovania jadra procesora.

IV. ŠTANDARDY CANOPEN A MICROCANOPEN

CAN vo svojej špecifikácii definuje len najspodnejšie vrstvy sieťovej architektúry a to fyzickú a linkovú. Vyššie vrstvy si môže dedefinovať užívateľ sám, avšak omnoho výhodnejšie z hľadiska kompatibility je použitie už vytvorených štandardov ako napr. J1939, DeviceNet, OSEK, CANopen a pod.

A. CANopen

CANopen je otvorený a voľne prístupný štandard. Nevyžaduje žiadne poplatky ani licencie a navyše môže byť rozšírený a upravený podľa aktuálnych požiadaviek. Veľmi často sa využíva v automatizácii. Z pohľadu OSI referenčného modelu CANopen implementuje sieťovú a vyššie vrstvy, pričom spodné vrstvy zabezpečuje väčšinou CAN zbernica, ale využívajú sa aj iné spôsoby komunikácie (napríklad Ethernet Powerlink alebo EtherCAT) [10]. Štandard CANopen pozostáva z týchto častí:

- Spôsob adresovania zariadení na sieti.
- Komunikačné protokoly, ktoré zabezpečujú obsluhu siete, monitorovanie a vzájomnú komunikáciu jednotlivých uzlov.
- Jednoduchá transportná vrstva pre rozdeľovanie a spájanie správ.
- Aplikačná vrstva definovaná profilom zariadenia.

Slovník objektov (z angl. Object Dictionary - OD) je základ pre komunikáciu pomocou CANopen. OD predstavuje tabuľku, v ktorej sa nachádza množstvo štandardizovaných položiek slúžiacich na vzájomnú komunikáciu a ovládanie jednotlivých uzlov siete.

Každá položka obsahuje 16-bitový index a 8-bitový subindex. Ďalej môže obsahovať premennú rôzneho typu a rôznej veľkosti. Normálne fungovanie uzla bude zabezpečené ak OD bude obsahovať minimálne položky, ktoré sú dané štandardom.

Medzi povinnými položkami sa v OD vyskytuje položka, v ktorej je udaný typ (profil) zariadenia. Definuje parametre komunikácie a položky OD, ktoré dané zariadenie podporuje. Existuje mnoho prednastavených profilov pre všeobecné I/O moduly, kodéry a iné zariadenia, ale je tu aj možnosť vytvoriť vlastné profily. Hlavný uzol potom po prečítaní profilu zariadenia vie určiť, ktoré položky OD sú definované pre konkrétne zariadenie a priamo k týmto položkám pristupuje. Hlavný uzol predstavuje zariadenie v sieti, ktoré kontroluje a monitoruje stav a správanie ostatných uzlov [6]. Pre prístup k OD vzdialených uzlov sa používajú nasledujúce protokoly:

- NMT (Network Management, správa siete)
- SDO (Service Data Object, servisný dátový objekt)
- PDO (Process Data Object, procesný dátový objekt)
- Ostatné - SYNC, TIME a EMCY

B. MicroCANopen

MicroCANopen (MCO) je minimálna a veľmi úsporná verzia štandardu CANopen. Oproti pamäťovým nárokom štandardu CANopen (viac ako 50 kB kódu a 1 kB RAM pamäte) zaberá MicroCANopen len 4 kB kódu a 200 B RAM pamäte, avšak má niekoľko obmedzení a odlišností:

- Originálne bol navrhnutý pre prácu bez hlavného uzla, ale môže ho využívať, zatiaľ čo pri CANopen je to presne opačne.
- Implementuje len protokol heartbeat, nie nodeguard.

- Uzly sú predkonfigurované a ich konfigurácia sa nedá meniť počas používania.
- Identifikátory položiek v OD môžu byť len 32-bitové čísla.
- Konfigurácia procesných dát nie je povolená.
- Premenné procesných dát nie sú definované v OD, ale len v správach procesných dát.
- Položka objektového slovníka nemôže byť väčšia ako 4 bajty.
- Dynamické zoradovanie viacerých premenných do jednej CAN správy nie je povolené.
- Posielanie CAN správy je založené len na čase a/alebo pri zmene stavu premennej.

V. OPERAČNÝ SYSTÉM COOCOX CoOS

CooCox CoOS [8] je viac úlohový (multitaskingový) operačný systém reálneho času s jadrom, ktoré zodpovedá štandardu CMSIS (Cortex Microcontroller Software Interface Standard) [4]. Je optimalizovaný pre vložené systémy, najmä pre aplikácie založené na mikroprocesoroch ARM Cortex M3.

Jeho zdrojové kódy sú voľne šíriteľné a preto je ľahko prispôsobiteľný podľa potrieb užívateľa. Podporuje preemptívne prideľovanie času aj systém round-robin, doba oneskorenia prerušenia je nulová, dokáže detekovať pretečenie zásobníka a predísť inverzii priorít (stav, keď úloha s vyššou prioritou čaká na dokončenie úlohy s nižšou prioritou). Na komunikáciu a synchronizáciu používa vlajky (flags), semaforey, mutexy, mailboxy, rady (queues) a podporuje prekladače GCC [5], ARMCC a ICCARM.

VI. HARDVÉROVÉ A SOFTVÉROVÉ PROSTRIEDKY

Jedným z cieľov práce bolo vytvoriť fungujúcu sieť niekoľkých uzlov, ktoré budú komunikovať pomocou CAN zbernice využívajúc protokol CANopen. Na to sú nutné zariadenia, ktoré sú schopné zabezpečiť túto komunikáciu. K dispozícii boli tieto vývojové dosky:

- MCBSTM32 od firmy Keil je vývojová doska postavená na báze mikrokontroléra STM32F103RB [9].
- STM3210C-EVAL (Obr. 5 – 2) [11] je vývojová doska od firmy ST Microelectronics. Na rozdiel od MCBSTM32 má táto doska oveľa rozsiahlejšie vybavenie. Pracuje s mikrokontrolérom STM32F107VC [10], ktorý patrí do radu Connectivity Line.
- KEMT STM32 (Obr. 5 – 3) bola vyvinutá na Katedre elektroniky a multimediálnych telekomunikácií (KEMT) Technickej univerzity v Košiciach, pričom bude využitá na cvičeniach špecializovaných predmetov tejto katedry. Jej jadro tvorí mikrokontrolér STM32F103C8 [11].

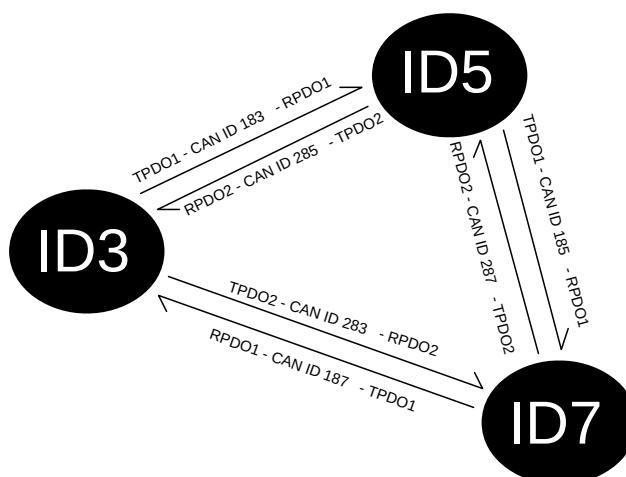
Ďalšie použité hardvérové zariadenia sú programátory/debuggery. Sú to adaptéry, ktoré umožňujú pripojiť počítač, na ktorom je nainštalované vývojové prostredie, s doskou využívajúcou rozhranie napríklad JTAG, SWD alebo iné. Pomocou nich je možné program jednoducho nahráť alebo ladiť priamo v cieľovom zariadení. Použité boli adaptéry Ulink a Ulink2 od firmy ARM Keil a STlink od ST Microelectronics.

Bolo použité integrované vývojové prostredie uVision od firmy Keil (verzia 4.03q) ako súčasť balíka MDK-ARM (verzia 4.11). Okrem vstavaného kompilátora bol využitý aj kompilátor GCC od firmy CodeSourcery. Programové vybavenie bolo vyvíjané tak, aby bol funkčný pre všetky hardvérové prostriedky a rovnako aj pre obidva použité kompilátory.

VII. TVORBA PROGRAMOVÉHO VYBAVENIA

A. Vývoj softvéru pre komunikáciu pomocou protokolu CANopen

Pre tvorbu softvéru, ktorý by bol schopný komunikovať pomocou CAN zbernice je nutné realizovať niekoľko dôležitých krokov. Prvým z nich je výber protokolu a keďže finálna aplikácia je jednoduchá, postačujúce je využitie protokolu MicroCANopen, ktorého kódy sú voľne prístupné a dajú sa získať na internetovej stránke spoločnosti ESA (verzia 3.30). Problémom je však to, že pre túto verziu MCO nie sú vytvorené ovládače pre STM32, čiže ďalším krokom bolo tieto ovládače vytvoriť. Po vytvorení ovládačov nasleduje fáza návrhu funkcionality MCO uzla, ktorá zahŕňa nielen to, čo bude daný uzol vykonávať, ale aj tvorbu tabuliek OD a návrh premenných pre PDO. Posledným krokom je použitie operačného systému reálneho času, ktorý ponúka z hľadiska programovania výhodu v tom, že všetky časti komplexného programu môžu byť samostatne vytvárané a ladené. Všetky zdrojové kódy, podrobný opis vývoja softvéru a kompletne projekty sú dostupné v [14]. Na Obr. 3 je znázornená jednoduchá sieť uzlov využívajúcich protokol MicroCAN open.



Obr. 3 Graf príkladu komunikácie medzi tromi uzlami MCO. Každý uzol posiela správy zvyšným dvom uzlom a zároveň od nich správy prijíma.

B. Tvorba testovacieho softvéru pre vývojovú dosku KEMT STM32

Tento softvér má za úlohu otestovať zariadenia, ktoré sú využité na vývojovej doske KEMT STM32. Program používa pre inicializáciu a periférie knižnice vyvinuté firmou Keil, ktoré sú veľmi jednoduché a boli vytvorené pre demonštračné účely. Štandardné knižnice STM32, ktoré boli využité pri tvorbe softvéru opísanej v predchádzajúcej časti sú oproti knižniciam Keil podstatne komplexnejšie, pričom ponúkajú plný prístup k zariadeniam mikrokontroléra a k ich nastaveniam.

Telo funkcie *main()* je veľmi jednoduché. Najskôr sa inicializuje hardvér, základná CAN správa a na sériovú linku sa vypíše úvodná hlavička, ktorá obsahuje informácie o ovládaní programu. Podľa prijatého znaku zo sériovej linky program vykonáva tieto činnosti:

- Znak „h“ - Vypíše úvodnú hlavičku.
- Znak „l“ - Zmení stav LED1.
- Znak „s“ - Skontroluje tlačidlá S1, S2 a S3 a ich stav vypíše na sériovú linku.
- Znak „c“ - Vyšle na CAN zbernicu správu s identifikátorom 33, ktorá obsahuje v prvom slotе pre dáta číslo 15 a zároveň vypíše obsah poslednej prijatej správy s tým istým identifikátorom.

Pomocou tohto softvéru a projektu Martin-a Thomas-a [13] bolo uskutočnené oživovanie a testovanie hardvéru. Programy odhalili technické problémy s konektorom pre SD kartu, CAN budičom a kryštálom pre RTC. Tieto problémy boli následne odstránené a po ďalšom testovaní bola overená plná funkčnosť všetkých komponentov a ich zapojenie.

VIII. ZÁVER

V príspevku boli úspešne predstavené softvérové prostriedky pre vývojové dosky na báze mikrokontrolérov STM32. Ukázalo sa, že tieto mikrokontroléry sú vhodné aj pre zložitejšie vložené aplikácie. Tiež bol overený nový a perspektívny operačný systém reálneho času CoCoX CoOS. Veľkou výhodou CoOS je jeho kompatibilita s jadrom Cortex-M3, ktorá umožňuje jeho využitie v mnohých zariadeniach a aplikáciách. Výsledok vývoja softvéru pre komunikáciu na MCO je šesť projektov, ktoré sú pripravené pre tri rôzne vývojové dosky a pre obidva prekladače. Navyše bol program vytvorený tak, aby bol uľahčený prechod na iné vývojové dosky alebo zariadenia, ktoré môžu byť priamo nasadené do praxe. Tieto projekty predstavujú len ukážku možností štandardu MicroCANopen a majú slúžiť ako základ pre tvorbu zložitých aplikácií komunikujúcich v komplexných sieťach na báze CANopen. Sieťe, ktoré využívajú len MCO sa dajú využiť na tvorbu zložitejšej senzorovej siete.

Jedným z najväčších prínosov pri tvorbe softvéru pre uzly MCO bolo vytvorenie ovládačov pre perifériu CAN mikrokontrolérov STM32. Tieto ovládače umožňujú vytvoriť zo zariadenia, ktoré je postavené na báze STM32 plnohodnotný MCO uzol. Pri vývoji softvéru bol kladený dôraz na to, aby jednotlivé programy boli funkčné na všetkých dostupných vývojových doskách a pre obidva použité prekladače. Pri tejto snahe boli odhalené niektoré problémy, ktoré súvisia najmä s nekompatibilitou použitých prostriedkov a s chybami v knižniciach výrobcov. V štandardnej knižnici STM32 pre perifériu CAN bola nájdená chyba, ktorá zatiaľ nebola výrobcom opravená. Chyba tiež vzniká pri preklade súborov CoOS prekladačom GCC s optimalizáciou pre veľkosť kódu. Posielanie na sériovú linku nie je korektné pri použití programátora ST-Link a prekladača GCC. Simulácia v prostredí uVision pre dosku STM3210C-

EVAL nie je zatiaľ podporovaná. Tiež sú to menšie problémy, ktoré sú spomenuté v readme súboroch a dočasne vyriešené v zdrojových kódach projektov dostupných v prílohe A a CD médiu. Niektoré problémy boli odstránené použitím iných hardvérových alebo softvérových prostriedkov a bude im venovaná pozornosť pri ďalšej práci. Ďalší vývoj sa môže uberať niekoľkými smermi. Jedným z nich je rozšírenie funkcionality MCO uzla, pričom samotný uzol by vykonával množstvo úloh a navyše by sa dal implementovať a odskúšať protokol MicroLSS, ktorý by rozšíril funkcie v rámci komunikácie medzi jednotlivými uzlami. Štandard MCO je však obmedzený a tak ďalším krokom by mohla byť implementácia kompletného štandardu CANopen. Keďže kompletný štandard CANopen aj protokol MicroLSS vyžadujú hlavný uzol, zámerom KEMT je tento uzol získať alebo vytvoriť. Sieť by potom mohli zabezpečovať mnoho zložitejšie procesy napríklad v automatizácii alebo v inteligentných domoch.

V zložitejších systémoch pri použití hlavného uzla by mohli byť ostatné uzly, ktoré vykonávajú jednoduché činnosti postavené na MicroCANopen. Nevýhodou však zostáva fakt, že komunikácia na CAN zbernici vyžaduje fyzické vodiče, aj keď existujú už aj systémy, ktoré fungujú na princípe CAN zbernice, ale využívajú bezdrôtové technológie.

IX. POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) (50%) a projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) (50%) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] MARTIN Trevor: The Insider's guide to the STM32 ARM based microcontroller [online]. Veľká Británia: Hitex, Február, 2008. [aktualizované September 2008]. [citované 15.5.2010]. Dostupné na internete: <<http://www.st.com/mcu/files/mcu/1221142709.pdf>>. ISBN 0-9549988 8
- [2] ST Microelectronics: STM32 Reference Manual [online]. April, 2010, Version 11. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.st.com/stonline/products/literature/rm/13902.pdf>>.
- [3] ARM: Cortex-M3 [online]. Máj, 2010. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.arm.com/products/processors/cortex-m/cortex-m3.php>>.
- [4] ARM: CMSIS [online]. Máj, 2010. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.arm.com/products/processors/cortex-m/cortex-microcontrollersoftware-interface-standard.php>>.
- [5] GNU: GCC [online]. Máj, 2010. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://gcc.gnu.org/>>.
- [6] PFEIFFER, Olaf - AYRE, Andrew - KEYDEL, Christian: Embedded Networking with CAN and CANopen. Spojené Štáty Americké: Copperhill Technologies Corporation, 2003. ISBN 978-0-9765116-2-5
- [7] ARM: SWD [online]. Máj, 2010. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.arm.com/products/system-ip/debug-trace/coresight-soccomponents/serial-wire-debug.php>>.
- [8] Coocox: Coocox CoOS User's Guide [online]. August, 2009. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <http://www.coocox.org/downloadfile/CoOS/PDF/Coocox_CoOS_User_Guide.pdf>.
- [9] ST Microelectronics: Datasheet for STM32F103x8/B [online]. September, 2009, Version 11. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.st.com/stonline/products/literature/ds/13587.pdf>>.
- [10] ST Microelectronics: Datasheet for STM32F105/107xx [online]. September, 2009, Version 4. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.st.com/stonline/products/literature/ds/15274.pdf>>.
- [11] ST Microelectronics: User Manual for STM3210C-EVAL [online]. Jún, 2009, Revision 3.0. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <<http://www.st.com/stonline/products/literature/um/15082.pdf>>.
- [12] VARGA Tomáš: CAN senzorová sieť pre monitorovanie vybraných parametrov v automobiloch: Diplomová práca. Košice: TUKE FEI, Máj, 2010. 60 s.
- [13] THOMAS Martin: Chan's FAT-Module with STM32 SPI [online]. Január, 2010. [s.a.]. [citované 15.5.2010]. Dostupné na internete: <http://www.siwawi.arubi.unikl.de/avr_projects/arm_projects/arm_memcards/index.html#chanfat_stm32>.
- [14] Martin Petrvalský: CAN zbernica vo vložených aplikáciách na báze STM32 Cortex-M3. Bakalárska práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky. 2010. 45 s.

Syntetický procesor NIOS s galvanicky oddeleným sériovým rozhraním

Peter HUSNAJ, Miloš DRUTAROVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

peter.husnaj@student.tuke.sk, milos.drutarovsky@tuke.sk

Abstrakt — Práca opisuje návrh a technickú realizáciu galvanicky oddeleného prevodníka USB (Universal Serial Buss) na UART (Universal Asynchronous Receiver/Transmitter) a overenie jeho funkčnosti pomocou vlozenej aplikácie implementovanej v syntetickom procesore Altera NIOS II. Na realizáciu prevodníka je využitý čip FT232RL firmy FTDI a ADuM1201 firmy Analog Devices. Aplikácia na overenie funkčnosti je realizovaná na vývojovom kíte UP3-1C6 osadeným obvodom FPGA (Field Programmable Gate Array) Cyclone od firmy Altera, pomocou programov Quartus II a NIOS II Embedded Design Suite.

Kľúčové slová — UART, USB, NIOS, Altera

I. ÚVOD

Súčasný moderný počítač podlieha trendu neimplementovať v dnešnej dobe už málo používané štandardné rozhrania ako je sériový port [1] alebo paralelný port [2], ktoré predstavujú dedičstvo z počítačovej minulosti. Napriek tomu pri vývoji vložených systémov aj v súčasnosti existuje reálna potreba prepojenia cieľovej aplikácie s počítačom za účelom napr. ladenia prostredníctvom sériového portu ktorý je stále obľúbený vďaka jeho jednoduchšej implementácii a nedávnej značnej rozšírenosti. Ak počítač neobsahuje sériový port, je možné vybrať jedno z riešení:

- zaobstarať si starší počítač, v ktorom sú tieto porty implementované,
- rozšíriť počítač o tieto porty rozširujúcimi kartami,
- realizovať prevodník pomocou špecializovaných obvodov, transformujúcich signály z USB (Universal Serial Buss) portu [3], ktorým je v dnešnej dobe vybavený každý počítač.

Táto práca je zameraná na posledné riešenie, a to realizáciu prevodníka medzi USB rozhraním a sériovým rozhraním s galvanickým oddelením, ktorým bude možné rozšíriť počítač bez sériového rozhrania.

Zrealizovaný prevodník by mal disponovať vývodmi pre vysielanie a prijímanie sériových dát a vývodmi pre hardvérové riadenie prenosu dát, teda nie je plnohodnotnou náhradou rozhrania RS232. Prevodník bude určený predovšetkým pre pripojenie mikroprocesorových systémov disponujúcich UART (Universal Asynchronous Receiver/Transmitter) rozhraním s prípadnou možnosťou hardvérového riadenia prenosu dát.

Aj keď je možné kompletne prevodníky USB-RS232 v dnešnej dobe zakúpiť v obchodoch, výhodou nami navrhnutého prevodníka je galvanické oddelenie prenášaných signálov sériového rozhrania a jeho realizačná cena.

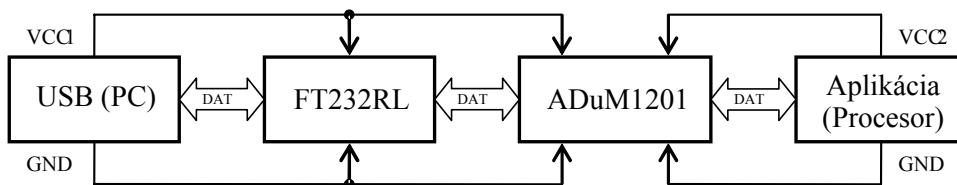
V práci bude navrhnutý kompletný vložený systém s využitím syntetického procesora Altera NIOS II [4], pomocou ktorého bude demonštrovaná činnosť a funkčnosť zhotovených modulov prevodníka.

II. NÁVRH MODULU USB/UART PREVODNÍKA

Základom modulu prevodníka sú dva integrované obvody. Obvod slúžiaci na prevod signálov z USB rozhrania na signály sériového rozhrania a obvod pre galvanické oddelenie signálov sériového rozhrania.

V súčasnosti existuje viacero integrovaných obvodov slúžiacich ako prevodníky medzi USB a sériovým rozhraním. Základom zapojenia modulu prevodníka je takýto integrovaný obvod

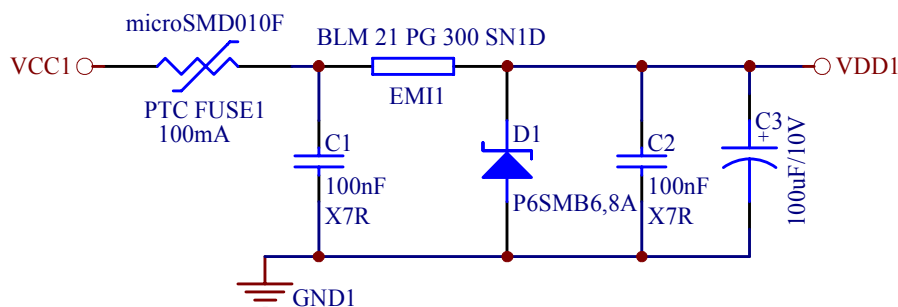
FT232RL [5] od firmy FTDI [6]. Obvod FT232RL pre svoju funkčnosť vyžaduje minimum externých súčiastok. Obvod bol zapojený podľa odporúčaného zapojenia z katalógového listu. Na galvanické oddelenie bol použitý obvod ADuM1201 [7] od firmy Analog Devices [8], ktorý používa na oddelenie primárnej a sekundárnej strany miniatúrny transformátor, vyleptaný priamo do štruktúry kremíka. Obvod vyžaduje dva napájacie zdroje, jeden zo strany USB a druhý zo strany pripojenej aplikácie. Prvá etapa pri návrhu schémy zapojenia bolo nakreslenie blokovej schémy (Obr. 1), ktorej základnou časťou sa stali vyššie popísané integrované obvody.



Obr. 1 Bloková schéma modulu prevodníka

Pre galvanické oddelenie sériového rozhrania sú potrebné dve napájacie napätia. Pre korektnú funkciu obvodu FT232RL bez externého kryštálového oscilátora je potrebné napájacie napätie minimálne 4 V a pre prácu s TTL (Transistor-Transistor Logic) signálmi je potrebné na jednom vstupe obvodu napájanie 5 V. Preto bolo zvolené napájacie napätie 5V, ktorým môže byť napájaný aj obvod ADuM1201. Keďže maximálny odoberaný prúd na strane USB rozhrania podľa výpočtov nepresiahol 100 mA je výhodné použiť napájanie priamo z USB portu počítača, ktorý disponuje napájacím napätím 5 V a je schopný tento prúd dodať.

Ďalšími súčiastkami, z ktorých pozostáva modul prevodníka sú súčiastky filtrujúce napájanie a súčiastky chrániace obvody pred ich zničením. Totožné zapojenie je použité na strane USB a aj na strane pripojenej aplikácie (Obr. 2).

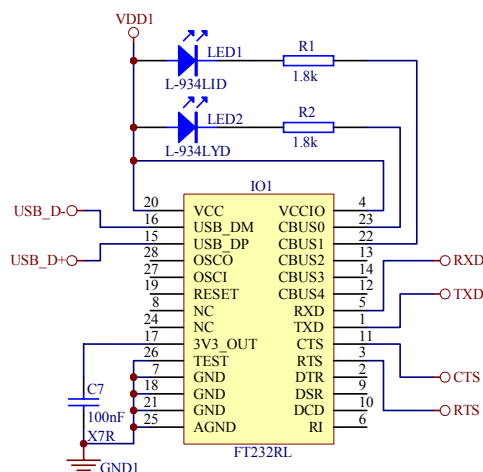


Obr. 2 Filtrácia napájania

Ako filtračné členy slúžia keramické kondenzátory C1, C2 s kapacitou 100 nF a dielektrikom X7R, EMI (ElectroMagnetic Interference) filter EMI1 typu BLM21PG300SN1D [9] a tantalový kondenzátor C3 s kapacitou 100 μ F. Na ochranu obvodu pred napäťovými špičkami slúži transil D1 typu P6SMB6,8A [10], filtrujúci napäťové špičky s amplitúdou väčšou ako 6,8 V a na ochranu pred skratovým prúdom slúži vratná poistka polyswitch PTC FUSE1 typu microSMD010F [11] obmedzujúca prúd väčší ako 100 mA. Vstupom pre nefiltrované napätie je VCC1 pre kladný pól a GND1 pre záporný pól. Výstupom filtrovaného napätia je VDD1. Záporný pól GND1 je spoločný pre vstup aj výstup napájania.

Na indikáciu pripojeného napätia sú použité dve LED (Light Emitting Diode) diódy indikujúce pripojené napätie z USB a pripojené napätie z aplikácie. Aby bola spotreba celého zapojenia čo najmenšia na indikáciu sú použité nízkopríkonové zelené LED typu L-934LGD [12], ktorých spotreba prúdu je 2 mA. Pre porovnanie spotreba bežne používaných LED diód je 20 mA.

Obvod FT232RL zabezpečuje prevod medzi USB a sériovým rozhraním. Obvod transformuje signály D+ a D- z USB na signály RXD, TXD, RTS, CTS, DTR, DSR, DCD a RI rozhrania UART používané pre sériový prenos dát. Zapojenie modulu prevodníka využíva iba prvé štyri signály RXD, TXD pre vysielanie a príjem sériových dát a RTS, CTS pre hardvérové riadenie prenosu sériových dát. Z celkového počtu 28 vývodov obvodu FT232RL je zapojených iba 16 vývodov. (Obr. 3)

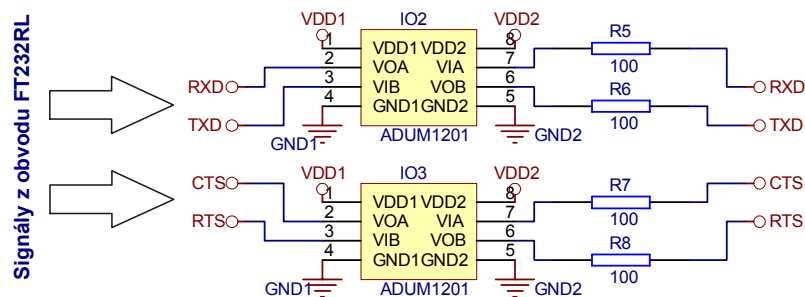


Obr. 3 Zapojenie obvodu FT232RL

Obvod pre správnu činnosť nepotrebuje žiadne externé súčiastky. Keramický kondenzátor s kapacitou 100 nF je potrebné použiť, ak sa nevyužíva výstup z integrovaného regulátora napätia (vývod 3V3_OUT). LED diódy LED1 a LED2 sú použité na indikáciu prijímaných a vysielaných dát. LED1 je nízkoprikonová červená LED typu L-934LID [13] indikujúca prijímané dáta a LED2 je nízkoprikonová žltá LED typu L-934LYD [14] indikujúca vysielané dáta. Vývod VCC slúži na pripojenie napájacieho napätia, ktoré je v našom prípade 5 V. Veľkosťou pripojeného napätia na vývod VCCIO sa volí logika výstupných signálov. Pri používaní TTL logiky je potrebné tento vývod pripojiť na napájacie napätie 5 V. Na komunikáciu s USB rozhraním slúžia vývody USB_DP a USB_DM. Na komunikáciu so sériovým rozhraním slúžia vývody RXD, TXD pre prijímanie a vysielanie dát a vývody RTS a CTS pre hardvérové riadenie komunikácie.

ADuM1201 je dvojkanálový digitálny oddeľovač založený na Analog Devices, Inc., iCoupler technológii [15]. Kombináciou vysokorychlostnej CMOS (Complementary Metal Oxide Semiconductor) a monolitické transformátorovej technológie tieto oddeľovacie komponenty poskytujú lepšie prevádzkové vlastnosti v porovnaní s inými alternatívami ako sú napríklad optické väzobné členy. Medzi tieto vlastnosti patrí hlavne prúdová spotreba a rýchlosť prenášaných signálov. Jedným obvodom je možné galvanicky oddeliť dva prenášané signály.

Zapojenie obvodu je veľmi jednoduché. Obvod pre svoju korektnú funkčnosť používa dve napájacie napätia. Keďže jeden obvod je schopný galvanicky oddeliť dva prenášané signály je potrebné v zapojení modulu prevodníka použiť dva takéto obvody, pretože je potrebné spracovať 4 signály sériového rozhrania (Obr. 4).



Obr. 4 Zapojenie obvodu ADuM1201

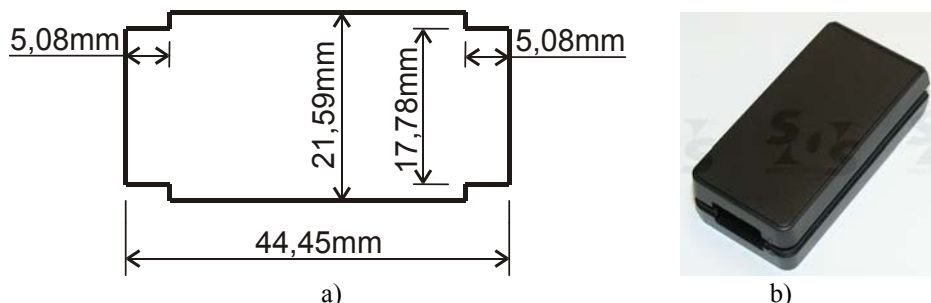
Rezistory s odporom 100 Ω použité v zapojení obmedzujú prúd pri náhodnom skrate a slúžia na ochranu obvodov pred ich prípadným zničením.

Pri výbere súčiastok rozhodovala predovšetkým ich veľkosť, aby bola navrhnutá doska plošných spojov čo najmenšia a vošla sa do vopred vybranej škatuľky P3A-201005U [16]. Keďže škatuľka má vopred pripravený otvor pre konektor typu USB-A [17], na prepojenie PC s modulom prevodníka bol zvolený práve tento konektor. Na prepojenie modulu prevodníka s pripojenou aplikáciou bola zvolená dvojradová rovná kolíková lišta WWD6G [18] so šiestimi

vývodmi potrebnými pre prepojenie štyroch signálov sériového rozhrania a napájania z aplikácie. Všetky použité súčiastky okrem indikačných LED diód sú typu SMD (Surface-Mount Devices).

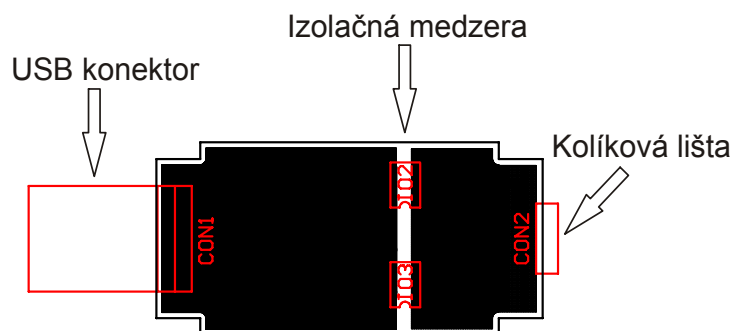
III. NÁVRH DPS A TECHNICKÁ REALIZÁCIA MODULU PREVODNÍKA

Na návrh dosky plošných spojov (ďalej len DPS) bol použitý program PROTEL 99 SE [19]. Návrh bol limitovaný veľkosťou DPS (Obr. 5a), ktorej rozmery boli zvolené tak, aby mohla byť vložená do škatuľky P3A-201005U (Obr. 5b).

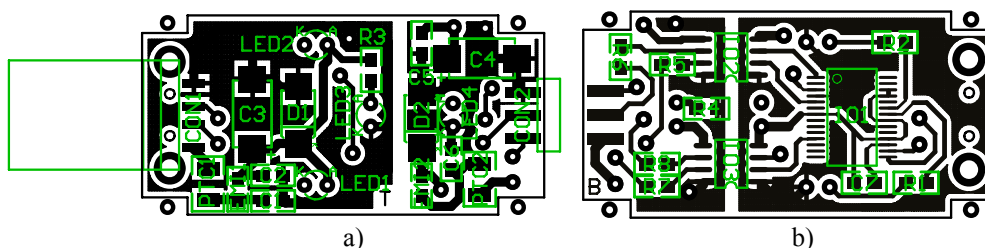


Obr. 5 a) Rozmery dosky plošných spojov, b) Škatuľka P3A-201005U

Ďalším obmedzením pri návrhu DPS bolo pevné umiestnenie oboch konektorov na koncoch DPS a galvanické oddelenie sériového rozhrania tak, aby mala izolačná medzera dostatočnú šírku (Obr. 6).



Obr. 6 Umiestnenie konektorov na DPS

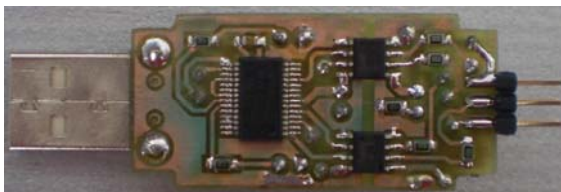


Obr. 7 Navrhnutá DPS s rozmiestnením súčiastok: a) TOP strana, b) BOTTOM strana

Pri návrhu modulu prevodníka boli taktiež vygenerované všetky potrebné výstupy pre profesionálnu výrobu. Kvôli realizačnej cene bola však použitá metóda výroby DPS fotocestou, ktorá je výhodná v amatérskych podmienkach. Na výrobu bola použitá obojstranná doska plošných spojov s profesionálne nanosenou fotocitlivou vrstvou.



Obr. 8 Osadená DPS – TOP strana



Obr. 9 Osadená DPS – BOTTOM strana



Obr. 10 Hotový modul prevodníka

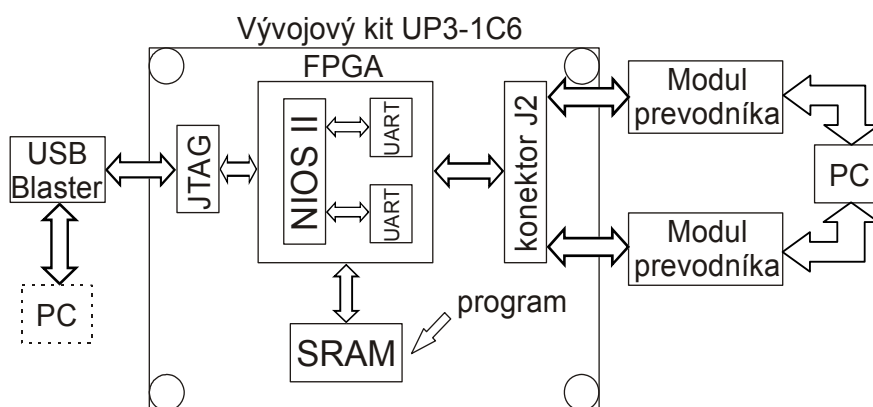
IV. PREPOJENIE MODULU PREVODNÍKA SO SYNTETICKÝM PROCESOROM ALTERA NIOS II

Ako reálna aplikácia pre overenie funkčnosti modulu prevodníka bol zvolený FPGA (Field Programmable Gate Array) obvod EP1C6Q240 [20] od firmy Altera a ich syntetický procesor NIOS II, ktorý je možné nakonfigurovať do FPGA obvodu. Výhodou pri použití syntetického procesora NIOS II je konfigurácia jeho rozhraní pre pripojenie periférií, čo umožňuje napr. pridanie dvoch UART rozhraní a tým je možné overiť funkčnosť dvoch modulov prevodníka súčasne.

Syntetický procesor Altera NIOS II je RISC (Reduced Instruction Set Computer) procesor vytvorený tak, aby bol v maximálnej možnej miere konfigurovateľný podľa požiadaviek používateľa. NIOS II predstavuje takzvaný IP (Intellectual Property) blok, čo je vlastne konfiguračný kód v jazyku HDL (Hardware Description Language), určený pre konfiguráciu FPGA obvodu do ktorého je procesor „vkladaný“.

Konfigurovateľnosť NIOS II procesora spočíva v tom, že je možné ho použiť podľa požiadaviek pre široké spektrum aplikácií od jednoduchého 16-bitového procesora pre riadenie nenáročných aplikácií až po výkonný 32-bitový systém s možnosťou definovania vlastných inštrukcií [21]. Práve možnosť definovať si vlastné inštrukcie a možnosť pripojenia vlastných periférií priamo na zbernicu Avalon [22] ho robia výrazne odlišujú od klasických mikrokontrolérov. [23]

Na realizáciu syntetického procesora NIOS II a na jeho prepojenie s navrhnutými modulmi prevodníkov bol použitý vývojový kit UP3-1C6 [24] a vývojové prostredie Quartus II [25] od firmy Altera [26], ktoré je vytvorené pre prácu s obvodmi CPLD (Complex Programmable Logic Device) a FPGA. Pomocou nich bol vytvorený syntetický procesor NIOS II s týmito hardvérovými vlastnosťami: NIOS II/s 32-bit, SRAM 128 kbit, JTAG UART, Avalon Tri State Bridge, časovač, výstupný port pre 4 LED a dve rozhrania UART. Taktovacia frekvencia procesora bola zvolená 48 MHz. Na konfigurovanie FPGA obvodu bol použitý adaptér USB Blaster, ktorý je jednou z možností ako konfigurovať FPGA obvody.



Obr. 11 Bloková schéma prepojenia modulov prevodníka so syntetickým procesorom NIOS II

Pre overenie funkčnosti modulov prevodníkov pripojených k procesoru NIOS II bol vytvorený jednoduchý program v jazyku C, ktorého úlohou je príjem sériových dát vysielaných z PC prostredníctvom jedného modulu prevodníka a vysielanie prijatých dát do PC prostredníctvom druhého modulu prevodníka. Pre programovanie syntetického procesora NIOS II bolo použité vývojové prostredie NIOS II EDS [27] od firmy Altera.

V. ZÁVER

Moduly sú výstupom bakalárskej práce [28]. Sú skonštruované, otestované a pripravené na použitie. Navrhnutý modul prevodníka je jednoduchý a obsahuje minimum súčiastok vďaka

obvodu FT232RL. Použitie SMD súčiastok prispelo k minimalizácii rozmerov hotového modulu. Taktiež bol úspešne navrhnutý vložený systém s využitím syntetického procesora Altera NIOS II, pomocou ktorého boli otestované moduly prevodníkov.

VI. POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) (50%) a projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) (50%) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Vít Olmr: Sériová linka RS-232 [online]. HW.cz, 2005. Aktualizované 12.12.2005. Dostupné na internete: <<http://hw.cz/rs-232>>.
- [2] Vít Olmr: Paralelný port - LPT (IEEE 1284) [online]. HW.cz, 2005. Aktualizované 12.12.2005. Dostupné na internete: <<http://hw.cz/lpt>>.
- [3] Redakcia HW serveru: USB–Universal Serial Bus [online]. HW.cz, 2002. Aktualizované 07.05.2002. Dostupné na internete: <<http://hw.cz/Teorie-a-praxe/Dokumentace/ART327-USB---Universal-Serial-Bus---Popis-rozhrani.html>>.
- [4] Altera: NIOS II Processor Reference Handbook [online]. Altera, 2009. Aktualizované November 2009. Dostupné na internete: <http://www.altera.com/literature/hb/nios2/n2cpu_nii5v1.pdf>.
- [5] Future Technology Devices International Ltd.: FT232R USB UART IC Data Sheet [online]. Future Technology Devices International Ltd., 2009. Aktualizované December 2009. Dostupné na internete: <http://www.ftdichip.com/Documents/DataSheets/DS_FT232R_V205.pdf>.
- [6] Internetová stránka spoločnosti Future Technology Devices International Ltd. <<http://www.ftdichip.com>>.
- [7] Analog Devices: Dual-Channel Digital Isolators [online]. Analog Devices, 2009. Aktualizované Január 2009. Dostupné na internete: <http://www.analog.com/static/imported-files/data_sheets/ADuM1200_1201.pdf>.
- [8] Internetová stránka spoločnosti Analog Devices <<http://www.analog.com>>.
- [9] BLM21 Series [online]. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/a/pdf/fil_BLM21.pdf>.
- [10] Semikron: P6SMB6,8A [online]. Semikron, 2004. Aktualizované 02.03.2004. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/b/pdf/p6smb.pdf>.
- [11] Farnell: Polyswitch Resettable Devices [online]. Farnell, Dostupné na internete: <<http://www.farnell.com/datasheets/68397.pdf>>.
- [12] Kingbright: L-934LGD [online]. Kingbright, 2002. Aktualizované 30.12.2002. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/pdf/king/L-934LGD.pdf>.
- [13] Kingbright: L-934LID [online]. Kingbright, 2002. Aktualizované 30.12.2002. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/pdf/king/L-934LID.pdf>.
- [14] Kingbright: L-934LYD [online]. Kingbright, 2002. Aktualizované 30.12.2002. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/pdf/king/L-934LYD.pdf>.
- [15] Pandatron: Co je a co nabízí iCoupler [online]. Pandatron.cz, 2009. Aktualizované 03.05.2009. Dostupné na internete: <http://pandatron.cz/?807&co_ja_co_nabizi_icoupler>.
- [16] New Age Enclosures: P3A-201005U [online]. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/i/USB31.pdf>.
- [17] Assmann: USB [online]. Dostupné na internete: <http://www.soselectronic.com/a_info/resource/f/assmann/a-usb-a.pdf>.
- [18] SOS electronic: Kolíkové lišty [online]. Dostupné na internete: <http://www.soselectronic.com/a_info/kat/f41.pdf>.
- [19] Altium: Protel99SE [online]. Altium, 2002. Dostupné na internete: <<http://www.altium.com/files/protel/legacy/handbookp99se.pdf>>.
- [20] Altera: Cyclone FPGA Family Data Sheet [online]. Altera, 2008. Aktualizované Máj 2008. Dostupné na internete: <http://www.altera.com/literature/hb/cyc/cyc_c5v1_01.pdf>.
- [21] Altera: NIOS II Custom Instruction [online]. Altera, 2008. Aktualizované Máj 2008. Dostupné na internete: <http://www.altera.com/literature/ug/ug_nios2_custom_instruction.pdf?GSA_pos=1&WT.oss_r=1&WT.oss=cust om%20logic%20nios>.
- [22] Altera: Avalon Interface Specifications [online]. Altera, 2009. Aktualizované Apríl 2009. Dostupné na internete: <http://www.altera.com/literature/manual/mnl_avalon_spec.pdf?GSA_pos=3&WT.oss_r=1&WT.oss=avalon%20 bus>.
- [23] Ing. Igor Hrušovský: Mikroprocesory, mikropočítače a mikrokontroléry/Softvérové procesory [online]. 2004. Aktualizované September 2004. Dostupné na internete: <http://www.kemt.fei.tuke.sk/predmety/KEMT411_ESM/_web/wwwfiles/str%2007.htm>.
- [24] System Level Solutions: UP3 Education Kit Reference Manual [online]. System Level Solutions, 2007. Aktualizované 07.09.2007. Dostupné na internete: <<http://www.slscorp.com/UP3Support/pages/documents.php>>.
- [25] Altera: Quartus II Handbook Version 9.1 [online]. Altera, 2009. Aktualizované November 2009. Dostupné na internete: <http://www.altera.com/literature/hb/qts/quartusii_handbook.pdf>.
- [26] Internetová stránka spoločnosti Altera <<http://www.altera.com>>.
- [27] Altera: Embedded Design Handbook [online]. Altera, 2010. Aktualizované Marec 2010. Dostupné na internete: <http://www.altera.com/literature/hb/nios2/edh_ed_handbook.pdf>.
- [28] P. HUSNAJ, “ Syntetický procesor NIOS s galvanicky oddeleným sériovým rozhraním“ Bakalárska práca, TUKE FEI KEMT, 2010.

Vývojová doska na báze Altera FPGA Cyclone III

Jozef MACEJ, Miloš DRUTAROVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

jozef.macej@student.tuke.sk, milos.drutarovsky@tuke.sk

Abstrakt — Príspevok opisuje návrh vývojovej dosky s FPGA obvodom z rodiny Altera Cyclone III s využitím moderných návrhových a simulačných metód, ako aj dostupných vývojových CAD nástrojov. Navrhnutá vývojová doska je koncipovaná tak, aby poskytla užívateľovi široké možnosti pre overenie komunikácie medzi FPGA obvodom a externými periférnymi obvody ako napríklad LED diódy, displej, FLASH pamäť, SDRAM a USB rozhranie. Zvláštna pozornosť je venovaná dimenzovaniu zdroja napájacieho napätia vývojovej dosky s využitím integrovaného impulzného a lineárneho stabilizátora spolu s jeho návrhom a simuláciou. Prezentovaný návrh tiež demonštruje problematiku návrhu dosky plošného spoja, v ktorom sú využité možnosti simulácie dosky plošného spoja na analýzu elektrických vlastností kritických vodivých ciest plošného spoja.

Kľúčové slová — Altera, FPGA, CAD, simulácia

I. ÚVOD

Na trhu existuje nespočetné množstvo digitálnych obvodov a každý obvod nachádza svoje uplatnenie v inej oblasti elektroniky. Špecifickou skupinou digitálnych obvodov sú FPGA (Field Programmable Gate Array) obvody ktoré sa vyznačujú flexibilitou a rýchlosťou, a práve to je dôvod ich využitia v aplikáciách, kde je užívateľom navrhnutý špecializovaný hardvér schopný v krátkom čase spracovať väčší objem dát. V neposlednom rade sa s využitím FPGA obvodov v elektronických zariadeniach znižuje množstvo použitých integrovaných obvodov, pretože FPGA obvody sa vyznačujú vysokou funkcionalitou.

Navrhnutá vývojová doska umožní návrhárovi prakticky overiť konfiguráciu FPGA obvodu a komunikáciu medzi FPGA obvodom a perifériami ako napríklad rýchle USB rozhranie, FLASH pamäť, dynamická RAM a podobne.

Aby bolo možné vykonať spoľahlivý návrh zariadenia, je potrebné poznať vlastnosti jednotlivých komponentov. Návrhár musí poznať najmä napájacie napätia použitých integrovaných obvodov, ich spôsoby konfigurácie a spôsoby komunikácie s ostatnými integrovanými obvody.

Na základe požiadaviek na napájacie napätia a odoberané prúdy je navrhnutý zdroj napájacieho napätia. Zdroj napájacieho napätia obsahuje niekoľko výstupných vetiev pre analógové alebo digitálne časti vývojovej dosky.

Veľký dôraz je kladený na správnosť návrhu schémy zapojenia, pretože od nej sa odvíja návrh DPS (dosky plošného spoja). Ak by bol návrh nesprávny a tento fakt by sa zistil až po fyzickej realizácii vývojovej dosky, v praxi by to viedlo k časovým stratám spojeným s opätovným návrhom a výrobou DPS, čo sa odzrkadlí na zvýšených výrobných nákladoch.

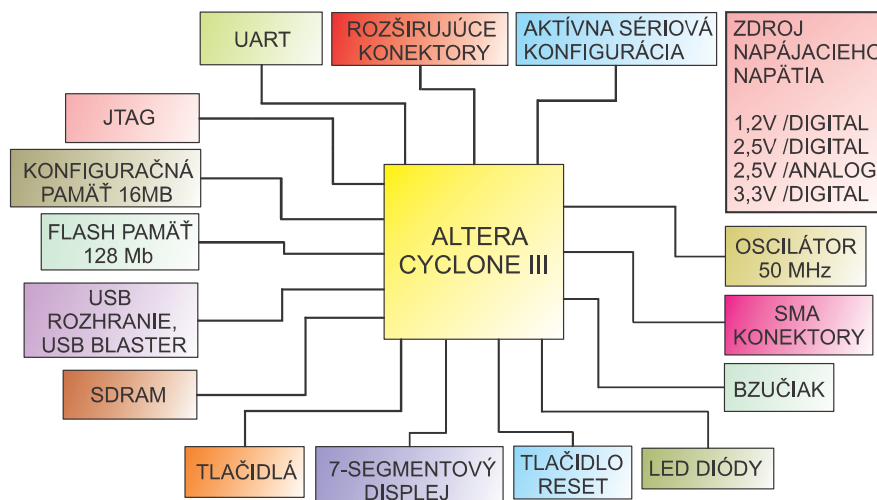
K návrhu dosky plošného spoja sa pristupuje ak je návrh schémy zapojenia kompletný a prekontrolovaný. V návrhovom programe je vhodné vykonať simulácie DPS, aby sa odstránili nedostatky z dôvodu rušení a odrazov v spojoch. Pri návrhu DPS je potrebné prihliadať na možnosti cieľovej výroby. Ak je plánovaná strojová výroba, DPS musí byť prispôbená výrobným obmedzeniam. Finálnym krokom po návrhu dosky plošného spoja v návrhovom programe je generovanie výstupných podkladov pre výrobu.

II. NÁVRH SCHÉMY ZAPOJENIA

Prvý krok pri návrhu schémy zapojenia elektronického zariadenia predstavuje návrh blokovej schémy zariadenia, ktorá umožňuje zjednodušený pohľad na celé zariadenie. V

d'alších krokoch návrhu schémy zapojenia sú rozpracované jednotlivé časti blokovej schémy, čo vedie k finálnej schéme zapojenia.

Pri návrhu opisovanej vývojovej dosky vznikli dve blokové schémy, a to bloková schéma celkového zapojenia (obr. 1) a bloková schéma napájacieho zdroja (obr. 2).



Obr. 1 Bloková schéma vývojovej dosky

Jadro vývojovej dosky tvorí FPGA obvod EP3C10E144C8 z rodiny obvodov Cyclone III [1] od spoločnosti Altera [2], ktorý sa vyznačuje vysokou funkcionalitou, nízkou cenou a okrem iného obsahuje aj 10320 logických elementov, dva obvody PLL, 94 užívateľských vývodov a 414 kBitovú integrovanú pamäť SRAM. FPGA obvod na vývojovej doske je možné konfigurovať prostredníctvom JTAG rozhrania, aktívnej sériovej komunikácie a zaujímavosťou je možnosť konfigurácie prostredníctvom USB rozhrania a integrovaného USB Blastra vyvinutého na Katedre elektroniky a multimediálnych telekomunikácií TU v Košiciach. Konfiguračné dáta sú uložené v sériovej konfiguračnej pamäti typu SRAM o veľkosti 16 MB. Vývojová doska ponúka 128 Mbitovú sériovú pamäť typu FLASH a 256 Mbitovú pamäť typu SDRAM. Na generovanie hodinových signálov pre FPGA obvod slúži kryštálový oscilátor s frekvenciou 50MHz a doska tiež ponúka možnosť pripojenia externého zdroja hodinových signálov cez SMA konektory. Vývojová doska disponuje klasickými perifériami ako LED diódy, UART, dvojité sedem-segmentový displej, akustický výstup vo forme bzučiaka a tlačidlá. Výrazným zvýšením flexibility je použitie rozširovacích konektorov, ktoré umožňujú rozšíriť vývojovú dosku o ďalšie periférie, prípadne umožnia prepojiť dve vývojové dosky navzájom.

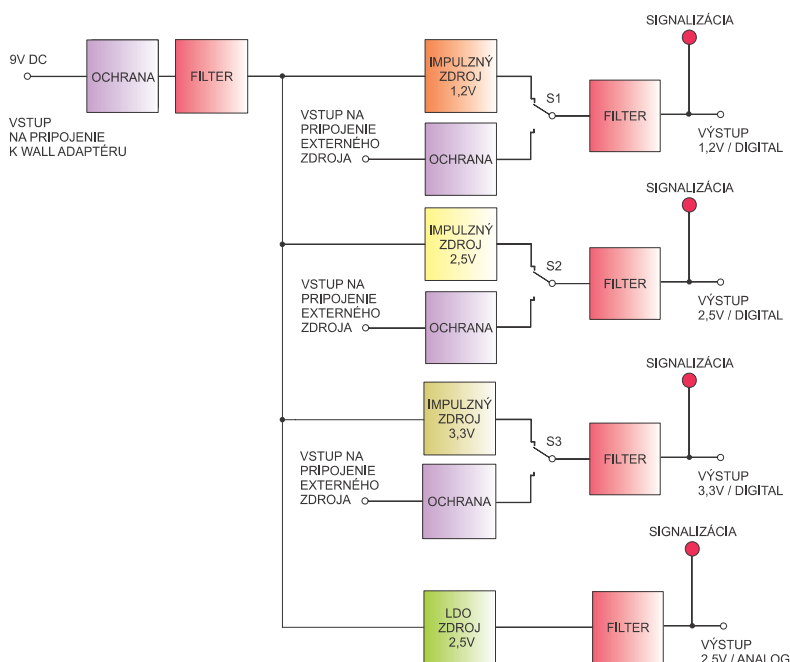
III. NÁVRH NAPÁJACIEHO ZDROJA

Použitý FPGA obvod EP3C10E144C8 z rodiny obvodov Cyclone III [1] vyžaduje pre svoju činnosť niekoľko napájacích napätí. Jadro FPGA obvodu je napájané napätím 1,2 V, obvod fázového závesu vyžaduje 1,2 V zdroj pre digitálnu časť obvodu a 2,5 V zdroj pre lineárnu časť obvodu. Pre ďalšie komponenty vývojovej dosky je potrebný 3,3 V zdroj napájacieho napätia. Pre prípadnú realizáciu komunikácie s 2,5 V logickými úrovňami cez rozširujúce konektory je potrebný zdroj s napätím 2,5V.

Bloková schéma zdroja napájacích napätí na obr. 2 znázorňuje usporiadanie jednotlivých napájacích vetiev, rozmiestnenie ochranných prvkov a filtrov. Na vstup zdroja je umiestnený blok s názvom „ochrana“ ktorý zabráňuje poškodeniu zariadenia pri prepätí na jeho vstupe, alebo pri zvýšenom odbere prúdu hlavne v prípade skratu. V praxi sú použité prvky transil a polyswitch. Blok s názvom „filter“ predstavuje filtračné prvky určené na odstránenie vyšších harmonických zložiek. Každá napájacia vetva má svoj zdroj napätia a je tiež možné použiť externý zdroj napájacieho napätia. Výstupy jednotlivých vetiev sú ošetrené filtermi a na signalizáciu jednotlivých napájacích napätí sú použité LED diódy. Okrem vetvy 2,5 V určenej na napájanie analógovej časti obvodu fázového závesu sú všade použité impulzné zdroje.

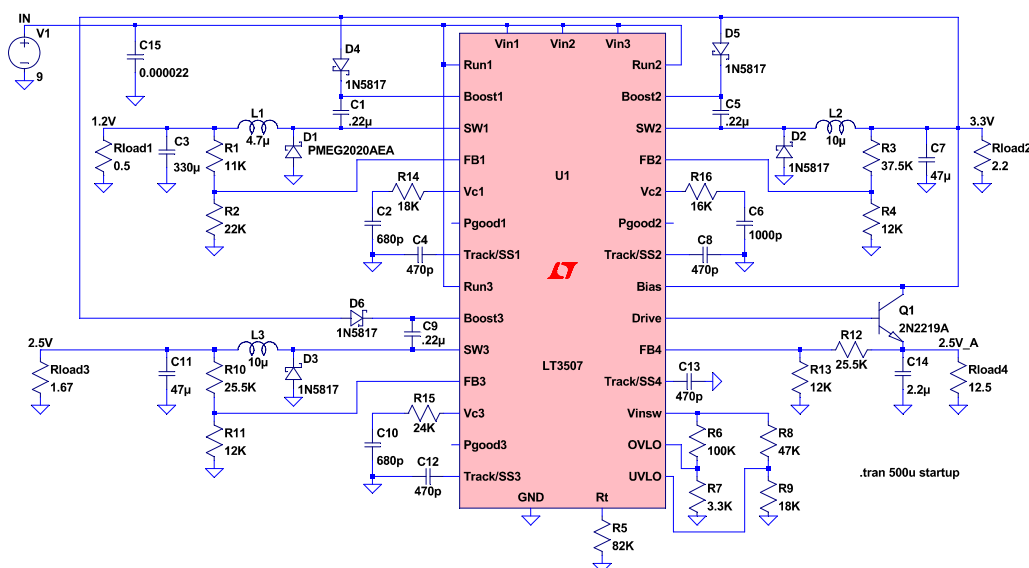
Pre dimenzovanie zdroja napájacích napätí je nutné poznať nielen výstupné napätia, ale aj prúdy ktoré odoberajú jednotlivé vetvy napájacieho zdroja. Spoločnosť Altera [2] na svojej internetovej stránke ponúka program PowerPlay Early Power Estimator [3] aj pre rodinu obvodov Cyclone III. Tento program umožňuje výpočet odhadu spotreby pre jednotlivé typy

FPGA obvodov na základe výstupného súboru generovanom programom Quartus II [4]. Na zistenie maximálnych odoberaných prúdov FPGA obvodu bola v prostredí Quartus II definovaná konfigurácia FPGA obvodu tak, aby bol maximálne využitý potenciál FPGA obvodu.



Obr. 2 Bloková schéma napájacieho zdroja vývojovej dosky

Na základe analýzy spotreby FPGA obvodu a všetkých periférií nachádzajúcich sa na vývojovej doske bol na realizáciu zdroja napájacích napätí vybraný integrovaný obvod LT3507 [6] od spoločnosti Linear Technology [5], ktorý je charakteristický širokým rozsahom napájacieho napätia od 4 V do 36 V, má jeden 2,4 A a dva 1,5 A impulzné stabilizátory a jeden LDO (Low Drop Out) stabilizátor. Vzhľadom na frekvenciu spínania od 200 kHz do 2,5 MHz je možné použiť rozmerovo menšie kondenzátory a cievky. Zdroj napájacích napätí s obvodom LT3507 [6] je realizovaný podľa zapojenia v katalógovom liste, no je potrebné vypočítať a upraviť hodnoty niektorých súčiastok tak, aby boli dosiahnuté požadované hodnoty výstupných napätí a prúdov. Správnosť návrhu napájacieho zdroja s obvodom LT3507 (obr. 3) bola overená pomocou programu LTspice IV verzia 4.06 [7], ktorý slúži na simuláciu zapojení s obvodmi od spoločnosti Linear Technology. Na stránke spoločnosti Linear Technology sú k dispozícii modely jednotlivých obvodov ktoré si užívateľ môže stiahnuť a simulovať v programe LTspice.

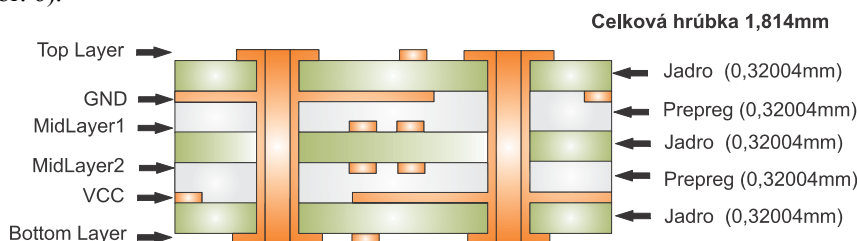


Obr. 3 Zapojenie napájacieho zdroja s obvodom LT3507 v programe LTspice

IV. NÁVRH A SIMULÁCIA DOSKY PLOŠNÉHO SPOJA

Ďalší krok k realizácii vývojovej dosky predstavuje návrh dosky plošného spoja. Pre návrh DPS bol zvolený návrhový program Altium designer 2009 [8], ktorého skúšobnú „trial“ verziu je možné získať zo stránky spoločnosti Altium [9]. Program umožňuje okrem iného návrh schémy zapojenia, návrh dosky plošného spoja a simuláciu návrhu.

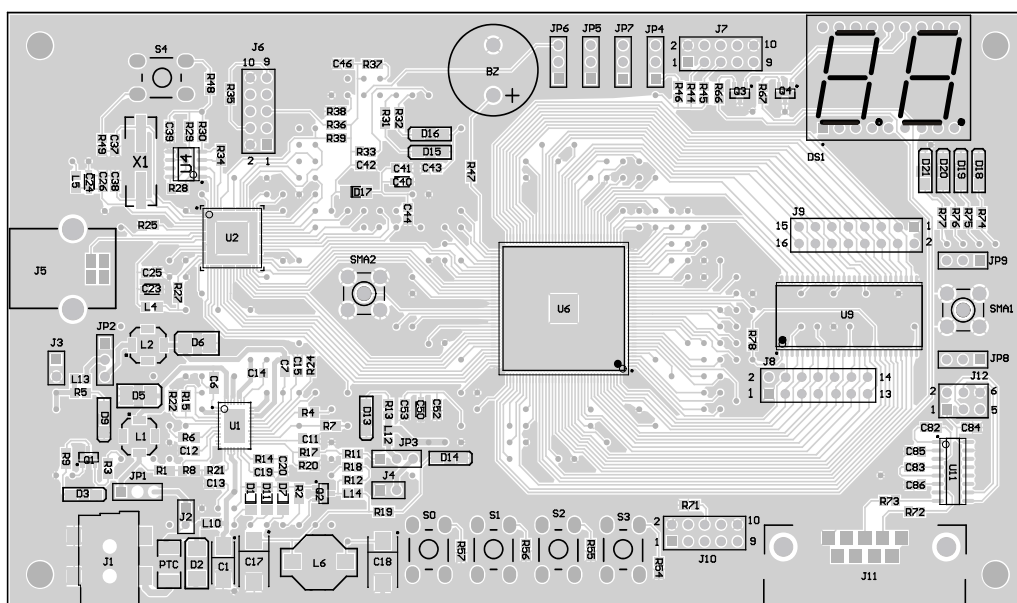
Na základe zložitosti zapojenia a potreby dobrého tienenia vývojovej dosky bol zvolený šesťvrstvový plošný spoj. Dve vrstvy sú použité na realizáciu napájacej vrstvy GND a 3,3 V napájacej vrstvy VCC ako je možné vidieť na obr. 4. Tieto dve vrstvy vytvárajú tienenie vývojovej dosky a tiež slúžia na distribúciu napájania pre prvky umiestnené na DPS. Najvyššiu precíznosť návrhu vyžadovalo pripojenie SDRAM [10], nakoľko SDRAM pracuje s vysokými frekvenciami signálov, ktoré sa približujú k oblasti rádiových frekvencií. Preto je opodstatnená otázka dobrého tienenia DPS, oneskorení a odrazov vo vodičoch pripojených k SDRAM. Vodiče pripojené k SDRAM boli vedené najkratšími možnými cestami a v blízkosti vrstvy so signálovými vodičmi SDRAM bola umiestnená tieniacia napájacia vrstva pripojená na GND napájacieho zdroja. Nevyužitú plochu jednotlivých vrstiev DPS boli vyplnené tzv. rozliatou medťou pripojenou k GND, ktorá prispieva k zlepšeniu tienenia DPS (obr. 5, obr. 6). Vodiče zdroja napájacieho napätia a silové časti vývojovej dosky boli riešené formou delených plôch a viacnásobných prekovených otvorov v dôsledku prúdového zaťaženia týchto vodivých ciest (obr. 5, obr. 6).



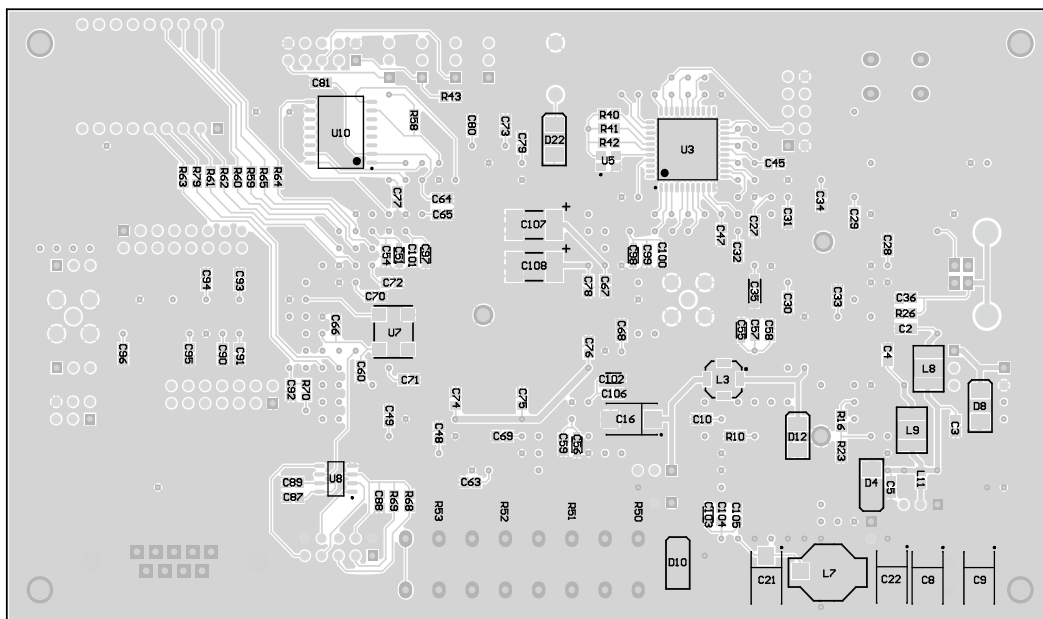
Obr. 4 Usporiadanie vrstiev DPS

Pri návrhu DPS sa dbalo na správnu distribúciu elektrickej energie rozmiestnením filtračných a blokovacích kondenzátorov na vývojovej doske, ktoré majú za úlohu pokrývať rýchly odber energie a tiež plnia úlohu filtra pri impulzných zmenách odberu prúdu. Pri každom napájacom vývode integrovaných obvodov bol umiestnený v čo možno najmenšej vzdialenosti malý keramický blokovací kondenzátor s dielektrikom X7R a tantalový kondenzátor s vyššou kapacitou. Každý z týchto kondenzátorov sa uplatní pre iné frekvencie, čo prispieva k vyššej stabilite zapojenia. Obmedzenie kolísania odoberaného prúdu tiež prispieva k obmedzeniu rušenia generovaného vývojovou doskou.

Správne pripojenie SDRAM na DPS je pre jej funkčnosť kritické. Program Altium designer 2009 [8] ponúka okrem iného aj tzv. Signal Integrity [11] analýzu, ktorá umožňuje skúmať vlastnosti vodičov DPS z pohľadu prenosu signálu.



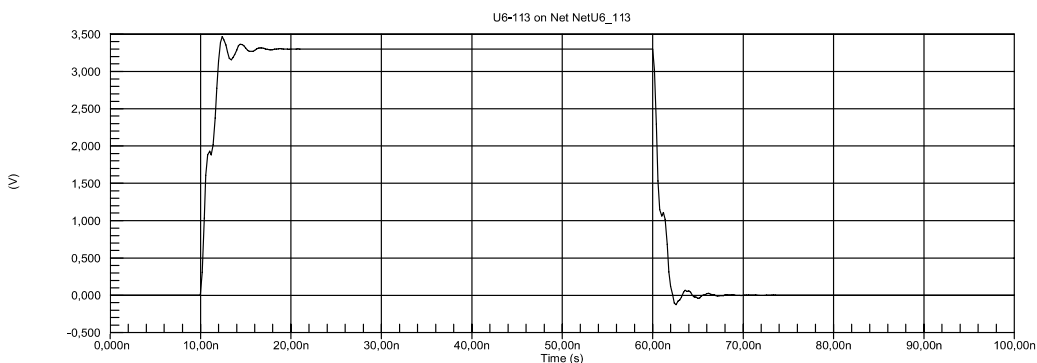
Obr. 5 Pohľad na DPS vývojovej dosky zo strany Top (rozмеры DPS : 157,5 mm x 90 mm)



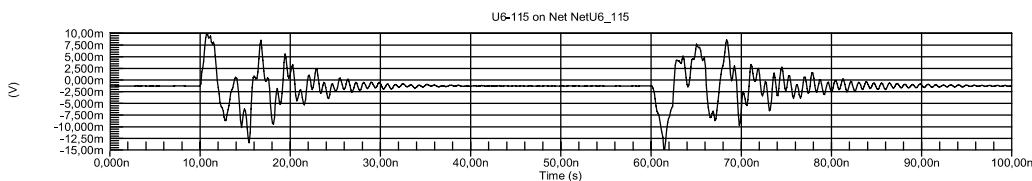
Obr. 6 Pohľad na DPS vývojovej dosky zo strany Bottom

Pred vykonaním Signal integrity analýzy niektorého vodiča medzi dvomi obvodmi ako napríklad v tomto prípade medzi FPGA (obr. 5-U6) obvodom a SDRAM (obr. 5-U9), je potrebné priradiť súčiastkam modely špecifikácie správania sa vstupno-výstupných vývodov (IBIS - IO Buffer Interconnect Specification) [12]. IBIS model súčiastky je možné získať priamo na stránke výrobcu, alebo konkrétne pri Altera FPGA obvodoch vygenerovať IBIS model pre danú aplikáciu FPGA programom Quartus II [4].

Po priradení IBIS modelov súčiastok, ktorých vzájomná komunikácia je objektom skúmania, je možné spustiť Signal Integrity analýzu. Analýza integrity signálu programu Altium designer 2009 [8] predstavuje komplexný simulačný nástroj ktorý umožňuje zistiť napríklad impedanciu, dĺžku vodiča, oneskorenie na tomto vodiči, amplitúdu zámkitov vo vodiči, zobrazíť testovací impulz ovplyvnený odrazmi vznikajúcimi v ľubovoľnom analyzovanom vodiči (obr. 7), zobrazíť rýchlu Fourierovú transformáciu tohto signálu a tiež umožňuje analyzovať presluchy vznikajúce v danom vodiči od iného vodiča ako je znázornené na obr. 8.



Obr. 7 Analýza odrazov signálu vo vodiči pomocou Signal Integrity analýzy



Obr. 8 Analýza presluchov signálu vo vodiči pomocou Signal Integrity analýzy

Na základe analýzy integrity signálu bolo zistené, že navrhnutá DPS by mala umožniť spoľahlivú komunikáciu medzi FPGA obvodom a SDRAM. Napriek tomu, že ide len o simuláciu DPS a správnosť návrhu sa overí až po realizácii vývojovej dosky v reálnych

podmienkach, analýza integrity signálu poskytuje návrhárovi užitočný nástroj pre overenie návrhu náročných vysokorychlostných digitálnych systémov.

V. ZÁVER

Hlavným cieľom prezentovaného návrhu bolo vykonať kvalitný návrh vývojovej dosky na báze Altera FPGA Cyclone III. Ukázalo sa, že návrh dosky plošného spoja, hlavne z pohľadu pripojenia SDRAM, si vyžaduje znalosti z oblasti návrhu vysokorychlostných digitálnych systémov a zvládnutie tejto problematiky je pre funkčnosť zapojenia kľúčové. Preto je výhodné využívať možnosti moderných vývojových CAD prostriedkov pre simuláciu vlastností DPS z pohľadu integrity signálu. Výstup tejto práce poskytuje podklady pre fyzickú realizáciu vývojovej dosky, kde sa ponúka možnosť porovnať výsledky simulácií s hodnotami nameranými v reálnej aplikácii. Podrobnejšie informácie a podklady realizovaného návrhu je možné nájsť v [13].

VI. POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) (50%) a projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) (50%) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Cyclone III Device Handbook Volume 1, cyclone3_handbook.pdf, verzia 3.3, január 2010, Altera. Dostupné na internete: <<http://www.altera.com/products/devices/cyclone3/literature/cy3-literature.jsp>>
- [2] Internetová stránka spoločnosti Altera. <<http://www.altera.com>>
- [3] PowerPlay Early Power Estimator, dostupné na internete: <http://www.altera.com/support/devices/estimator/cy3-estimator/cy3-power_estimator.html?GSA_pos=2&WT.oss_r=1&WT.oss=PowerPlay%20Early%20Power%20estimator>
- [4] Quartus II, Introduction to the Quartus II software. Dostupné na internete: <http://www.altera.com/literature/manual/intro_to_quartus2.pdf?GSA_pos=2&WT.oss_r=1&WT.oss=quartus%20II%209.0%20intro>
- [5] Internetová stránka spoločnosti Linear Technology. <<http://www.linear.com>>
- [6] LT3507 : Triple Monolithic Step-Down Regulator with LDO, 3507fa.pdf, Linear Technology Corporation 2008. Dostupné na internete: <<http://cds.linear.com/docs/Datasheet/3507fa.pdf>>
- [7] Internetová stránka Linear Technology venovaná simulačnému programu LTspice IV (verzia 4.06) s možnosťou jeho získania: <<http://www.linear.com/designtools/software/ltspice.jsp>>
- [8] Altium designer 2009, možnosť získať ako trial verziu. Dostupné na internete: <http://www.altium.com/products/altium-designer/en/altium-designer_home.cfm>
- [9] Domovská stránka spoločnosti Altium s informáciami o produktoch: <<http://www.altium.com/>>
- [10] K4S561632 : 256Mbit SDRAM. 62360_1.pdf, rev. 0.1, Aug. 2002, Samsung electronics. Dostupné na internete: <http://www.datasheetcatalog.org/datasheets2/62/62360_1.pdf>
- [11] Signal Integrity, Performing Signal Integrity Analyses, tutorial v1.2, December 12, 2005, dostupné na internete: <<http://www.altium.com/files/AltiumDesigner6/LearningGuides/TU0113%20Performing%20Signal%20Integrity%20Analyses.PDF>>
- [12] Ibis Models: Background and Usage, informácie o IBIS modeloch, dostupné na internete: <http://www.actel.com/documents/Ibis_AN.pdf>
- [13] J.Macej, "Vývojová doska na báze Altera FPGA Cyclone III," Bakalárska práca, TUKE FEI KEMT 2010.

Radarové systémy v automobiloch

Ján GAMEC, Iveta GLADIŠOVÁ, Mária GAMCOVÁ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

jan.gamec@tuke.sk, iveta.gladisova@tuke.sk, maria.gamcova@tuke.sk

Abstrakt — Článok sa zaoberá analýzou problematiky radarových senzorov v automobiloch. Aplikácia mikrovln a milimetrových vln (MiliMeter Waves) v radarových systémoch má obrovský potenciál v automobilovom priemysle a bude zohrávať dôležitú bezpečnostnú úlohu v moderných inteligentných automobilových a transportných systémoch. Spracovanie informácií o vzdialenosti, rýchlosti a uhle tvorí tri hlavné úlohy pri spracovaní signálov z automobilového radaru.

Kľúčové slová — adaptívny tempomat, radarový senzor, systém kolíznej výstrahy

I. ÚVOD

S narastajúcou hustotou premávky na pozemných komunikáciách narastá aj dopyt po stále sofistikovanejších a hlavne cenovo dostupnejších prvkoch aktívnej bezpečnosti, ktoré by znížili počet dopravných nehôd a smrteľných zranení. Takýmto, a dnes už v automobiloch vyšších tried dostupnými systémami, sú moderné radarové systémy [1], často v kombinácii s obrazovými senzormi. Ich hlavnou úlohou je predísť kolíznym situáciám v stále hustejšej a dynamickejšej premávke. Na rozdiel od pasívnych systémov bezpečnosti, ktoré reagujú vtedy, keď kolízia už nastala, radarový systém môže upozorniť na riziko zrážky, a tak ju vylúčiť, respektíve minimalizovať jej následky. Bývajú súčasťou adaptívnych tempomatov, systémov snímajúcich vybočenie z jazdného pruhu, systémov snímajúcich mŕtvy uhol spätného zrkadla a pod. Za posledných pár rokov sa pre automobilové radary stala dôležitou a zaujímavou oblasť milimetrových vln (MMW- MiliMeter Waves) [2], [3].

Existujú dva dominantné faktory, ktoré predurčujú MMW technológiu pre použitie v automobilových radaroch: *cena a rozmery*. Kľúčovým faktorom pre spotrebiteľa je nízka cena, ktorá z radaru robí dôležitý bezpečnostný a cenovo dostupný prvok. Minimálne rozmery sú základom pre jednoduchú integráciu radaru do vozidla bez nutnosti väčšieho zásahu do dizajnu a výkonu. Použitie milimetrových vln - MMW (30-300 GHz) ponúka oproti mikrovlnám v pásme 0,3-30 GHz (frekvenčné pásmo je preplnené, väčšie rozmery antén) množstvo výhod: *vyššia smerovosť, redukcia rušenia a interferencií, kompaktnejšie a ľahšie zariadenie (obzvlášť antény), vyššia presnosť merania rýchlosti, lepšie rozlíšenie v diaľke, koexistencia niekoľkých systémov blízko seba redukujúca rušenie (v dôsledku pomerne veľkej absorpcie prostredím) a dobré uhlové rozlíšenie*.

Zavedenie nových a lacných radarových systémov s lepšími charakteristickými vlastnosťami je podmienené inováciou polovodičových výrobných procesov s vysokou opakovateľnosťou a efektívnosťou, integrovaných MMW obvodov s veľkou účinnosťou aj citlivosťou a moderných anténových riešení. Pokroky v GaAs alebo kremíkových integrovaných obvodoch operujúcich v pásme milimetrových vln dovoľujú zmenšiť rozmery, znížiť hmotnosť a výrobné náklady. Lepšie rozlíšenie radaru a spolupráca s inými senzormi (obrazovými) zvýši výkonnosť a funkčnosť automobilových elektronických systémov budúcnosti, ktoré vytvoria nové inteligentné prvky ako napr. autonómne vedenie vozidla.

II. ŠTANDARDIZAČNÉ PROCESY

Oblasť milimetrových vln pokrýva frekvenčné pásmo od 30 do 300 GHz, čomu zodpovedá vlnová dĺžka od 10 mm do 1 mm.

V roku 1979 konferencia *WARC (World Administrative Radio Conference)* po prvý krát ustanovila pre radary 10 radarových pásiem situovaných nad 40 GHz. IEEE zaviedla dokument, v ktorom oblasť 33-36 GHz pre radar pomenovala ako pásmo Ka, oblasť od 46 do

56 GHz ako pásmo V a 56 - 110 GHz ako pásmo W. Ostatné oblasti pásma milimetrových vln neboli dokumentom presne špecifikované [3].

CEPT (The European Conference of Postal and Telecommunications Administrations - Európska konferencia poštových a telekomunikačných administratív) reprezentuje 43 európskych regulačných inštitúcií. CEPT v rozhodnutí (92)02 Európskeho rádiokomunikačného výboru (**European Radiocommunications Committee**) [4] stanovila, že pásmo 76 – 77 GHz by malo byť určené pre radarové systémy automobilov. Toto rozhodnutie taktiež podporila správa č.3 Európskeho rádiokomunikačného výboru „Harmonizácia frekvenčných pásiem určených pre dopravné informačné systémy“.

Ustanovenia **FCC (Federal Communications Commission – Federálny komunikačný úrad)** v časti 15.253 [5] podporujú používanie frekvenčných pásiem 46,7 – 46,9 GHz a 76 – 77 GHz v rámci USA pre automobilové senzory a pre použitie v automobilových radarových systémoch. Časť 15.255 [6] FCC výslovne zakazuje používanie „senzorov, ktoré ovplyvňujú elektromagnetické pole, vrátane automobilových radarov“ v pásme 59 – 64 GHz, s výnimkou špecifikovaných činností.

Norma EN 301 091 (1998-06) **ETSI (European Telecommunications Standards Institute - Európskeho inštitútu pre telekomunikačné normy)** špecifikuje požiadavky [7] pre 76 GHz až 77 GHz radar, ktorý je určený pre aplikácie Telematiky cestnej dopravy a prepravy (Road Transport and Traffic Telematics - RTTT), medzi inými aj systémy automobilových tempomatov (Automotive Cruise Control - ACC), výstrahy pred kolíziou (Collision Warning - CW) a antikolízne systémy (Anti-Collision - AC).

Odporúčanie M.1310 **ITU-R (International Telecommunication Union - Medzinárodná telekomunikačná únia)** v sebe zahŕňa požiadavky [8] z hľadiska **TICS** systémov dopravných informácií a riadenia dopravy - (**Transport Information and Control Systems - Informačné a riadiace systémy dopravy**). TICS obsahujú systémy riadenia v núdzových situáciách, služby elektronického mýta, systémy riadenia vozového parku, systémy verejnej dopravy (kyvadlová doprava) a pokročilé systémy riadenia vozidiel. Automobilový radar je klasifikovaný ako element pokročilých systémov riadenia vozidiel. Predbežné odporúčanie ITU-R navrhuje pre automobilové radarové systémy pásma 60 - 61 GHz a 76 - 77 GHz. Toto predbežné odporúčanie opisuje technické a pracovné charakteristiky radaru s nízkym výkonom, ktorý je súčasťou väčšieho dopravného a riadiaceho systému. Taktiež sa odvoláva na to, že:

- FCC používa pre tento účel pásmo 76 – 77 GHz,
- Japonské Ministerstvo pôšt a telekomunikácií používa pre tento účel pásma 60 - 61 GHz a 76 – 77 GHz,
- dokument ETSI EN 301 091 (1998-06) [7], ktorý hovorí o aplikácii pásma 76 - 77 GHz pre účely automobilového radaru a
- **ASTAP (Asia-Pacific Telecommunications Standardization Program – Ázijsko – pacifický telekomunikačný normalizačný program)** schválil predbežný návrh na normu pre „Automobilový radar s nízkym výkonom operujúci v pásmach 60 - 61 GHz a 76 - 77 GHz“.

Zhrnutie frekvenčných pásiem pre automobilové radary a inštitúcií, ktoré ich podporujú, je uvedené v Tab. 1.

Tab. 1
FREKVENČNÉ PÁSMO URČENÉ PRE AUTOMOBILOVÉ RADARY

Podporované frekvenčné pásma	Organizácia
76 - 77 GHz	CEPT (Európa)
76 - 77 GHz	ETSI (Európa)
46.7 - 46.9 GHz, 76 - 77 GHz	FCC (USA)
60 - 61 GHz, 76 - 77 GHz	ITU-R
60 - 61 GHz, 76 - 77 GHz	ASTAP (Ázia, Tichomorie)

III. APLIKÁCIE AUTOMOBILOVÝCH RADAROV A ICH KLASIFIKÁCIA

Vývoj prediktívnych systémov zamedzujúcich potenciálnej zrážke (CAS - Crash Avoidance Systems) je momentálne zameraný na systémy, ktoré by boli absolútne nezávislé na vozidle, v ktorom sa nachádzajú, a takisto, aby nevyžadovali externú podporu od ostatných dopravných prostriedkov alebo od cestnej infraštruktúry. Pre kooperatívne a plnoautomatické cestné systémy ďalšieho desaťročia budú nevyhnutné autonómne systémy, ktoré budú alebo sú nasadené v týchto aplikáciách: 1) *adaptívny tempomat (Adaptive Cruise Control - ACC)*; 2) *systém predčasného varovania použitý pri riadení spúšťania airbagov, predpínačov bezpečnostných pásov atď.*; 3) *antikolízny systém*; 4) *systém „Stop and Go“*; 5) *systém výstrahy*

zmeny jazdného pruhu; 5) systém monitorovania mŕtveho uhla; 6) asistent parkovania a cúvania.

Automobilové radarové systémy môžeme z hľadiska rozsahu, v ktorom sú schopné pracovať, rozdeliť na **LRR** (Long Range Radar) – radar dlhého dosahu, ktorý využíva napr. systém ACC a **SRR** (Short Range Radar) – radar krátkého dosahu, ktorý zahŕňa senzory snímajúce zmenu jazdného pruhu, senzory snímajúce mŕtvy uhol spätného zrkadla, a pod. Systémy fungujúce na princípe LRR operujú vo frekvenčnom pásme 76 -77 GHz, zatiaľ čo systémy na báze SRR operujú v pásme aj v pásme 24 GHz, aj keď už boli prijaté rozhodnutia a opatrenia, ktoré by mali zabezpečiť prechod na pásmo 79 GHz. Kombináciou oboch týchto radarových systémov sa z automobilu stáva oveľa bezpečnejší a komfortnejší dopravný prostriedok, nehovoriac o prínosoch pre životné prostredie.

Hlavné parametre jednotlivých kategórií radarov sú [19]:

- LRR
 - o dosah do 150 – 200 m,
 - o detegovateľná rýchlosť nad 30 km/h do 250 km/h,
 - o úzky vyžarovací diagram pre riadenie jazdnej dráhy pred vozidlom, tak aby sa určila vzdialenosť ku pred nim idúcemu vozidlu (pre udržiavanie minimálnej bezpečnej vzdialenosti),
 - o šírka frekvenčného pásma menej ako 1GHz a typické priestorové rozlíšenie 0,5 m.
- SRR
 - o dosah do 30 m,
 - o detegovateľná rýchlosť nad 5 km/h do 150 km/h,
 - o široký vyžarovací diagram,
 - o šírka frekvenčného pásma menej ako 5GHz a typické priestorové rozlíšenie 0,1 m.

Pre kategóriu LRR sa častejšie uvažuje o FMCW (**F**requency **M**odulated **C**ontinuous **W**ave - frekvenčne modulovaná spojitá vlna) technike a pre kategóriu SRR aj o UWB (**U**ltra **W**ide **B**and – extrémne široké pásmo) technike.

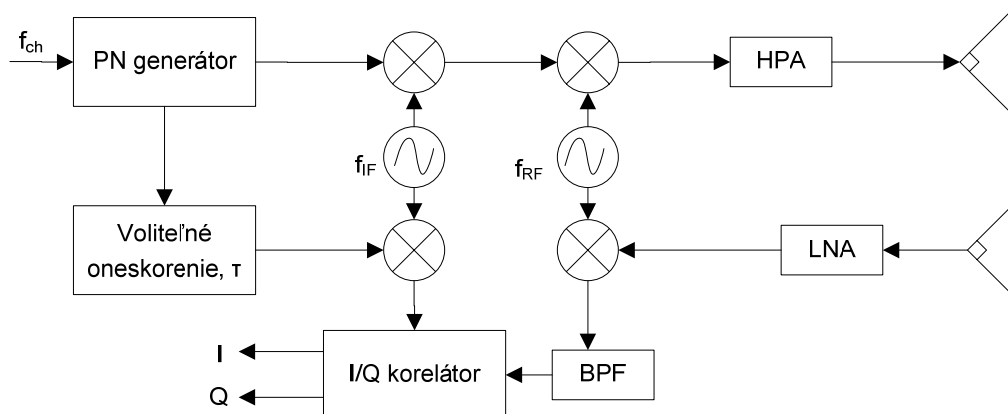
IV. KONCEPCIA UWB RADAROV

Časť automobilového radaru, v ktorom sa spracováva signál má tri hlavné úlohy:

- Spracovanie informácií o vzdialenosti.
- Spracovanie informácií o rýchlosti.
- Spracovanie informácií o uhle.

Požiadavky kladené na tieto úlohy sú presnosť a doba spracovania nameraných údajov. Bloková schéma automobilového UWB radaru s je na Obr. 1.

Vzdialenosť od vozidla k cieľu je vypočítavaná pomocou polovice času, ktorý je potrebný na prekonanie vzdialenosti postupujúcou elektromagnetickou vlnou medzi vysielačom, cieľom a prijímačom. Tento čas sa vypočítava pomocou rôznych techník výpočtu korelácie medzi prijatou pseudonáhodnou (PN) postupnosťou a referenčnou PN postupnosťou, ktorá je známa v prijímači. Presnosť výpočtu vzdialenosti je nepriamo úmerná dobe trvania jedného symbolu PN postupnosti (chip) a priamo úmerná počtu jej symbolov [9], [10].



PN (PseudoNoise) – pseudo náhodná postupnosť, HPA (High Power Amplifier) – výkonový zosilňovač, LNA (Low Noise Amplifier) - zosilňovač s nízkym šumom, BPF (Band-Pass Filter) – pásmový priepust.

Obr. 1 Bloková schéma automobilového UWB radaru

Výhody: 1) jednoduchá implementácia, 2) odpadá problém so synchronizáciou referenčnej PN postupnosti a vysielanej PN postupnosti (pretože obe využívajú rovnaký zdroj signálu), 3) krátky čas výpočtu, ktorý nie je závislý na „chipovej“ rýchlosti.

Nevýhody: 1) presnosť je limitovaná šírkou frekvenčného pásma, ktorá môže byť riadená modulátorom / demodulátorom v IF časti, zatiaľ čo stav fázy a amplitúdy sa uchováva, 2) prílišné zväčšenie frekvenčného pásma prispieva k zvýšeniu šumu, 3) pevne stanovená presnosť (okrem prípadu variabilnej „chipovej“ rýchlosti [11]).

Metódy výpočtu korelácie však čelia problému, a to že presnosť je fixná a po implementácii systému sa nedá meniť, čo je veľmi významná požiadavka pri prepínaní systému z fázy detekcie do fázy sledovania pohybu objektov [12], [13].

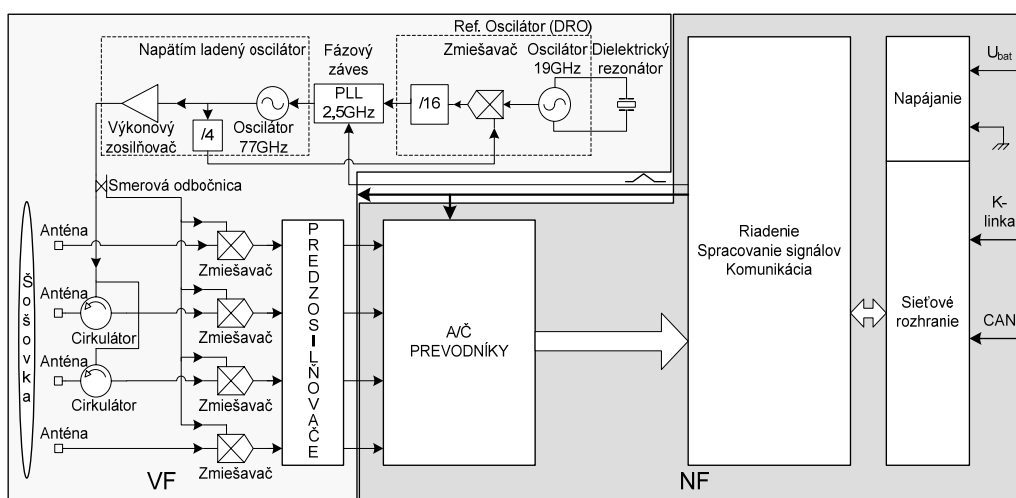
Pri **spracovaní údajov o rýchlosti** v radare sa využíva fakt, že keď sa signál odrazí od pohybujúceho sa objektu, je frekvencia vysielaného signálu ovplyvnená Dopplerovým posuvom. Konvenčným postupom je možné Dopplerovu frekvenciu vyrátať výpočtami rýchlej Fourierovej transformácie (FFT), avšak problémom tohto konvenčného postupu je, že sa nedá určiť znamienko Dopplerovej frekvencie, to znamená že nie je jasné, či sa objekt približuje alebo vzdďaľuje. Ďalším problémom je, že doba merania zázneových signálov je závislá na rozlíšení z hľadiska rýchlosti, a preto každá snaha o zvýšenie rozlíšenia z hľadiska rýchlosti so sebou prináša zvýšenie meracieho času. To zvyšuje nároky na systém. Problémom metódy pre meranie Dopplerovej frekvencie využitím výhod gradientných charakteristík frekvencie je skutočnosť, že zázneový signál je najprv vedený cez hornopriepustný filter kvôli odstráneniu nízkofrekvenčných zložiek [14]. Tým sa však strácajú dôležité údaje v situácii, keď sa cieľ pohybuje rovnakou rýchlosťou ako vozidlo s radarom.

Spracovanie údajov informujúcich o uhlovej pozícii cieľa je nevyhnutné na to, aby sa detegované objekty dali do súvislosti s ich okolím, a aby tieto objekty boli klasifikované ako nebezpečné. Na určenie potenciálnych kolíznych trajektórií, musia byť objekty, ako aj ich aktuálne a predpokladané trajektórie, vopred odhadnuté a navyše priradené k jednotlivým jazdným pruhom. Automobilové proti-kolízne radarové systémy sú predmetom skúmania už dlhé roky, ale až experimenty ukázali, že množstvo nesprávne vyhodnotených informácií - falošných alarmov, v tomto systéme je príliš vysoké. Jednou z príčin sú nedostatky pri rozlíšení v uhle. Tento nedostatok sa rieši anténnymi systémami so syntetickou apertúrou alebo fázovanými anténnymi sústavami so šošovkami.

V. KONCEPCIA FMCW RADAROV

FMCW radar (**F**requency **M**odulated **C**ontinuous **W**ave - frekvenčne modulovaná spojitá vlna) je radar s maximálnym uvádzaným dosahom 200 metrov pre aplikácie v automobiloch. Jedná sa o tzv. homodynový radar využívajúci moduláciu FMCW.

Na Obr.2 je schéma prototypu homodynového FMCW radaru typu LRR3 (Bosch) [19], [20].



Obr. 2 FMCW radarový systém automobilu s homodynovým prijímačom (Bosch: LRR3-Prototype)

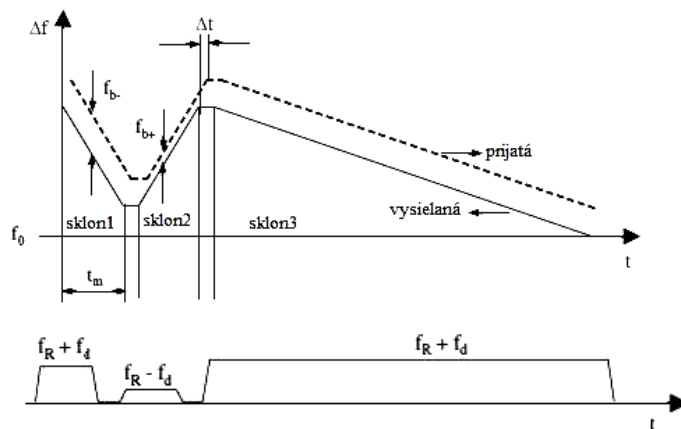
Výhody FMCW radaru v porovnaní s inými typmi radarov sú: 1) modulácia (FMCW) je kompatibilná so širokou škálou polovodičových vysieláčov, 2) meranie frekvencie nutnej pre meranie vzdialenosti môže byť realizované digitálne použitím FFT, 3) bežnými prijímačmi je pomerne ťažké zachytiť FMCW signály.

Medzi ďalšie výhody FMCW radaru patria: tvar vlny FMCW (poskytuje jednoduchú moduláciu), vysoký priemerný výkon, široké frekvenčné pásmo (umožňuje veľmi dobré rozlíšenie v diaľke a Dopplerovo spracovanie), dobré vlastnosti aj na krátkych vzdialenostiach, vysoká presnosť. Taktiež môže pracovať na extrémne krátkych vzdialenostiach, umožňuje použitie nízkonapäťového napájania, kompaktnej vysielačej jednotky a vysokofrekvenčného prvku vysielača [3], [15].

Hlavnými *nevýhodami* FMCW radaru sú: 1) vysoké nároky na výrobu hardvéru, 2) pre požadovanú presnosť je potrebný výpočet FFT s veľkým počtom bitov na vzorku, 3) privysoká spotreba energie (výkonné procesory pre spracovanie signálov), 4) rozsiahly softvér pre riešenie algoritmov, 5) rozptýlený signál a jednosmerný posun signálov môžu zamaskovať malé objekty [16], [18].

Vzdialenosť k cieľu a rozlíšenie v diaľke sa v FMCW systémoch získava rozmiataním vysielanej frekvencie, čo je presnejší spôsob, ako generovať, vysielat' a prijímať veľmi úzke impulzy. Z toho vyplýva, že veľmi vysoké rozlíšenie vo vzdialenosti je oveľa ľahšie dosiahnuť metódou FMCW ako impulzovým radarom. Kritickou úlohou pri metóde FMCW je zaistiť linearitu frekvenčného rozmiatania 76,5 GHz zdroja. To je dôvod použitia fázového závesu (Phase Locked Loop – PLL). U novších návrhov sa používa napätím riadený oscilátor na báze mikrovlnných monolitických integrovaných obvodov [3].

Frekvencia nepretržite vysielaného FMCW signálu je modulovaná ako funkcia času s periodickým tvarom modulačného signálu, akým je napríklad trojuholníkový signál (Obr. 3). Frekvencia vysielaného signálu je $f_i(t) = f_0 + \Delta f(t)$. Signál odrazený od cieľa sa oneskorí o čas Δt , ktorý je priamo úmerný vzdialenosti cieľa a má frekvenciu $f_r(t) = f_0 + \Delta f(t - \Delta t)$. Demoduláciou týchto dvoch frekvencií $f_i(t)$ a $f_r(t)$ v prijímači radaru vzniká výstupná frekvencia, ktorá je známa ako prechodná (stredná) frekvencia $f_R = \Delta f(\Delta t)$ alebo ako *záznejová frekvencia* $f_b = f_R$. Pre pohybujúci sa objekt bude záznejová frekvencia obsahovať okrem frekvencie f_R aj Dopplerovu frekvenciu f_d , t.j. $f_b = f_R + f_d$. V takom prípade je záznejová frekvencia iná pre klesajúci nábeh (sklon1) a iná pre stúpajúci nábeh (sklon2) modulačnej krivky na Obr.3. Výsledkom spriemernenia prechodnej (strednej) frekvencie je frekvenčný údaj, ktorý zodpovedá vzdialenosti cieľa, zatiaľ čo rozdielom sa získa informácia o rýchlosti cieľa [15].



Obr. 3 Schéma modulácie FMCW automobilového radaru a frekvenčný rozdiel medzi vyslaným a prijatým signálom (Bosch: LRR2-Prototype)

Rozptyl z vysielača k prijímaču a nežiaduce viacnásobné odrazy („clutter“), vznikajúce na prekážkach medzi vysielačom a cieľom, je prakticky najvýznamnejším faktorom, s ktorým treba uvažovať pri návrhu FMCW radaru. Tento sa nevyskytuje pri impulzových radaroch. Na zvýšenie dosahu merania v prítomnosti viacnásobných nežiaducich odrazov, môže byť použitá istá forma digitálnej filtrácie ako priestorový filter, ktorý by eliminoval nežiaduce odrazy.

Veľké uhlové pokrytie (charakteristické pri použití v automobiloch), spôsobuje, že každý senzor musí v rámci CFAR procedúry spracovať množstvo detekcií každého kanála radaru, čo je spôsobené situáciami s viacerými cieľmi a špecifickými situáciami. Experimentálne výsledky dokázali, že pre spoľahlivú **detekciu cieľa** v prostredí s viacerými cieľmi je vhodná metóda ordered-statistics OS - CFAR [15]. V reálnych aplikáciách sú viacnásobné odrazy komplikovaným variantom stochastického procesu, čo sa týka času a priestoru. Všetky tieto okolnosti si vyžadujú adaptívnu metódu v detekcii a spracovaní signálu, ktorá by pri detekčnom procese nepracovala s fixnou ale variabilnou prahovou hodnotou, a ktorá by rozhodovala v závislosti od lokálne pozorovaných nežiaducich odrazov o rôznej vzdialenosti, intenzite a fluktuácii. Pri reálnom použití radaru je priemerná hladina výkonu šumu a rušenia neznáma a

musí byť odhadnutá procesom detekcie ako prvá v poradí. Na tento odhad sa používa niekoľko známych CFAR techník, ako sú: CFAR s technikami posuvného okna, CA-CFAR (CA - Cell Averaging - spriemerňovanie bunky), CAGO-CFAR (CAGO - Cell Averaging with Greatest Of – spriemerňovanie bunky z ich najväčších hodnôt), OS-CFAR. (OS - Ordered Statistic – poriadková štatistika), WCA-CFAR („vyvažovaná“ CA-CFAR z angl. „weighted“) a iné [15].

VI. ZÁVER

Na podčiarknutie dôležitosti a opodstatnenosti radarových systémov v automobile je potrebné poukázať na štúdiu SeiSS (Socio Economic Impacts of Intelligent Vehicle Safety System - Prieskumná štúdia o možných pozitívnych sociálnych a hospodárskych vplyvoch zavedenia inteligentných bezpečnostných systémov v cestných vozidlách), v ktorej sú naznačené značne pozitívne dopady zavedenia inteligentných bezpečnostných systémov [16],[17].

Automobilový radar je dôležitou časťou inteligentných dopravných systémov, od ktorých sa očakáva veľké zlepšenie v bezpečnosti cestnej dopravy a jej efektívnosti. ITS sú vo fáze celosvetového vývoja a v budúcnosti budú prostriedkom pre redukciu strát na životoch a obmedzenie nákladov. Bude potrebné spomínané navrhnuté riešenia rozšíriť aj o uvažovanie v situáciách s viacerými cieľmi, t.j. použitie systému anténnych polí pre identifikáciu cieľa v situáciách s viacerými jazdnými pruhmi [14].

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku / Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

LITERATÚRA

- [1] R. H. Rasshofer, K. Gresser: „Automotive Radar and Lidar Systems for Next Generation Driver Assistance Functions“. Advances in Radio Science, 3, pp. 205–209, 2005.
- [2] J. Grubert, J. Heyen, C. Metz, L. C. Stange, A. F. Jacob: „Planar millimeter wave radar frontend for automotive applications“. Advances in Radio Science 1: pp. 125–129, 2003.
- [3] William David: „FMCW MMW Radar for Automotive Longitudinal Control“. University of California, Berkeley, pp.1055-1425, Máj 1997. <http://www.path.berkeley.edu/path/publications/pdf/PRR/97/PRR-97-19.pdf>
- [4] ECC Decision of 15 March 2002 on the withdrawal of the ERC Decision (92)02 “Decision on the frequency bands to be designated for the co-ordinated introduction of Road Transport Telematic Systems”, www.erodocdb.dk/docs/doc98/official/word/DEC0202.doc
- [5] Federal Communications Commission § 15.253, http://edocket.access.gpo.gov/cfr_2007/octqtr/pdf/47cfr15.253.pdf
- [6] Federal Communications Commission § 15.255, http://edocket.access.gpo.gov/cfr_2009/octqtr/pdf/47cfr15.255.pdf
- [7] A Review of Automotive Radar Systems - Devices and Regulatory Frameworks, http://acma.gov.au/webwr/radcomm/frequency_planning/spps/0104spp.pdf
- [8] DSRC Standards Development in ITU-R and Japan, <http://www.itsforum.gr.jp/Public/E4Meetings/P06/SS30%20oyama.pdf>
- [9] W. Mezel, J. Buechler, a J. Taech, „An experimental 24 GHz radar using phase modulation spread spectrum techniques“. 28th European Microwave Conference, vol.2, pp.56-60, Oct. 1998.
- [10] V. Filimon and J. Buechler, „A pre-crash radar sensor system based on pseudo-noise coding“. Microwave Symposium Digest., 2000 IEEE MTT-S International, vol.3, pp.1415-1418, 2000.
- [11] S. Lindenmeier, S. Mayer, and D. Knopfle, „Communicating near range sensor system for automotive application“. MTT-S Workshop, Dig., 6, 2002.
- [12] M. Watanabe, K. Inomata, S. Noda, K. Okazaki, T. Fukae, H. Yamabuchi, „A sideways-looking radar and its measuring principles“. 31st European Microwave Conference, London 2001, pp.261-264, Sept 2001. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4140083>
- [13] O. A. Aly and A. Omar, „New implementation of the correlation function of the pn code for application in automotive radars“. Antennas and Propagation Society International Symposium, 2003. IEEE, vol.3, pp. 272-275, 22-27 June 2003. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1219841>
- [14] Alebel Arage Hassen: „Indicators for the Signal Degradation and Optimization of Automotive Radar Sensors under Adverse Weather Conditions“. PhD Thesis, Darmstadt 2006. <http://tuprints.ulb.tu-darmstadt.de/epda/000765/PhdThesis.pdf>
- [15] H. Rohling: „Some radar topics: Waveform design, range CFAR and target recognition“. Technical University Hamburg, Germany, 2005. <http://nato-us.org/sensors2005/papers/rohling.pdf>
- [16] G. Bootsma, L. Dekker, „Road to the Future“. Rijkswaterstaat, NL, april 2007.
- [17] “Towards Europe-wide Safer, Cleaner and Efficient Mobility: The First Intelligent Car Report.” Commission of the European Communities, Brusel 2007. <http://repositories.cdlib.org/its/path/reports/UCB-ITS-PRR-2001-13/>
- [18] World Intellectual Property Organisation, WO/2001/013141, „Neural network radar processor“, <http://www.wipo.int/pctdb/en/wo.jsp?IA=US2000022007&DISPLAY=DESC>
- [19] R. Lachner, „Development Status of Next Generation Automotive Radar in EU“. ITS Forum, Tokyo, 2009. <http://www.itsforum.gr.jp/Public/J3Schedule/P22/lachner090226.pdf>
- [20] D.Hoetzer, D. Freundt, B. Lucas, "Automotive Radar and Vision Systems – Ready for the Mass Volume Market". Department | 10/3/2008 | © 2008 Robert Bosch LLC and affiliates. All rights reserved. Vehicle Dynamics Expo 2008.

Elektronické systémy automobilov a trendy ich vývoja

Ján GAMEC, Iveta GLADIŠOVÁ, Mária GAMCOVÁ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

jan.gamec@tuke.sk, iveta.gladisova@tuke.sk, maria.gamcova@tuke.sk

Abstrakt — Rozmach elektronických systémov v automobiloch neustále narastá. Rezonujúcou ideou ich budúceho smerovania je pojem infotainment, ktorý v sebe skrýva spojenie termínov informácie, zábava a komunikácia. Jedným z východísk vývoja elektronických systémov v automobiloch je snaha integrovať a viacnásobne využívať hlavné časti softvéru a hardvéru od viacerých dodávateľov, pričom nové prevádzkové funkcie by mali byť realizované prostredníctvom interakcie pôvodne autonómnych jednotiek. Nové systémy musia byť vybavené prostriedkami pre digitálne prenosy, sieťovú prenositeľnosť funkcií a musia poskytovať možnosť údržby počas životnosti produktu. Prínosmi nástupu sietí vo vozidlách sú vyššia spoľahlivosť, viac prevádzkových funkcií za nižšiu cenu, štandardizácia prepojení a komponentov, rýchlejšie zavádzanie nových technológií a rozšíriteľnosť prevádzkových funkcií.

Kľúčové slová — infotainment, vložené elektronické systémy, senzory, siete vo vozidlách

I. ÚVOD

Elektronické systémy automobilov zaznamenali najväčší rozmach za posledných 25 rokov a podľa predpokladov ich podiel bude v nasledujúcej dekáde aj naďalej narastať. Hnacím motorom vývoja elektronických systémov boli tri faktory, a to bezpečnosť, čistota - ekologickosť a ekonomickosť. Tieto je možné rozvinúť ešte podrobnejšie, a to na: *bezpečnosť, zákonodarstvo, spotreba, komfort a dynamika*.

Bezpečnosť: Nárast požiadaviek zákazníkov na bezpečnosť viedol k skutočnosti, že airbag, ABS (Antilock Breaking System) a ESP (Electronic Stability Program) sú ponúkané takmer vo všetkých vozidlách minimálne ako voliteľná výbava. Kým súčasné najbežnejšie systémy zväčša poskytujú iba pasívnu bezpečnosť, nákladnejšie systémy dneška a systémy budúcnosti sa radia do kategórie aktívnych. Aktívne systémy bezpečnosti pracujú tak, aby zabránili vzniku nebezpečných situácií pre pasažierov. Súčasný inovatívny systém na trhu sú predikčné bezpečnostné systémy a patria sem napríklad brzdné asistenčné systémy, protikolízne výstražné systémy a núdzové brzdné systémy.

Zákonodarstvo: Spoločnosť prostredníctvom zákonodarcov vyzýva a núti inžinierov automobilového priemyslu už od začiatku 70-tych rokov minulého storočia k nepretržitému znižovaniu emisií. Právne požiadavky boli stanovené legislatívou USA (EPA, CARB) [1]. Vysoký štandard dosiahol taktiež v Európe a v otvárajúcich sa trhoch krajín ako sú Čína, Rusko, Ukrajina a pod. Na základe tohto tlaku sa v súčasnosti dosiahol veľmi dobrý stav v emisii výfukových plynov, emisii častíc a v zamedzovaní produkcie škodlivých materiálov, čím sa zvýšila kvalita životného prostredia, hlavne v husto osídlených oblastiach.

Ekonomickosť: Znižovanie spotreby paliva v automobiloch je primárne dosahované znižovaním emisií, zamedzovaním parazitných strát motora a používaním ľahších materiálov. Nasledujúcim krokom je snaha, aby motor pracoval v účinnejšej oblasti pracovného rozsahu a tiež aby brzdná energia bola rekuperovaná, ako je to možné pozorovať v hybridných koncepciách vozidiel. Samozrejme priame porovnanie dnešných automobilov, a tých spred 15 rokov, nie vždy vykazuje absolútnu redukciu spotreby, ale jeho vlastnosti (bezpečnosť, emisie, komfort, dynamika) sú celkovo efektívnejšie.

Komfort: Elektrické ovládanie okien, sedadiel, zrkadiel atď. sú v súčasnosti relatívne bežné. Takmer všetci používatelia sa pri kúpe nového vozidla rozhodujú pre klimatizáciu. Nízka hladina hluku a vibrácií v interiéri garantuje príjemnejšie podmienky jazdy, hlavne na dlhšie vzdialenosti. Vozidlá vybavené elektrickými motormi a elektronickými zariadeniami sa

postupne rozširujú a stávajú sa samozrejmosťou aj v modelových radoch menších rozmerov.

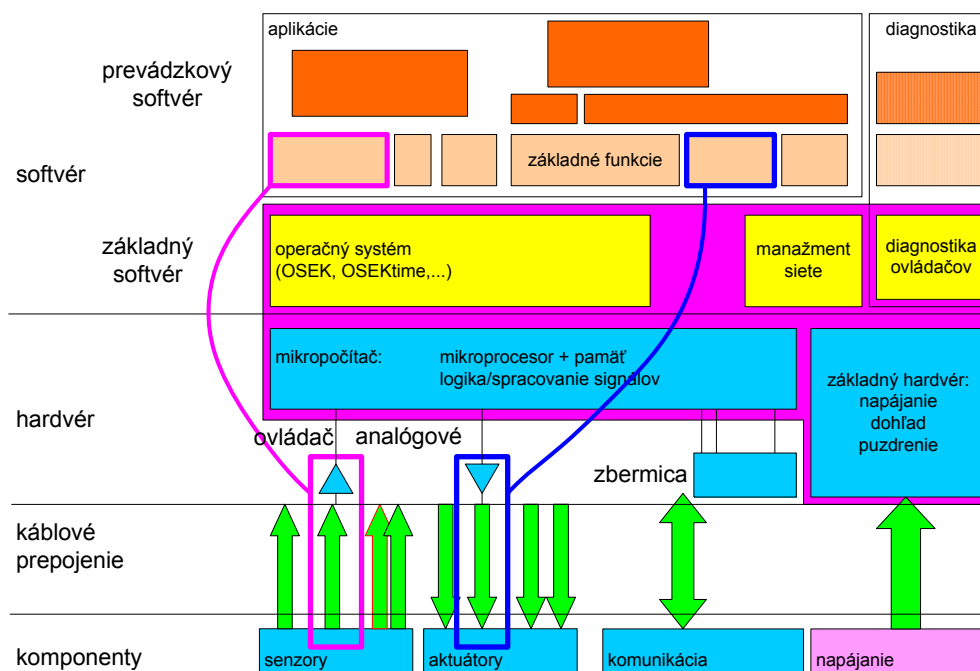
Dynamika: Automobil už nie je iba jednoduchý prostriedok pre dopravu. Zákazník sa ním identifikuje a vytvára si ním svoj imidž. Poskytovanie jazdnej dynamiky používateľovi sa stáva dôležitým obchodným faktorom výrobcov. Súčasný a budúci vývoj v oblasti automobilov sa pokúša zbližovať protichodnosť ekonomickosti a dynamiky jazdy. Údaje o predaji súčasných vozidiel s turbo-dieselovými motormi alebo preplňovanými zážihovými motormi, ktoré sú navrhované so zreteľom na znižovanie rozmerov (malé motory s vysokým špecifickým výkonom) naznačujú, že vývojové trendy sa ubierajú týmto smerom. V oblasti podvozkov je možné zadať orientáciu na aktívne tlmenie a pruženie ako aj individuálnu distribúciu momentov.

II. EXPANZIA ELEKTRONICKÝCH SYSTÉMOV V AUTOMOBILOCH

Dynamiku nárastu zastúpenia elektroniky v automobiloch možno rozčleniť do niekoľkých časových období. Predchádzajúce obdobie sa vyznačovalo predovšetkým vývojom a zdokonaľovaním elektronických systémov pre riadenie motora, jazdnej dynamiky vozidla (ABS, ESP, ...) a systémov pasívnej a aktívnej bezpečnosti. K dominantným témam, ktoré sú riešené v súčasnosti a predstavujú určitý revolučný skok v architektúre elektroniky vozidiel, patria systémy X- by-wire. Rezonujúcou ideou budúceho smerovania je pojem *infotainment*, ktorý v sebe skrýva spojenie termínov informácie, zábava a komunikácia.

Analytici odhadujú, že viac než 80% všetkých inovácií v automobiloch sa opiera o inovácie v elektronike. Jedným z hlavných vývojových trendov automobilových elektronických systémov sú vložené elektronické systémy (embedded systems) [2, 3].

Vývojové trendy v oblasti riadiacich jednotiek (ECU - Electronic Control Unit) charakterizuje bloková schéma na obr.1.



Obr. 1 Nárast dôležitosti softvéru vo vozidle

Z hľadiska celkového systému sa pri jeho vývoji musí brať zreteľ na tieto požiadavky: štandardizácia prepojení prevádzkových funkcií, spoločné a viacnásobne využitie existujúcich komponentov, komplexná bezpečnosť, vysoká úroveň komfortu, nízka spotreba energie a minimalizácia znečisťujúcich emisií.

Jedným z východísk vývoja systémov je snaha integrovať a viacnásobne využívať hlavné časti softvéru a hardvéru od viacerých dodávateľov, pričom nové prevádzkové funkcie by mali byť realizované prostredníctvom interakcie pôvodne autonómnych jednotiek (rekonfigurovateľné distribuované systémy/mechatronika). Nové riešenia by mali vyhovovať z hľadiska modularnosti a rozširiteľnosti pre rozličné vozidlá a varianty platforiem. Nástrojmi pri vývoji sa stávajú štandardizované vývojové prostredia, akými sú CMMI (Capability Maturity Model® Integration) alebo SPICE (Software Process Improvement and Capability dEtermination) [1], ktoré pomáhajú zvyšovať kvalitu, spoľahlivosť a integrovať tradične oddelené organizačné

funkcie. Nové systémy musia byť vybavené prostriedkami pre digitálne prenosy, sieťovú prenositeľnosť funkcií a musia poskytovať možnosť údržby počas životnosti produktu. V typickom systémovom reťazci, ktorý je tvorený senzorom, prenosovou cestou a riadiacou jednotkou, je možno pozorovať posun realizácie jednotlivých operácií spracovania signálov od riadiacej jednotky smerom k senzoru. Ak v konvenčnom reťazci bol výstupom zo senzora analógový signál, potom prenosová cesta mala analógový charakter a prenášaný signál bol citlivý na rušenia. Úprava signálu, A/Č prevod a ďalšie operácie sa pri takomto konvenčnom prístupe realizovali v riadiacej jednotke. Nárastom úrovne integrácie sa do senzorov prenáša funkcia predspracovania signálu, jeho A/Č prevod a senzor je vybavovaný mikropočítačom, ktorý tiež zabezpečuje pripojenie na prenosovú cestu. Túto reprezentuje zbernica (najčastejšie so sériovými protokolmi) a informácia, ktorá sa ňou prenáša má číslíkový charakter, čím môže byť viac odolná voči rušeniu. Do riadiacej jednotky sa tak dostávajú už informácie v číslíkovom tvare, kde sa realizujú definované algoritmy.

V novodobých vozidlách sa očakáva nárast podielu a významu softvéru [2]., napr. kým v roku 2000 tvoril podiel softvéru a hardvéru 20:80, v roku 2010 vzrástol tento podiel softvéru na 38:62.

Hlavnými východiskami pri vývoji softvéru sú:

- operačný systém reálneho času,
- komponentovo-orientovaný softvér,
- aktualizácie – updates a zvyšovanie úrovni – upgrades softvéru počas životnosti vozidla,
- minimalizácia nákladov a doby výpočtu komponentov softvéru a
- jednotný formát dát a priama prepojitelnosť komponentov softvéru.

Nosnými sa stávajú štandardizované systémy, ktoré sú otvorenými systémami (AUTOSAR - Automotive Open System Architecture) [4]. Tie majú umožňovať manažment komplexnosti elektronických systémov automobilov, spojený s nárastom funkčných možností. Majú napomáhať flexibilitu pri modifikácii produktu, zvyšovaní úrovne a aktualizácii, taktiež zlepšovať modularnosť riešení nielen v rámci výrobných sérií, ale aj medzi výrobnými sériami a zvyšovať kvalitu a spoľahlivosť elektronických systémov automobilov.

III. ROZMACH SIETÍ VO VOZIDLÁCH

Siete vo vozidlách prepájajú elektronické zariadenia vo vozidle, čím umožňujú využívať spoločné informácie a prostriedky medzi distribuovanými aplikáciami. Tieto riadiace a komunikačné siete sú založené na sériových protokoloch, čím nahrádzajú množstvo konektorov a vodičových prepojení, a tak menia prepojenia typu bod-bod centralizovaných riadiacich jednotiek na distribuované riadiace jednotky.

Hlavnými cieľmi sietí vo vozidle sú jednoduchosť použitia, redukcia nákladov, zvyšovanie kvality a z dôvodu zvýšenia diverzifikácie aj otvorenosť štandardov. Prínosmi sietí vo vozidle sú vyššia spoľahlivosť, viac prevádzkových funkcií za nižšiu cenu, štandardizácia prepojení a komponentov, rýchlejšie zavádzanie nových technológií a rozšíriteľnosť prevádzkových funkcií. Okrem iného ich zavádzanie znižuje hmotnosť vodičových prepojení (káblkových zväzkov) a ich zložitnosť, pričom sa tiež zmenšujú elektronické riadiace jednotky a sú aplikované priamo na senzoroch a aktuátoroch. V nasledujúcej tabuľke č. 1 sú uvedené moderné a najčastejšie používané zbernice s ich najdôležitejšími parametrami a ich pôvodným nasadením.

Automobilová AdHoc sieť VANET (Vehicular AdHoc Network) je technológia, ktorá využíva pohybujúce sa autá ako sieťové uzly pre vytvorenie mobilnej siete [5], [6], [7]. VANET premení každý zúčastnený automobil na sieťový smerovač alebo uzol, ktorý umožní autám na vzdialenosť desiatok až stoviek metrov pripojiť sa a následne vytvoriť sieť so širokým rozsahom. Tieto aplikácie predpokladajú vo vozidlách vysokú mieru vybavenosti komunikačnými jednotkami. Štandardy služieb a protokolov musia byť striktné rešpektované všetkými výrobcami palubných jednotiek, ako aj jednotiek infraštruktúry. Aplikácie budú využívať medzivozidlovú komunikáciu V2V (vehicular-to-vehicular) ako aj komunikáciu medzi vozidlami a jednotkami dopravnej infraštruktúry V2I (vehicular-to-infrastructure). Ich prvé nasadenie sa predpokladá v záchranných a bezpečnostných zložkách. Komunikačný systém bude môcť využívať základné typy bezdrôtových komunikačných technológií: *konvenčné bezdrôtové technológie na báze štandardov IEEE 802.11 a/b/g/n, komunikačné technológie na báze štandardu IEEE 802.11p, 802.16, bunkové rádiové komunikačné technológie GSM, GPRS, UMTS, HSPA+, LTE a pod. (viď. Tab.1).*

Tab. 1
 ZÁKLADNÉ TYPY BEZDRÔTOVÝCH KOMUNIKAČNÝCH TECHNOLOGIÍ

		Prenosová rýchlosť
Zbernice	D2B (elektrická alebo optická hlavne pre digitalne audio)	5 Mbit/s
	MOST (audio, video, riadenie)	22.5 Mbit/s
	FlexRay (x-by-wire, riadenie kritické z hľadiska bezpečnosť)	10 Mbit/s
	Byteflight (konštantné oneskorenie, airbag, bezp. pásy)	10 Mbit/s
	TTP (distribúované/poruchám odolné aplikácie reálneho času)	5 - 25 Mbit/s
	CAN (riadenie)	50 – 1000 kbit/s
	J1850 (riadenie)	10,4 kbit/s a 41,6 kbit/s
	LIN (riadenie)	20 kbit/s,
Bezdrôtové siete	Bluetooth (prepojenie zariadení infotainmentu)	10 Mbits/s
	WiFi (WLAN) IEEE 802.11 a/b/g (5 / 2,4 / 2,4 GHz)	54 Mbit/s a 11 Mbit/s
	IEEE 802.11n (2,4 a 5 GHz)	600 Mbit/s
	IEEE 802.11p (5,85-5,925 GHz)	3; 4,5; 6; 9; 12; 18; 24; 27 Mbit/s.
	WiMAX IEEE 802.16 (10 - 66 GHz)	134 Mbit/s
	GPRS	max. 21,4 kbit/s
	UMTS (s HSPA+)	45 Mbit/s
	LTE (1,4 MHz - 20 MHz)	100 Mbit/s

IV. ZÁVER

V príspevku boli predstavené niektoré úvahy o budúcnosti elektronických zariadení, ich infraštruktúry a ich vplyvy na vývoj a výrobu automobilov vo svete.

V posledných 20- tich rokoch bolo možné zaznamenať dramatický nárast objemu a zložitosti elektronických systémov v automobiloch, ktoré priamo súviseli s:

- vysokým počtom riadiacich jednotiek a ich sieťovým prepojením v rámci vozidla,
- veľkým počtom prevádzkových funkcií a nárastom zložitosti softvéru a
- nárokmi na kapacitu pamäti a výpočtový výkon.

Pre niektorých menších výrobcov sa stávajú náklady na vývoj softvéru vyššími, než náklady na elektronický hardvér. Bude to taktiež náročná úloha pre veľkovýrobcov, aby uspokojili požiadavky budúcich zákazníkov, a pritom zachovali náklady na vývoj (hardvér aj softvér) na primeranej úrovni. Napriek tomu budú požiadavky na vývoj elektronických systémov celosvetovo narastať. Vývoj elektronických systémov, ktorý zohľadňuje viacero aspektov, otvára veľký potenciál pre pomoc výrobcov pri vytváraní jedinečných produktov pre svojich zákazníkom.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku / Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

LITERATÚRA

- [1] S. Rathmann – R. Fischerkeller, *Latest Trends in Automotive Electronic Systems – Highway meets Off-Highway?* Agricultural Engineering International: the CIGR Ejournal, vol. IX, July, 2007.
- [2] B. Hardung – T. Kölzow – A. Krüger, *Reuse of Software in Distributed Embedded Automotive Systems*. Proceedings of the 4th ACM international conference on Embedded software, ACM New York 2004, pp. 203 – 210.
- [3] A. Sangiovanni-Vincentelli, *Automotive Electronics: Trends and Challenges*. Convergence 2000, Detroit, October 2000.
- [4] AUTOSAR. *Automotive Open System Architecture*. <http://www.autosar.de>, Apr. 2004.
- [5] T. KOSCH, *Technical Concept and Prerequisites of Car-to-Car Communication*. Germany, Munich, 2006.
- [6] R. H. MENOVAR – M. LENARDI, *Reseaux VANETs et norme 802.11p*, Paris, 2008.
- [7] Bo Yu – Chengzhong Xu, *Vehicular Ad-Hoc Networks: An Information-Centric Perspective*. http://www.wen.zte.com.cn/endata/magazine/ztecommunications/2010Year/no3/articles/201009/t20100913_191839.html

OPNET Modeler: Testing of the New Integration Model of QoS and Security for MANET

Ján PAPAĽ, Ľubomír DOBOŠ, Anton ČIŽMÁR

Dept. of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

jan.papaj@tuke.sk, lubomir.dobos@tuke.sk, anton.cizmar@tuke.sk

Abstract — The OPNET Modeler is a simulation tool that enables modeling different type of large-scale networks with in detail defined parameters. In mobile ad hoc network (MANET), the security and QoS are essentials to satisfy the communication constraints and requirements. Current trends in the field of providing new services require the providing of services with QoS and security together. In this article, the simulation studies of the new designed model, which is used to integration of QoS and security as a one parameter in MANET, is discussed. Main parts of our model are modified security service vector and cross layer design model. These two elements are used to interaction between user and system and also to provide interaction between system's elements. The new model enables the cooperation between QoS and security mechanisms. The OPNET modeler simulator is used to simulation of the new designed model.

Keywords — OPNET, MANET, QoS, Security

I. INTRODUCTION

OPNET Modeler is very popular commercial network simulation and analytical tool, which is used for network design, modeling and simulation [1]. The users have the possibilities to design and study many type of communication networks, devices, protocols and applications with flexibility and scalability. OPNET also provide very useful tools for simulations of the MANET networks (Fig. 1). OPNET modeler provides a graphical editor interface to build models for various network entities from physical layer modulator to application processes [1]. All the protocols and algorithms are modeled in an object-oriented approach. Process models and algorithms are specified in the PROTO-C language which uses a graphical editor to design the structure of the process in the form of a finite state machine (FSM). The FSM contains the logic of the process model within its states and transitions. Process models are using a library of many kernel procedures which enables access to packets, network variables, statistic collection, packet communication and other simulation service [1].

OPNET Modeler consists of the group of the hierarchic editors [1]. The network editor is used to design and specify networks models, which consist of subnetwork and node objects. The Node editor is used to define the behavior of each network object. Behavior is defined using different modules. The process editor provides the process models, which control the underlying functionality of the node models created in the Node Editor. There are another types of editor used to specify specific use in OPNET, namely: Packet editor, link editor, the modulation curve editor, the demand editor, the path editor, the antenna pattern editor, the ICI editor, the Probability Density Function editor, probe editor etc.

A MANETs are characterized as a dynamic multi-hop wireless networks that consists of mobile nodes that communicate with each other by cooperatively sharing a common wireless medium. The terms of Quality of Service (QoS) and security provides important areas of the research in MANET. There are many definitions of QoS, but we can say that QoS is defined as a set of service requirements that need to be met by the network while transporting a packet stream from source to destination. Research in the field of QoS includes following areas: QoS models, QoS resource reservation signaling, QoS routing and QoS Medium Access Control (MAC) [2]. Security also provides very important part of research. In this field, research community is oriented to following research sectors: Secure routing, Key management and

Intrusion Detection System [3]. The main goals of security requirements are following: Confidentiality, Authentication, Availability, Integrity and Non-repudiation.

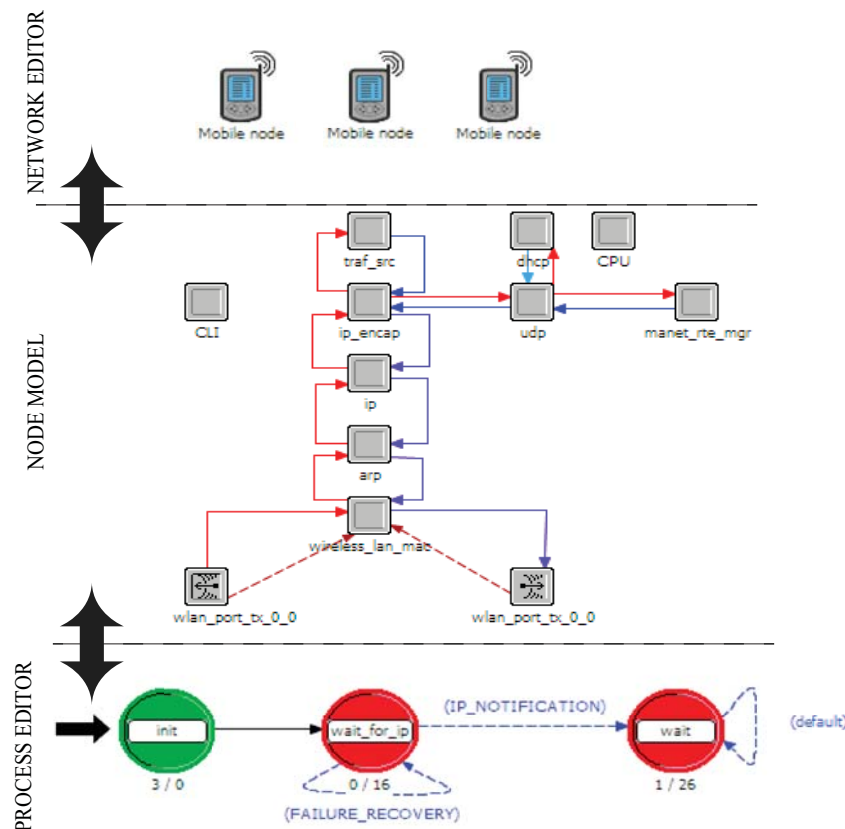


Fig. 1 OPNET modeler hierarchy

The integration of the QoS and security as a one parameter provides very important and new dynamic part of research in MANET. Today, the QoS schemes, for example QoS routing, QoS models, QoS signalizations, don't solve problem how can be integrated with security mechanisms to the one integrated system. Moreover, the research in the field of security provides mechanism to eliminate different type of security attack and do not solve problem of integration.

The new service, that can MANET provides, requires the providing security and QoS algorithms together. In this case, the integration enables proper functioning of both mechanisms in terms of QoS and security. In this article the new integration model is introduced. We design the new cross layer interface which is used to cooperation between application and network layer and is used to collect of relevant information. The modified security service vector (SSV) is used for cooperation between several blocks of the new model. This model is implemented to the OPNET modeler simulation environment. The performance analysis is also introduced.

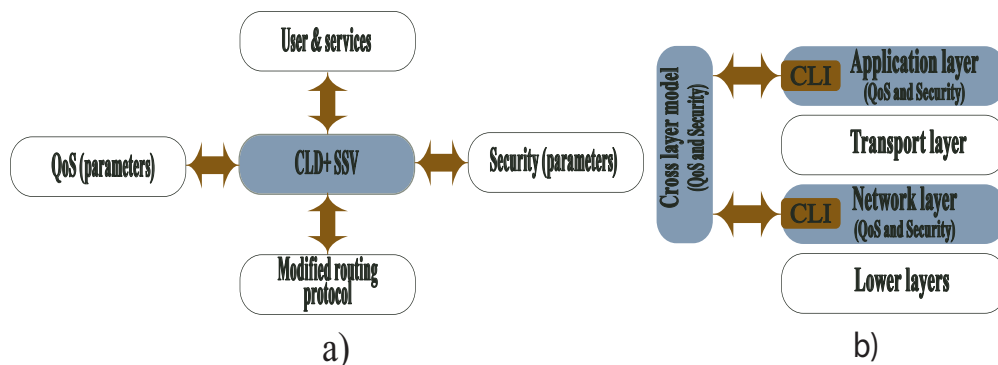


Fig. 2 Integration of the security and QoS in MANET: a) New model to integration of QoS and security in MANET, b) new cross layer model (CLD) in MANET

II. NEW INTEGRATION MODEL USED TO INTEGRATION OF SECURITY AND QoS IN MANET

A. New integration and cross layer model in MANET

We design the new integration model, which allows integrating of the security and QoS as a one parameter via modified SSV and CLD (Cross layer model) in MANET (Fig. 2a). Model includes all components for the interaction between user and system to integrate security as a one parameter in MANET. The main block of our model is block *CLD+SSV*. The CLD (Cross layer design model) is used to create interactive environment between users and system and is used to support interactions between the routing protocol and modified security service vector (SSV) [4]. The SSV presents modified security service vector used to collecting information about QoS and security. Block *QoS (parameters)* represents a mechanism for delivering of QoS in MANET. It defines and specifies the QoS parameters necessary to provide the required services or provide information about what type of service can node provide. Block *Security (parameters)* represents a mechanism to provide security-related services and also defines the necessary parameters used to process services providing. Block *User&Service* enables the interaction between user and system. The interaction with user means, that user can define parameters for the type of service, which has to be achieved for services. Block *Modified routing protocol* represents the routing protocol with implemented modified SSV algorithm to selection the optimal way based on user defined requirements (QoS and Security).

The new cross layer model (CLD) or cross layer interface (CLI) between non-adjacent layers was created for the purpose of collecting QoS and security related data that are necessary for processing of the modified SSV and modified routing protocol [5]. Modified SSV was implemented to the dynamic source routing protocol (DSR). The task of our cross layer design is to enable transferring and collecting data from the application layer to the routing protocol operating on network layer. The collected SSV attributes consist of information about security and QoS parameter that a node is able to provide [6]. The basic concept is depicted in Fig. 2b. Process of collecting QoS and security related information is the same for all 3 types of nodes. In the case of source node, the user defines QoS and Security via CLI interface located on application layer (Fig. 2b). Collected data are marked as QoS_1 and $Security_1$, and are stored to modified route cache and to the RP (position rSSV) as well (Fig. 3).

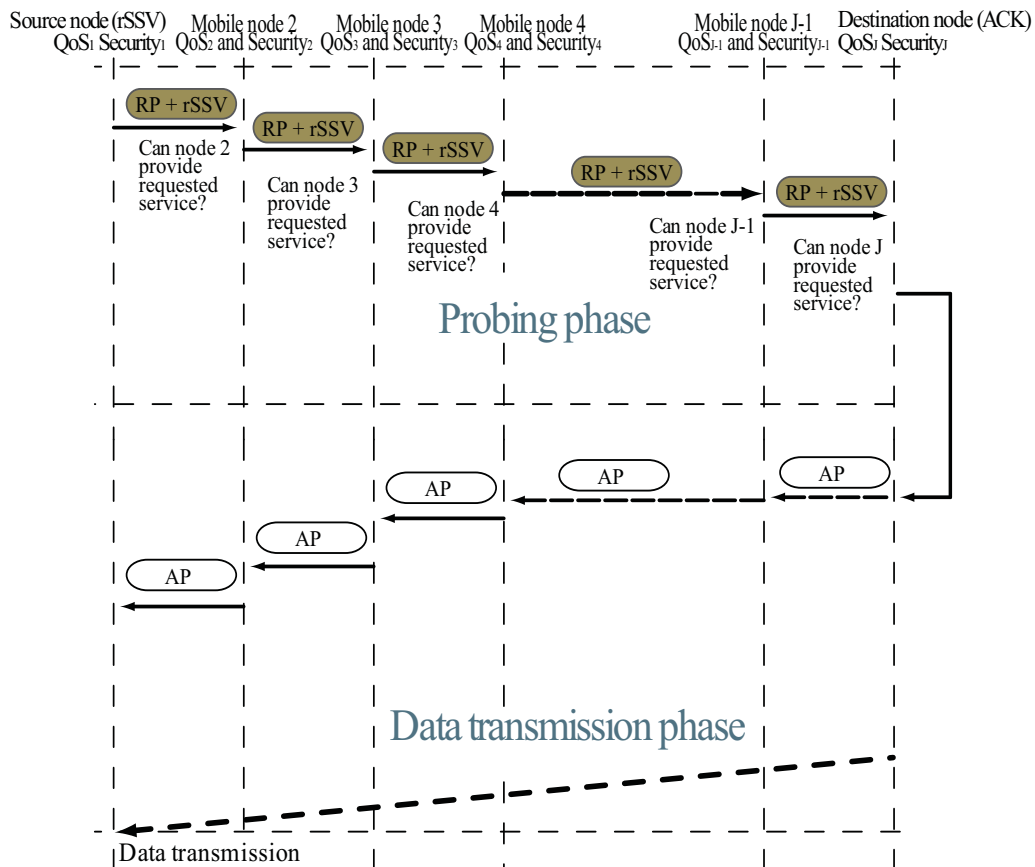


Fig. 3 Basic idea of modified SSV in MANET [4]

B. Implementation of the CLD model to the OPNET modeler simulator

The following process model architecture is designed to integration of the CLD model to the MANET layer model in OPNET modeler. The cross layer interface process model (CLI), shown in Fig. 4, enables the bi-directional interactions and communications between application and network layer in MANET layer model. This new CLI model enables interaction with user and enables to collect relevant data about QoS and security via modified SSV. The CLI is described as finite state machine (FSM). The CLI process model consists of four states. State INIT consists of the initialization of the process model.

This state initializes every variable, statistic, table, and user parameter that is used by the CLI process model (Fig. 4b). State *IDLE* is used to analyzing the type of requests and this is the default state where the process waits for an event. State *From_APPL* is activated when *CLI_REQUEST_APPL* request from application layer is obtained. This request is used in the case of source node, when user specifies the QoS and security requirements. State *From_IP* indicates, that network layer activate request for collecting information about QoS and security from the application layer on the node.

III. SIMULATIONS AND RESULTS

The main goals of the simulations were to verify the possibilities of implementing a new designed model to the MANET modeler. All behavior of the proposed model was simulated in OPNET simulator. Three types of simulations were used to evaluate the effectiveness of integrating a new model with CLD and modified SSV:

- *Model where the nodes used routing protocol DSR without modified SSV and CLD (DSR)* – data are transmitted by each layer without CLD and modified SSV.
- *Model where the nodes used modified routing protocol with implemented the modified SSV (DSR+SSV)* - data are transmitted by each layer without CLD with implemented modified SSV.
- *Model where the nodes used modified routing protocol with implemented modified SSV and CLD (DSR+SSV_CLD)* - data are transmitted by new CLD interface and modified SSV is implemented.

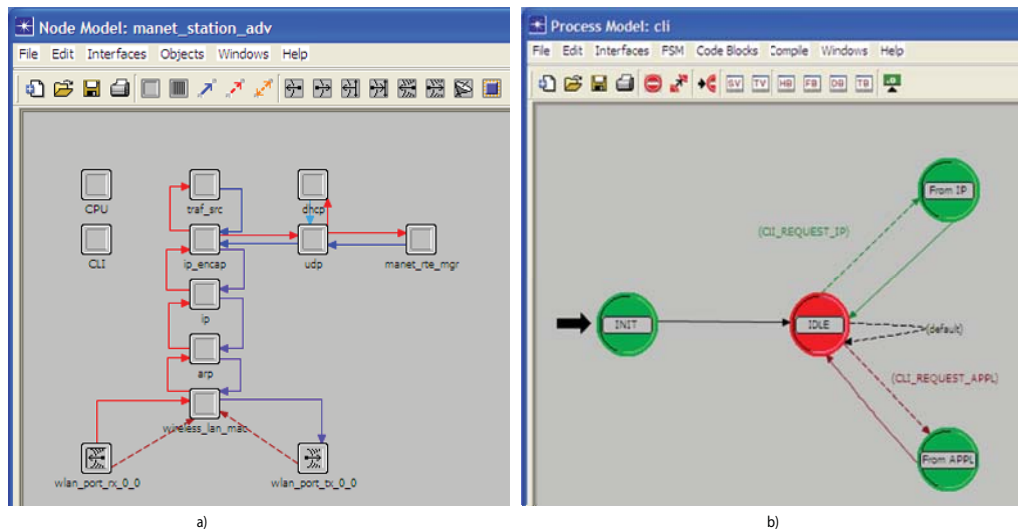


Fig. 4 Implementation of integration model to OPNET: a) Layered model of MANET terminal with implanted CLD, b) new designed process model of CLD

The 5 separated simulation scenarios that were formed of 20, 40, 60, 80 and 100 nodes were created to check the effectiveness of the modified SSV and CLD in MANET. Transmit power was setup to 1mW. The random mobility model was used to simulate the mobility of the nodes. Speed range ranged from 0 to 2m.s⁻¹. Simulation period was in all cases 1000 seconds. Free environment without affecting interference was used as a simulation environment. Changing parameter was the initial value of movement, which gives a different initial position of individual nodes in the simulated project. The result of each simulation was a set of values, which were then statistically processed and evaluated. The number of values can be chosen in the simulator environment and in our case, each sample was made up of a set of 100 values from each simulation (10 000 values were recorded). The averages values are shown in all

graphs. We investigated the effect of the increasing of the traffic (burden) and applying the new designed model with modified SSV and CLD on delay of MANET and total packet processing delay in simulations. The burden is seen as the number of the nodes generating traffic (packets) in this case. In each simulation is randomly generated set of nodes (20%, 40%, 60%, 80%) that generates the packet traffic.

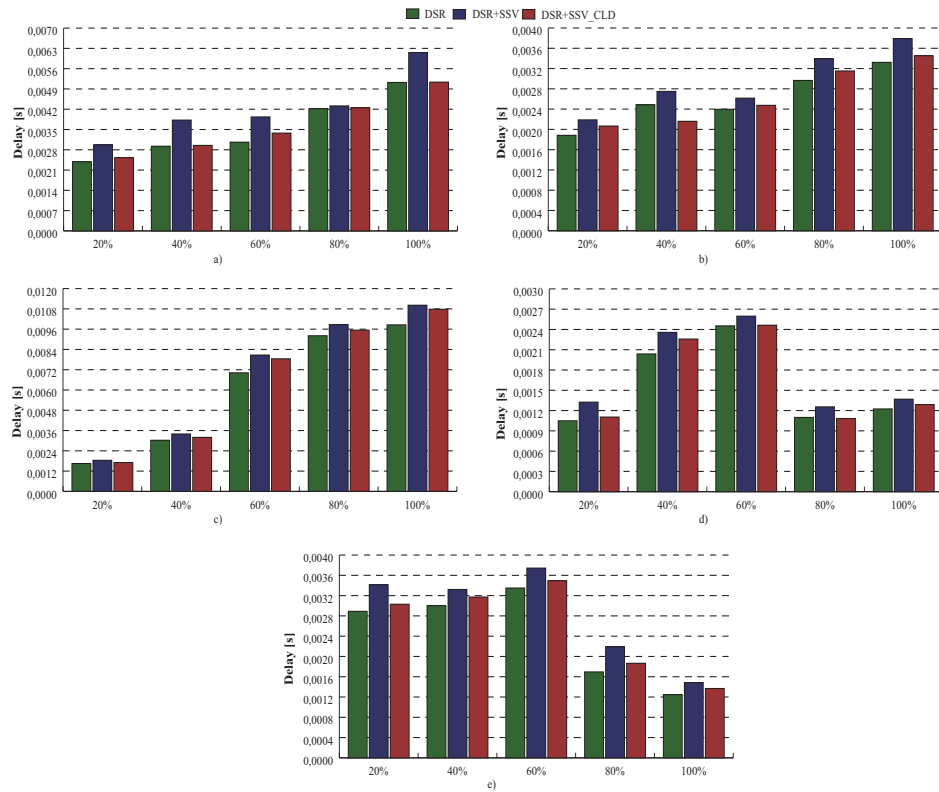


Fig. 5 Delay of MANET depending on the number of traffic generates nodes:
 a) 20%, b) 40%, c) 60%, d) 80% and e) 100%

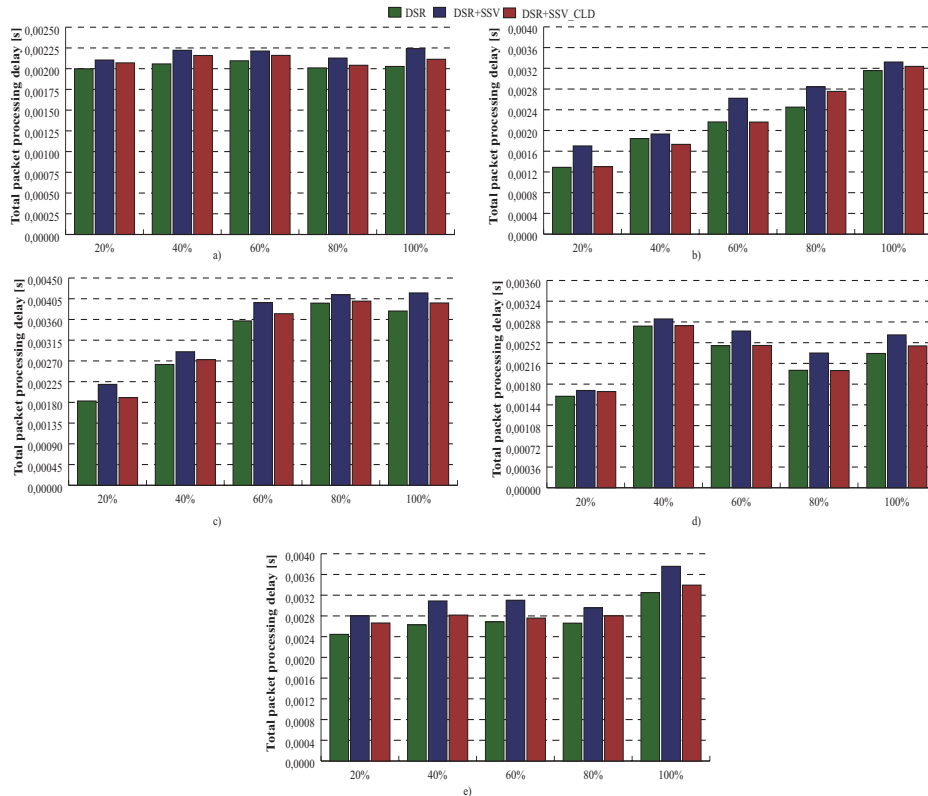


Fig. 6 Total packet processing delay of MANET depending on the number of traffic generates nodes:
 a) 20%, b) 40%, c) 60%, d) 80% and e) 100%

Fig. 5 shows the results of delay of MANET for varying number of the nodes (from 20 to 100) that generates the traffic. Fig. 8 also shows the delay of MANET comparison between the DSR, DSR+SSV and DSR+SSV_CLD. We can see that implementation of the DSR+SSV_CLD to the MANET layer model represents the reduction of the delay in compares with DSR+SSV model, but on the other side provides the increasing of the values of delay in compare with DSR.

Fig. 6 shows total packet processing delay versus the increasing of the traffic in MANET. In all cases, the model DSR+SSV increases of the total packet processing delay values cooperated to DSR model. We can see the implementation of the DSR+SSV_CLD provides reductions of the values in compare with DSR+SSV, and provides minimally increasing of these values in compared with DSR. These situations, in all cases, are caused with following factors:

- *number of nodes on simulation area,*
- *speed of the mobile terminals,*
- *number of operations necessary for SSV and CLD on the nodes.*

IV. CONCLUSION

The OPNET modeler provides well know simulation tool for simulation of mobile ad hoc networks. OPNET modeler contains a large set of predefined simulation attributes, but it also makes it possible to build user-defined network elements with proprietary functions.

In this article the new model used to integration of the QoS and security as a one parameter was introduced and tested in OPNET modeler simulation environment. This new integration model provides a new way how could be provided QoS and security related services in parallel. The conclusions reached on the basis of these results are: if the new model with implemented the new modified SSV and CLD (DSR+SSV_CLD) was used, the values of delay and total packet processing delay of MANET increased minimally compared with standard DSR protocol. All simulations are made in OPNET modeler.

ACKNOWLEDGMENT

Research described in the paper was financially supported by INDECT (FP7-No.218086) (30%). This work is also the result of the project implementation Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120020) supported by the Research & Development Operational Program funded by the ERDF (70%).

REFERENCES

- [1] OPNET Modeler Simulation Software, <http://www.opnet.com>
- [2] A. Roche, C. B. Westphall, PACK-So Graf von Mecklenburg, "Quality of Service for Ad-hoc Wireless Network", Computer Science Society, SCCC 2002. Proceedings, 22nd International Conference of the Chilean, pp.100-105, 2002.
- [3] D. Djenouri, L. Khelladi, A.N. Badache, "A Survey of Security Issues in Mobile Ad Hoc and Sensor Networks". In Communications Surveys & Tutorials, IEEE, volume 7, issue 4, pp. 2 - 28, 2005.
- [4] J. Papaj, L. Doboš, A. Čižmár, "Security service vector in MANET" In AEI'2008 International Conference on Applied Electrical Engineering and Informatics : September 8-11, Athens, Greece, 2008. Košice : FEI TU, pp. 130 -138, ISBN 978-80-553-0066-5, 2008.
- [5] J. Papaj, L. Doboš, A. Čižmár, "New cross layer model to integration QoS and security as a one parameter in mobile ad hoc network", In MCSS 2010: Multimedia Communications, Services and Security, IEEE international conference, Krakow, pp.1-6., ISBN 978-83-88309-92-2, 6-7 May 2010.
- [6] J. Papaj, L. Doboš, A. Čižmár, "Using Cross layer design to collecting of SSV data in MANET", In AEI '2009 : International Conference on Applied Electrical Engineering and Informatics: September 7-11, Genova, Italy, 2009. Košice : FEI TU, pp. 66 - 73, ISBN 978-80-553-0280-5, 2009.

Prediktívne prepínanie videotokov v internetovej televízii

*Branislav HRUŠOVSKÝ, Pavol KOCAN,
Stanislav MARCHEVSKÝ, Martin LIPTAJ*

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

branislav.hrusovsky@tuke.sk, pavel.kocan@gmail.com
stanislav.marchevsky@tuke.sk, martin.liptaj@tuke.sk

Abstrakt — Tento článok sa zaoberá prediktívnym prepínaním videotokov v internetovej televízii. Analyzuje súčasný stav používania internetovej televízie na širokopásmových linkách. Bližšie sa zameriava na problémy vznikajúce pri prepínaní video tokov a s tým vznikajúce oneskorenie prenosu vysielania. Cieľom článku je navrhnúť spôsob zjednodušenia prepínania videotokov použitím štatistického spracovania divákom sledovaných staníc a následného použitia predikcie voľby stanice, ktorá by mohla byť zvolená prepnutím užívateľa. Pomocou počítačových simulácií skúma možnosť použitia náhľadových videotokov pre potrebu jednoduchšieho prepínania staníc užívateľom. Vzhľadom na aktuálnosť a zložitosť tejto problematiky ponúka práca v závere návrh ďalších krokov pre výskum v tejto oblasti.

Kľúčové slová — internetová televízia, prepínanie paketov, riadenie zahltenia, algoritmus

I. ÚVOD

S príchodom stále viac dostupného širokopásmového pripojenia do celosvetovej siete internetu si multimediálna komunikácia pomocou internetového protokolu aj v našej krajine nachádza stále viac priaznivcov. Po rokoch telefonovania pomocou pevnej linky či mobilného prístroja začíname objavovať aplikácie ako video hovory, lacné hlasové hovory s použitím počítača a v poslednom čase prináša vplyv digitalizácie aj postupný prechod na digitálne terestriálne vysielanie, či televízne vysielanie s použitím internetového protokolu [3].

K masívnemu rozšíreniu takzvanej IP televízie prispelo aj neustále zrýchľovanie prístupu do celosvetovej siete internetu, umožňujúce ponúkať stále náročnejšie a kvalitnejšie služby. Okrem celkového zrýchľovania sietí sú to aj efektívnejšie kompresné algoritmy používané v súčasných kódovacích a dekodovacích programových prostriedkoch, ktoré výrazne prispeli k rozšíreniu služby prenosu zvukových a obrazových tokov v podobe počúvania internetového rádia alebo sledovania televízie. Aj napriek pokroku vo vývoji takýchto sietí, častokrát dochádza pri prenose k ich zahlteniu, a to najmä v ich najslabších miestach s najmenšou šírkou pásma. Aj preto sa stále viac kladie dôraz na výskum algoritmov, ktoré dokážu predchádzať takýmto stavom, v tom lepšom prípade dokážu prispôsobiť prenášaný dátový tok aktuálnym podmienkam prenosovej siete.

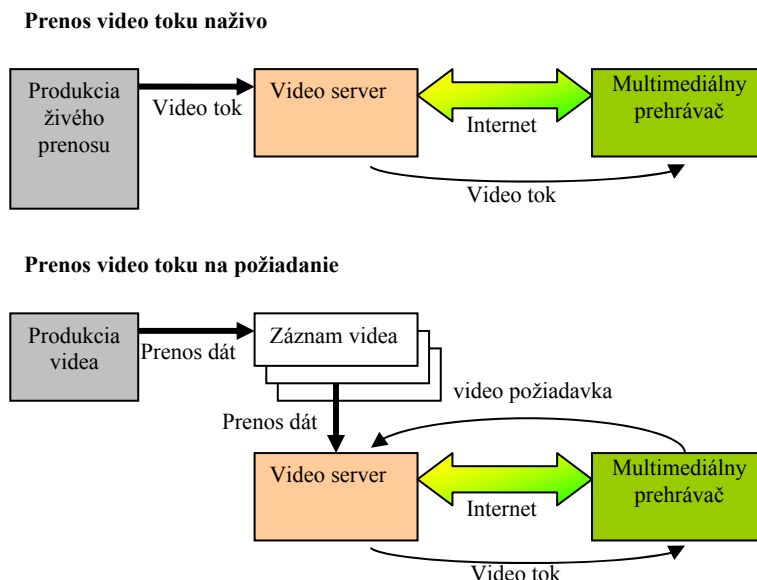
II. TELEVÍZIA CEZ INTERNET

A. Prenos toku dát a jeho spracovanie

Pri spracovávaní tokov videa sa stretávame s viacerými základnými problémami. Používanie tejto služby v prostredí internetu je problematické, pretože postupné sťahovanie ponúka len službu tzv. “best effort“, čo znamená, že sa snaží podľa možností zabezpečiť čo najlepšiu kvalitu. Negarantuje šírku prenosového pásma, oneskorenie “jitter-a“ alebo stratu rýchlosti, nakoľko tieto parametre sú vopred neznáme a dynamicky sa menia [5]. Práve tieto parametre sú pri postupnom sťahovaní cieľom pre návrh systému spoľahlivého doručovania vysoko kvalitného videa cez sieť internetu.

Prenosová šírka pásma medzi dvoma bodmi internetu je v podstate neznáma a neustále sa mení. Ak začne odosielateľ posilať rýchlejšie, než akú má dostupnú šírku pásma, potom dochádza k zahlteniu, pakety sa začínajú strácať a pre video to znamená viacero výpadkov a

zhoršenie kvality. Ak sa však posiela menšie množstvo dát, než aké je sieť schopná preniesť bez straty, prijímač je schopný vytvoriť subjektívne optimálnu kvalitu videa [1]. Z tohto dôvodu je vhodné predchádzať problémom s nedostatkom kapacity prenosového kanála odhadovaním jeho šírky pásma a následným prispôbením toku dát, ktorý sa bude pomocou neho posilať. Do úvahy treba však vziať aj ďalšie aspekty odvíjajúce sa od nestabilnej šírky pásma prenosového kanála, pôsobiace najmä na iné súbežné toky internetu pri riešení problému skupinového vysielania, pri ktorom jeden zdroj vysiela dátový tok viacerým klientom s rôznymi prístupovými rýchlosťami. Na Obr.1 je znázornený prenos videotoku naživo a videotoku na požiadanie.



Obr.1 Porovnanie prenosu videotoku naživo a videotoku na požiadanie

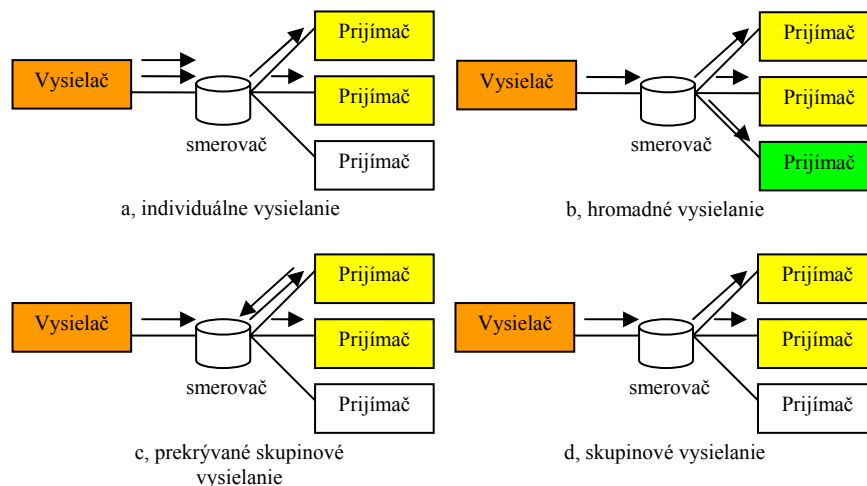
B. Skupinové vysielanie

Štandardná komunikácia v sieti internetu pri používaní služieb je formou individuálneho vysielania (z angl. pojmu unicast), t.j. každý jeden používateľ je obsluhovaný serverom samostatne, zo strany poskytovateľa služby. V prípade strímingu bežného televízneho vysielania, by to predstavovalo príliš veľké zaťaženie nielen samotnej prenosovej siete, ale aj serverov poskytujúcich posielanie toku dát [2]. Z tohto dôvodu využíva IP televízia prenos skupinovým vysielaním, kde môže byť jeden prenášaný dátový tok obrazu a zvuku simultánne spracovávaný viacerými užívateľmi, pričom na strane vysielača je postačujúci výkon porovnateľný s jedným užívateľom. Táto úspora umožňuje v reálnom čase spracovávať a šíriť niekoľko paralelných tokov s použitím jedného servera pre veľké množstvo užívateľov [6].

Na riadenie zahľtenia vplývajú aj zložitejšie štruktúry sieťových prenosov medzi vysielačim a prijímacím bodom. Konkrétne prípady sú znázornené na Obr.2. Základnou myšlienkou týchto modelových situácií je jeden zdrojový tok dát, ktorý je prijímaný viacerými koncovými účastníkmi, napríklad prenos zvukového toku rádio stanice. Všeobecne môžeme povedať, že tieto príklady predstavujú najbežnejšiu prevádzku pre multimediálnu komunikáciu v reálnom čase. Odlišnosť prenosu skupinového vysielania a súčasne jeho efektívnosť oproti individuálnym užívateľským sú znázornené na Obr.2. V prípade (a) je zhodný tok dát, dvakrát prenášaný pomocou prvých dvoch prepojení, čím sa mrhá prenosovou kapacitou, ba zvyšuje sa aj riziko vzniku zahľtenia [3]. Riešením problému je prenos skupinovým vysielaním, pri ktorom sa tok prenáša až k druhému smerovaču, ktorý následne rozpošle tok koncovým užívateľom. Tento spôsob prenosu vlastne vytvára stromovú štruktúru oproti typickému spojeniu dvoch bodov (end-to-end) medzi vysielačom a jednotlivými prijímačmi.

Medzi ďalší spôsob komunikácie patrí prenos hromadného vysielania (broadcast), ktorý je porovnateľný s klasickým rádio rozhlasom, kde prijímame všetky možné stanice frekvenčného pásma, ale pomocou filtra ovládaného otočným ovládacím prvkom si navolíme len jednu stanicu, ktorú chceme práve počúvať. Na obrázku je znázornená neefektívnosť takéhoto prenosu, pretože sa zbytočne plytvá prenosovou kapacitou medzi druhým smerovačom a spodným užívateľom (zelený), ktorý sa nepodieľa na prijímaní daného toku. Prenos typu

prekrývaného skupinového vysielania (overlay multicast), prípad (c), je najpoužívanejšie, prechodné riešenie, kým nedôjde k celosvetovému rozšíreniu skupinového vysielania: niektoré koncové systémy plnia úlohu prepojavacích systémov a berú si na starosť úlohu pôvodne predurčenú smerovačom. Schéma znázorňuje menšiu efektívnosť voči skupinovému vysielaniu na sieťovej vrstve, tu sa plytvá šírkou pásma vrchného užívateľa smerom späť k smerovaču [4]. V tejto modelovej sieti je jasným víťazom skupinové vysielanie. Nie je ťažké si predstaviť, že negatívne efekty ostatných prenosových typov by boli v oveľa väčšej a zložitejšej sieti zreteľnejšie. Veľkou nevýhodou pri prenose skupinového vysielania je zložitá otázka riadenia.



Obr. 2 Typy vysielaní

C. Popis procesu prediktívneho prepínania

Všeobecne prevláda mylná predstava, že skupinové vysielanie v IP sieťach spôsobuje pomalé prepínanie kanálov. Je nesporné, že zastavenie príjmu kanálu skupinového vysielania a začatie nového nie je okamžitým procesom, toto oneskorenie je veľmi nízke vzhľadom na rozšírenie smerovačov, ktoré obsluhujú tisíce požiadaviek zároveň. Pretože existujú miestami aj tisíce dostupných kanálov, zo strany poskytovateľov je veľmi malá pravdepodobnosť, že by sa vyskytol enormný záujem o jeden konkrétny kanál. Čo však v prípade, keď hromadná väčšina účastníkov sleduje práve takýto žiadaný kanál a dôjde k situácii, kedy ukončia sledovanie a začne hromadné prepínanie na iný kanál? Predstava prediktívneho systému prepínania v reálnych podmienkach môže byť ovplyvnená mnohými faktormi, medzi ktoré určite patrí oneskorenie spôsobené prenosom alebo spracovaním dekódovaním prijatého video toku ako aj ďalšie nemenej dôležité faktory. Nasledujúca časť popisuje predstavu realizácie prediktívneho prepínania ako aj vývojový diagram algoritmu.

1) Sledovanie záujmu diváka

Predstavme si športového fanúšika, oddaného sledovaniu po väčšine času hlavne športových prenosov na športových stanicích, alebo diváka sledujúceho vo voľnom čase historické dokumenty, prípadne cestovateľa, ktorý túži neustále spoznávať nové kúty sveta. Toto je len hrstka modelových užívateľov, pre ktorých môže predstavený systém priniesť pohodlnejšie sledovanie televízneho vysielania. Systém prediktívneho prepínania staníc zakladá práve na myšlienke, že daný typ užívateľa bude zameraný výhradne na daný okruh relácií alebo dokumentov. Tento predpoklad ponúka poskytovateľom káblových rozvodov novú službu v oblasti starostlivosti o zákazníka. V ideálnom prípade je vhodné umiestniť monitorovacie zariadenia na strane poskytovateľa. Výsledok monitorovania je potrebné uchovávať pre potrebu neskoršieho použitia, pričom sa využívajú časové značky vysielaných programov. Na základe týchto značiek je možné vyhodnotiť celkový záujem diváka.

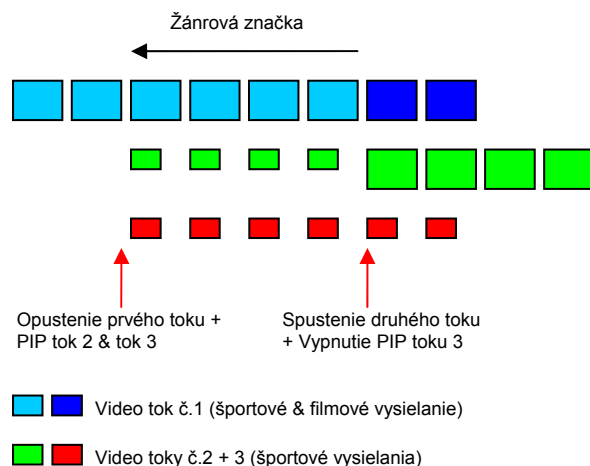
2) Predikovanie ukončenia relácie

Zákazník (účastník) sleduje športovo zameraný program na televíznej stanici (ďalej len video tok s poradovým číslom). Systém na základe informácie získanej z EPG alebo časového kódu poskytovateľa televízneho vysielania do káblového rozvodu, signalizuje nastávajúci koniec sledovaného programu, resp. signalizuje zmenu vysielaného žánru na taký, ktorý už nie je zaujímavý pre daného užívateľa/zákazníka. Toto sa deje v dostatočnom predstihu, takže systém

má dostatok času reagovať na túto zmenu žánru a teda ponúknuť alternatívu k práve končiacemu programu.

3) Hľadanie alternatívneho kanálu

Pri hľadaní alternatívneho kanálu sa pracuje na strane poskytovateľa s výsledkom monitorovania účastníka. Priorita navrhovaných alternatív je nastavená podľa prednastavených hodnôt prostredníctvom televízneho prijímača. Do úvahy sa tak môže brať žánrové zameranie účastníka, ale aj presnejšie zameranie, napríklad v prípade športovo založeného účastníka na konkrétny druh športu, prípadne pevné nastavenie ponúkaných náhľadov podľa vlastného výberu.

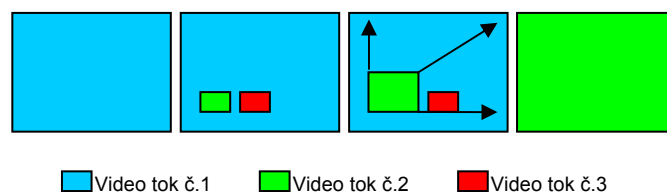


Obr. 3 Proces prepnutia videotokov vrátane náhľadov

Je preto dobré, ak sa samotná štatistika získaná z monitorovania priebežne spracúva a neustále aktualizuje, a tým sa prispôsobuje aj samotná služba aktuálnym potrebám užívateľa. V ďalšom kroku prichádza na rad pravdepodobnostný model užívateľa, ktorý vyhodnotí štatistické záznamy a určí spomedzi dostupných staníc najpravdepodobnejšiu voľbu, ktorú by vykonal samotný užívateľ, a teda ktorú stanicu by si vyvolil.

4) Alternatívne náhľady

Keďže predstavovaná služba funguje na báze poskytnutia viacerých náhľadov alternatívnych staníc formou obrazu v obraze (PIP – Picture in picture), je potrebné posielat' dodatočné videotoky smerom k užívateľovi, ale len v dostatočnej náhľadovej kvalite, čím sa výrazne šetrí na potrebnej šírke prenosovej kapacity siete [7]. Pravdepodobnostný model teda zvolí len niekoľko staníc/tokov, v našom prípade uvažujeme dodatočné dva až štyri toky, ktoré začne posielat' v dostatočnej náhľadovej kvalite. Takúto kvalitu je možné získať znížením rozlíšenia alebo zmenou počtu snímok. Takto vysielané a formou PIP zobrazené toky je možné na obrazovke sledovať niekoľko sekúnd pred ukončením sledovaného programu. Počas tejto chvíle je na výbere užívateľa, ktorú voľbu si zvolí.



Obr. 4 Proces na televíznej obrazovke

5) Prepnutie stanice

Po voľbe požadovanej stanice začína server poskytovateľa vysielat' už plnú kvalitu zvoleného videotoku, a zároveň sa preruší vysielanie prvého toku, ktorý bol doposiaľ sledovaný. Proces prepnutia videotokov vrátane náhľadov je znázornený na Obr. 3. Počas tohto prepnutia sa v priebehu krátkej doby zastavuje aj náhľad prípadných ďalších video tokov v PIP. Proces je teda ukončený hneď, ako sa začne v plnej kvalite na obrazovke sledovať prepnutý druhý videotok a ďalšie pomocné video toky sú už odstavené. Systém je možné v prípade potreby automatizovať a viac prispôbovať nárokom užívateľa. Celý proces prebiehajúci na obrazovke je načrtnutý na Obr. 4.

III. SIMULÁCIE

Nastavenie simulácií prenosovej siete v simulačnom programe NS-2 bolo nasledovné: prenos prebiehal po sieti tvorenej z piatich bodov, dvoch koncových a troch priebežných. Koncové body (n0, n4) predstavovali server na strane poskytovateľa a klienta. Body n1, n2 a n3 predstavovali smerovače. Medzi jednotlivými bodmi siete boli prednastavené nasledujúce parametre liniek (prepojenie, rýchlosť, oneskorenie):

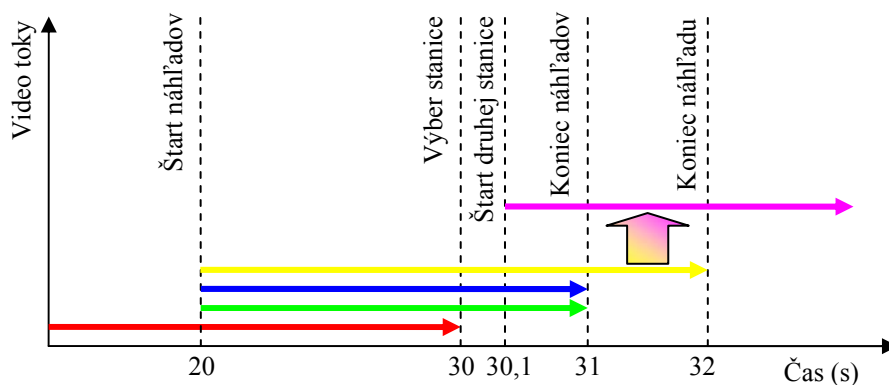
n0 -> n1	100Mbit/s	20 ms
n1 -> n2	100Mbit/s	30 ms
n2 -> n3	25Mbit/s	50 ms
n3 -> n4	5Mbit/s	100 ms

Hodnoty pre vyrovnávacie pamäte smerovačov, pre body n1 a n2 boli 10 paketov a pre bod n3 to bola hodnota 5 paketov. Nasledujúce simulácie obsahovali zakaždým základné dva toky, prvým bol prenos videa prostredníctvom UDP protokolu rýchlosťou 3.5 Mbit/s, čo približne odpovedalo súčasnemu prenosu videa v štandardnom rozlíšení prostredníctvom IPTV služieb, druhým tokom bol TCP riadiaci kanál pre potrebu spúšťania požiadaviek IPTV služby ako aj na ostatné funkcie riadenia služby. Veľkosť paketu UDP prenosu bola podľa [5] zvolená na 1368 bajtov, rýchlosť prenosu na tzv. poslednej míli bola zvolená na základe dostupných meraní reálnych liniek. Ako najvhodnejší algoritmus riadenia zahŕňajúci prenos multimediálnych tokov sa javil TFWC.

A. Proces voľby a konečného prepnutia

Na základe dosiaľ uskutočnených pokusov je táto simulácia celkového procesu ponúknutia náhľadov a zároveň prepnutia medzi video tokmi realizovaná s týmito parametrami: rýchlosť linky 11 Mbit/s, hlavný video tok 3,5 Mbit/s, náhľadové toky 400 kbit/s. Červenou je znázornený hlavný sledovaný tok, ktorý je vystriedaný v závere ružovým tokom, po prechode zo zvoleného žltého náhľadového toku. Modrou a zelenou sú zobrazené dodatočné náhľadové toky, ako znázorňuje Obr. 5. Farebné zobrazenie je jednotné pre časové znázornenie priebehu tokov ako aj pre výsledky simulácie.

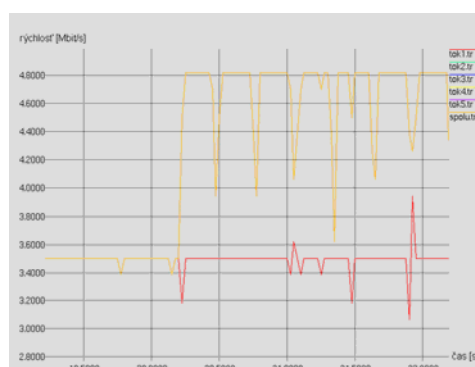
Na Obr. 6 sú zobrazené detaily simulácie. Obrázok (b) je začiatok prenosu náhľadov v čase 20 sekúnd, ktorý výrazne neovplyvňuje rýchlosť sledovaného prvého toku. Obrázok (c) zobrazuje stabilný priebeh rýchlosti počas zobrazenia troch náhľadov, (d) spolu s (e) moment prepnutia dvoch hlavných tokov. Na obrázku (f) je detail priebehu rýchlostí troch náhľadových tokov.



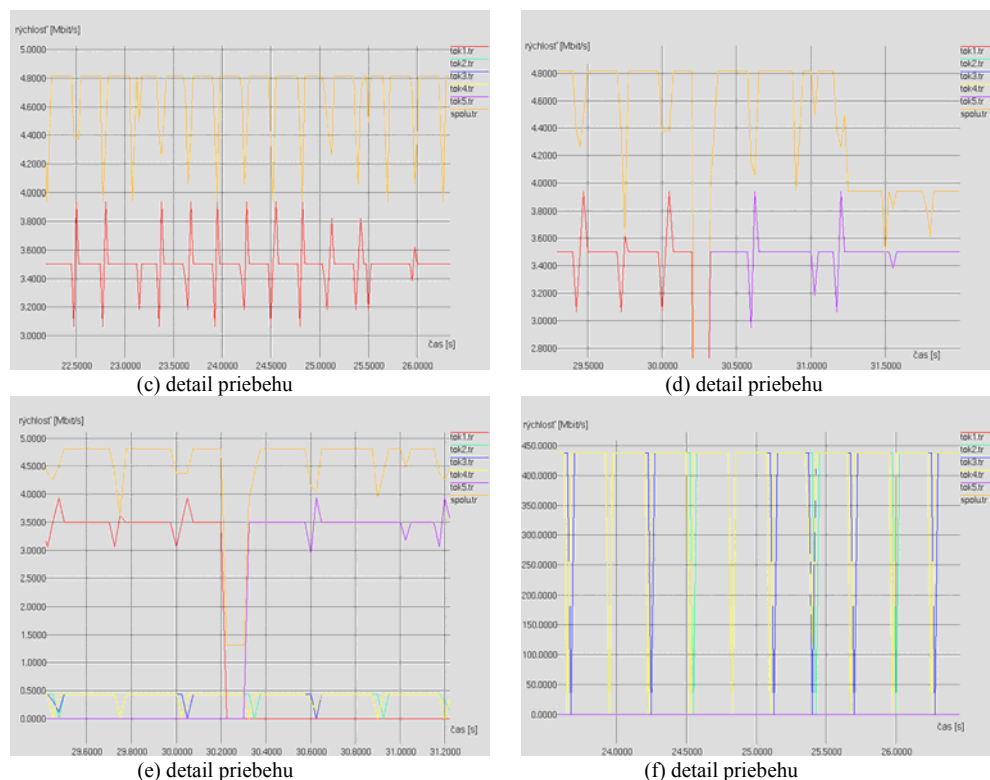
Obr. 5 Proces prepnutia sledovaného kanálu s použitím náhľadov



(a) kompletný priebeh



(b) detail priebehu



Obr. 6 Kvalita hlavného videotoku pre celkovú rýchlosť siete 11 Mbit/s

IV. ZHRNUTIE

Na základe uskutočnených simulácií, z ktorých viaceré pre nedostatok miesta nie sú uvedené v tomto článku, môžeme potvrdiť reálnosť nasadenia nielen prediktívneho prepínania stanic, ale aj použitie viacerých náhľadových videotokov za účelom skvalitnenia prepínania medzi jednotlivými tokmi.

Aby bol poskytovateľ schopný zabezpečiť dostatočne rýchlu službu prediktívneho prepínania, je pre neho nevyhnutné k smerovaču, ktorý je najbližšie na ceste k užívateľovi, použitie fyzickej linky s minimálnou odozvou. Táto odozva sa totiž môže výrazne prejaviť hlavne na "poslednej míli" spojenia. Sieťové zariadenia, no najmä použité smerovače, ktoré spracovávajú prioritu viacerých súperiacich tokov musia byť nastavené tak, aby mal najvyššiu prioritu hlavný sledovaný videotok v prípade, že dôjde k zhoršeniu podmienok prenosovej linky.

POĎAKOVANIE

Táto práca bola podporovaná Kultúrnou a edukačnou grantovou agentúrou Ministerstva školstva Slovenskej republiky VEGA na základe projektu č. 1/0045/10 (50%) a realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] J. G. Apostolopoulos – W. Tan – S. J. Wee, "Video streaming: Concepts, Algorithms, and Systems", Hewlett-Packard Laboratories, 2002, s. 1-14.
- [2] Coding of moving pictures and associated audio for digital storage media a tup to about 1.5 Mbits/s - ISO/IEC 11172, International Organization for Standardization, 1993.
- [3] P. Kocan, Prediktívne prepínanie video tokov v internetovej televízii", Dizertačná práca, Technická univerzita v Košiciach, jún 2010.
- [4] M. Welzl, "Network Congestion Control, Managing Internet Traffic", John Wiley & Sons, Ltd., 2005, ISBN-13 978-0-470-02528-4.
- [5] Advanced video coding for generic audiovisual services, ITU-T Recommendation H.264, International Telecommunication Union, 2010.
- [6] D. Austerberry, "The Technology of Video & Audio Streaming", Focal Press, 2005, s. 3-12, ISBN 0240805801.
T. Anttalainen, "Introduction to Telecommunications Network Engineering", Artech House, 2003, s.269-277, ISBN 1-58053-500-3.

Metódy maskovania chýb založené na filtrácii pohybových vektorov

*Branislav HRUŠOVSKÝ, Stanislav MARCHEVSKÝ,
Bohumil FRANKO, Martin LIPTAJ*

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

branislav.hrusovsky@tuke.sk, stanislav.marchevsky@tuke.sk,
bohumil.franko@student.tuke.sk, martin.liptaj@tuke.sk

Abstrakt — Videokomunikácia prostredníctvom bezdrôtových kanálov je stále veľkým problémom v dôsledku obmedzenej šírky pásma a prítomnosti kanálových chýb. Nízka prenosová rýchlosť mobilných sietí si vyžaduje zníženie použitého video rozlíšenia ako aj použitie vysokej účinnej techniky videokompresie. Výsledkom prenosu videa prostredníctvom bezdrôtového kanála je strata dát. To spôsobuje vznik vizuálnych artefaktov, a teda zreteľný pokles kvality obrazu. Keďže prenos videotoku v reálnom čase je obmedzený oneskorením prenosového kanála, nie je možné opätovne zasielať všetky chybné alebo stratené informácie. Je preto nevyhnutné zakryť tieto chyby na strane dekodéra pomocou vhodných metód, akými sú napríklad metódy maskovania chýb.

Kľúčové slová — makroblok, bloková chyba, maskovanie chýb, pohybový vektor, pixel

I. ÚVOD

Stratovosť paketov a jej vplyv na dekódované video je možné obmedziť viacerými spôsobmi. Existujú viaceré nástroje na zabezpečenie videotoku pred jeho prenosom. Ďalším riešením môže byť použitie opätovných prenosov stratenej informácie. Takéto metódy však zvyšujú množstvo informácie, ktorú je nutné preniesť cez sieť, respektíve zavádzajú oneskorenie. Preto do popredia vystupujú metódy maskovania chýb, ktoré sa zameriavajú na obnovu stratenej informácie z informácie, ktorá je k dispozícii. Kým prvá skupina metód spoločne s opätovnými prenosmi sa snaží o bezstratovú obnovu, metódy maskovania chýb nasadené v dekodéri, prípadne až za ním, sa snažia získať čo najbližšiu aproximáciu originálneho obrazu. Kým pri dátových prenosoch je takýto prístup neprijateľný, pri maskovaní strát v poškodenom videu je možné využiť určitú nedokonalosť ľudského oka. Pri strate paketov dochádza k strate dodatočnej kódovacej informácie, ku ktorej patria aj pohybové vektory. Jedným z možných spôsobov ich opätovného získania je použitie metód číslicovej filtrácie.

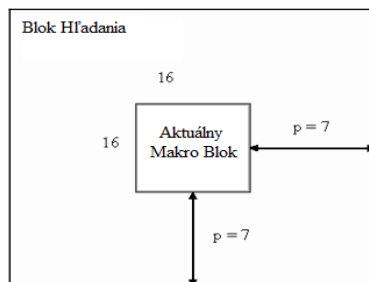
II. MASKOVANIE CHÝB

Metóda maskovania chýb založená na filtrácii pohybových vektorov sa zaraďuje do skupiny maskovania chýb v časovej oblasti. Delíme ich na dve skupiny. Prvá skupina súvisí s poruchami bitového toku inter-snímky [3]. Patrí sem obnova informácií na strane interpretácie, akými sú vektory pohybu alebo referenčný počet snímok. Druhá skupina súvisí s obnovou úplne zničených makroblokov v intra-snímках aj inter-snímках [4].

A. Pohybové vektory

Základným predpokladom odhadovaného pohybu je to, že vzory zodpovedajúce objektom a pozadia v snímke videosekvencie sa pohybujú v rámci snímky tvorenej odpovedajúcimi objektmi na ďalšej snímke. Myšlienkou porovnávania blokov je rozdeliť aktuálne snímky na matice - 'makrobloky' a tie následne porovnávať s odpovedajúcim blokom a jeho priľahlými susedmi v predchádzajúcej snímke. Získame vektor, ktorý stanovuje pohyb z jedného miesta na druhé v predchádzajúcej snímke pre každý makroblok aktuálnej snímky [3]. Oblasť hľadania pre porovnávanie makroblokov medzi predchádzajúcou a aktuálnou snímkou je obmedzená vyhľadávacím parametrom p pre všetky štyri strany odpovedajúceho makrobloku. Väčšie pohyby vyžadujú väčší parameter p a väčší parameter p je výpočtovo náročný na odhad

pohybu. Makroblok je obvykle chápaný ako štvorec, ktorého strana je veľká 16 pixelov a vyhľadávaci parameter p pre takúto veľkosť štvorca je rovný hodnote 7. Myšlienka je znázornená na Obr. 1.



Obr. 1 Makro blok o veľkosti 16 pixlov a parameter hľadania $p = 7$ pixelov

Pomocou odhadovej funkcie sa určí výsledný pohybový vektor pre aktuálny makroblok. Existujú rôzne odhadové funkcie, z ktorých najpoužívanejšie sú MAD (1) a MSE (2). Spomedzi týchto dvoch odhadových funkcií je najmenej výpočtovo náročná MAD funkcia [6].

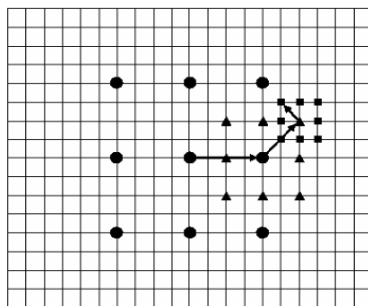
$$MAD = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |C_{ij} - R_{ij}| \quad (1)$$

$$MSE = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (C_{ij} - R_{ij})^2 \quad (2)$$

Kde N je veľkosť strany makro bloku, C_{ij} a R_{ij} sú pixely, ktoré sú porovnávané medzi aktuálnym a predchádzajúcim makroblokom.

B. TSS – algoritmus pre získanie pohybových vektorov

Tento algoritmus je známy ako takzvaná trojkroková metóda (z angl. three step search). Bol to jeden z prvých pokusov vytvorenia rýchleho algoritmu a siaha do polovice roku 1980. Začína sa s hľadaním miesta v centre makrobloku. Veľkosť kroku S sa nastaví na 4 pre parameter hľadania $p=7$. Algoritmus potom hľadá na ôsmich pozíciách $\pm S$ pixelov v okolí miesta (0,0). Z týchto lokalít sa vyberie jedna lokalita, ktorá ponúka najmenší odhad. Potom sa nastaví veľkosť kroku na $S/2$ a celý cyklus sa opakuje, až kým veľkosť kroku $S=1$. Získaná lokalita ponúka pre aktuálny makroblok najzhodnejšiu oblasť pixelov z predchádzajúceho makrobloku. Táto myšlienka je znázornená na Obr. 2 [6].



Obr. 2 Princíp TSS, pohybový vektor (5,-3)

C. Obnova stratených pohybových vektorov pomocou Kalmanovej filtrácie

Straty pohybových vektorov vznikajú dôsledkom toho, že snímka v sebe nesie chybu (obsahuje chybné makrobloky). Hodnoty pohybového vektora pre takúto snímku nie sú korektné. Hodnota pohybového vektora oproti pohybovému vektoru, ktorý by bol získaný z originálnej snímky (snímka nenesie v sebe chybu) je o smer alebo veľkosť pohybu makrobloku rozdielna. Existuje viacero spôsobov ako získať tieto stratené pohybové vektory späť. Jedným zo spôsobov je použitie Kalmanovho filtra [5].

1) Kalmanov filter

Kalmanov filter je optimálny rekurzívny algoritmus na spracovanie údajov. Tento algoritmus je nástroj, ktorý zo série neúplných zašumených meraní odhaduje súčasný stav dynamického systému. Princíp Kalmanovho filtra je založený na dvoch neustále sa opakujúcich základných fázach: predikcia (z angl. predict) a aktualizácia odhadu (z angl. update) [7]. Princíp rekurzívnych filtrov spočíva v tom, že na odhad parametrov v okamihu určenia nie je potrebné poznať predchádzajúce hodnoty observácií, ale len hodnoty posledného odhadu a nové meranie, ktoré systém aktualizuje [2]. Koefficienty Kalmanovho filtra sa v každom takte upravujú na základe dostupnej informácie tak, aby poskytl optimálny odhad signálu.

a) Predikcia

Predikcia predpovedá stav systému (hodnotu odhadovanej veličiny) a jeho kovariančnú maticu v epoche k na základe jeho odhadu pre epochu $k-1$, ako uvádzajú (3) a (4).

$$\hat{x}_k^- = \Phi_k \hat{x}_{k-1} + B_k u_k \quad (3)$$

$$P_k^- = \Phi_k P_{k-1} \Phi_k^T + B_k \Gamma B_k^T + Q \quad (4)$$

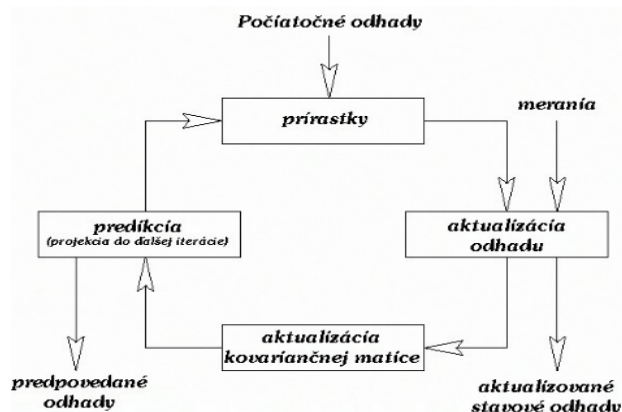
Kde $\hat{x}_k^-$ predstavuje odhad určovanej veličiny, P je kovariančná matica, B je riadiaca matica, ktorá zväzuje známy vektor u s odhadovanými parametrami, Γ je kovariančná matica predstavujúca neistotu určenia známych vstupných veličín u , Φ je transformačná (tranzitná) matica, ktorá definuje vzťah medzi predchádzajúcou a súčasnou epochou, Q je kovariančná matica bieleho šumu určovaného procesu [7]. Pomocou predikcie vypočítame apriórne predikcie stavu a chyby kovariancie [1].

b) Aktualizácia odhadu

Matica prírastkov K_k vyjadruje váhu súčasného merania vzhľadom na odhadovanú veličinu a určí sa podľa (5)

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \quad (5)$$

Kde R_k je kovariančná matica merania. Po jednoduchšej analýze (5) je zrejmé, že s presnejším meraním (tj. znižujúcou sa kovariančnou maticou meraní) vzrastá jeho váha a naopak. Ak sa k nule blíži apriórna kovariančná matica, potom meraniu, resp. apriórnemu reziduu bude prisúdená malá váha. Princíp Kalmanovho filtra je znázornený na Obr. 3.



Obr. 3 Princíp Kalmanovho filtra

Apriórne reziduum $\tilde{e}$ sa definuje ako rozdiel medzi skutočnou hodnotou aktuálneho merania a očakávaným meraním, určeným z posledného odhadu parametrov, ako vyjadruje (6).

$$\tilde{e} = z_k - H_k \hat{x}_k^- \quad (6)$$

potom aktualizovanú hodnotu odhadovanej veličiny určíme podľa (7):

$$\hat{x}_k = \hat{x}_k^- + K_k \tilde{e}_k \quad (7)$$

Aktualizovanú kovariančnú maticu určíme podľa (8):

$$P_k = (I - K_k H_k) P_k^- \quad (8)$$

Extrapoláciu pre epochu $k+1$ vyjadrujú (9),(10):

$$\hat{x}_{k+1}^- = \Phi_k \hat{x}_k \quad (9)$$

$$P_{k+1}^- = \Phi_k P_k \Phi_k^T + Q \quad (10)$$

c) Inicializácia

Každý rekurzívny filter vyžaduje inicializáciu. Ide o jednoduchý proces, ktorý na začiatku výpočtu definuje hodnoty základných veličín. Vektoru určených parametrov x_0 priradíme hodnotu určenú z počiatočného merania, alebo definujeme ako nulovú. Pri inicializácii kovariančnej matice P_0 je najvhodnejším prístupom definovanie diagonálnej matice s dostatočne veľkými členmi, ktorá v ďalšej iterácii nebude mať takmer žiadnu váhu [7].

III. SIMULAČNÉ VÝSLEDKY

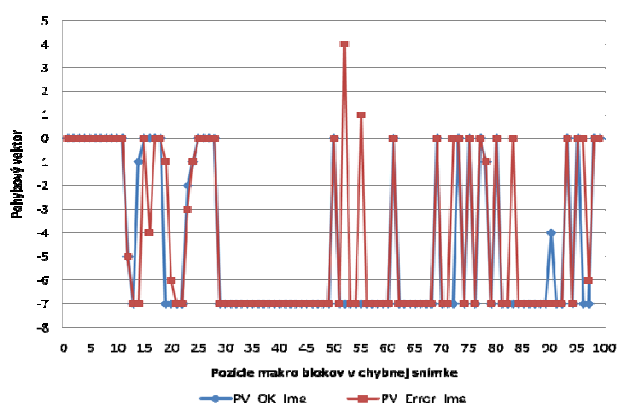
Simulácia bola realizovaná pre videosekvencie *Mother_Daughter.avi*, *Suzie.avi* a *Foreman.avi*. Pre účel simulácie sme veľkosť makrobloku nastavili na 16x16 pixelov a pre tento makroblok bola vytvorená bloková chyba, v našom prípade to bol blok veľkosti 8x8 pixelov. Pre názornosť uvádzame výsledok zamaskovania snímky videosekvencie *Mother_Daughter*. V Tab.1 uvádzame kvalitu rekonštrukcie všetkých troch videosekvencií pri rôznom počte chybných blokov v snímke.



Obr. 4 Chybná a opravená snímka č. 174 videosekvencie *Mother_Daughter*. Počet chybných blokov je 80

Opravená snímka na pravej strane obsahuje čierne stopy po chybných blokoch. Dôvod nedokonalého zamaskovania niektorých blokov je popísaný v časti Nezamaskovanie čiernej stopy. Dôležité informácie ako obrysy tváre ostali zachované, nerozmazali sa, ako vidno na Obr. 4. Kvalitu opravených snímok nám určuje najmä hodnota PSNR. Hodnota PSNR napríklad pre opravenú snímku č.174 videosekvencie *Mother_Daughter* bola 24.05 dB.

Kvalita rekonštrukcie snímky závisí okrem iných faktorov od množstva pohybu a dynamiky v danej skupine snímok. Graficky zobrazené rozdiely medzi získaným pohybovým vektorom z originálnej snímky (bez chýb) a opraveným pohybovým vektorom sú znázornené len pre hodnotu riadku pohybového vektora. Pre získanie strateného vektora bola použitá Kalmanová filtrácia. Pohybové vektory boli získané pomocou metódy TSS.



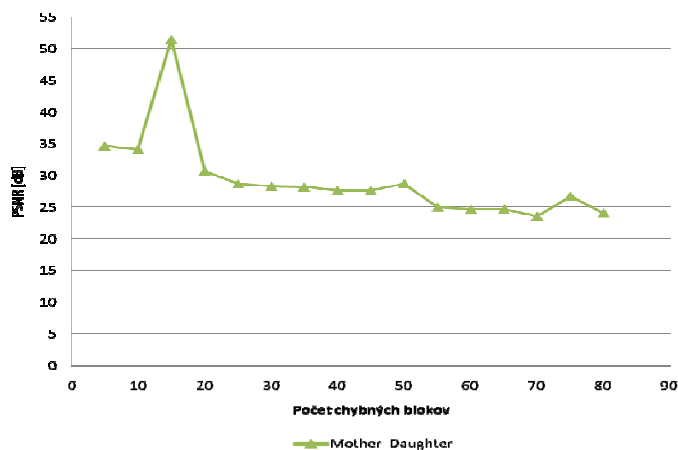
Obr. 5 Pohybové vektory pre snímku č. 174 videosekvencie *Mother_Daughter* s 80 chybnými blokmi

Z grafického zobrazenia sú zreteľné odchýlky medzi hodnotami získané pomocou Kalmanovej filtrácie stratených pohybových vektorov a hodnotami pohybových vektorov získaných z originálnej snímky, čo znázorňuje Obr. 5. Pri zohľadnení vysokého počtu stratených pohybových vektorov môžeme konštatovať, že pomocou Kalmanovej filtrácie sme získali pohybové vektory, ktoré sú totožné s pohybovými vektormi získanými z originálnej snímky.

S takto presne získanými hodnotami stratených vektorov vieme dosiahnuť pomerne kvalitné maskovanie chýb, čoho dôkazom je aj hodnota PSNR.

Tabuľka 1
 KVALITA OPRAVENÝCH SNÍMOK JEDNOTLIVÝCH VIDEOSEKVENCIÍ

Počet chybných blokov v snímke	PSNR [dB]		
	Suzie	Foreman	Mother Daughter
5	40.38	35.28	34.63
10	35.15	34.62	34.07
15	33.56	25.67	51.42
20	29.66	27.06	30.64
25	28.00	31.62	28.64
30	35.55	25.82	28.27
35	28.96	37.00	28.18
40	29.65	23.05	27.61
45	27.74	23.19	27.61
50	26.98	21.02	28.68
55	25.31	24.88	24.97
60	25.43	32.34	24.65
65	30.38	22.46	24.68
70	27.25	21.34	23.52
75	26.39	21.48	26.73
80	24.46	23.13	24.05
Priemer PSNR[dB]	29.68	26.87	29.27



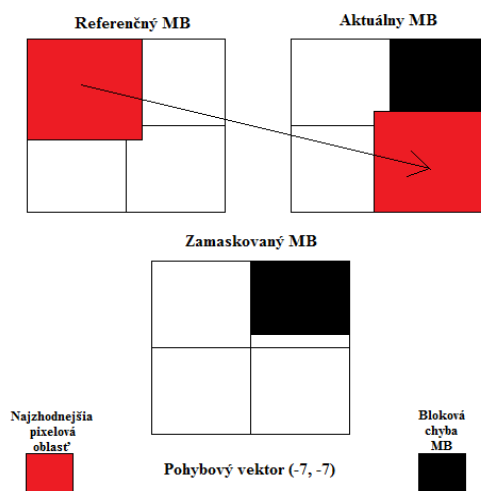
Obr. 6 PSNR pre snímky s rôznym počtom chybných blokov

A. Nezamaskovanie čiernej stopy

Z Obr. 6 je zrejmé, že najlepšie hodnoty kvality sme dosiahli pri najmenšom počte chybných blokov. Vtedy sa totiž zachovala väčšia časť snímky ako v prípade s najväčším počtom chybných blokov. Priemerná hodnota PSNR je 29.27 dB, čo predstavuje kvalitne zamaskovanú snímku aj napriek tomu, že zamaskovaná snímka obsahovala čierne stopy po chybných blokoch. Algoritmus vytvorený pre metódu maskovania chýb založenú na filtrácii pohybových vektorov vytvára snímky na základe princípu, ktorý je znázornený na Obr. 7. Pomocou metódy TSS sa vygenerujú pohybové vektory aktuálnej snímky. Aktuálna snímka však obsahuje makrobloky, v ktorých sa nachádza bloková chyba. Na získanie stratených pohybových vektorov sa použije Kalmanová filtrácia. Hodnoty pohybového vektora poskytujú informácie o smere a veľkosti posunutia referenčného makrobloku, vzhľadom ku aktuálnemu makrobloku.

Predpokladajme tento prípad: pohybový vektor má hodnoty (-7, -7). Tieto dve hodnoty nám hovoria o tom, že referenčný makroblok sa vzhľadom ku aktuálnemu makrobloku posunie o 7 pixelov od nultého bodu v smere nadol a o 7 pixelov od nultého bodu v smere doprava. Pozícia nultého bodu sa nachádza na ľavom hornom okraji makrobloku. Takýmto pohybom sme preniesli oblasť pixelov z referenčného makrobloku do aktuálneho makrobloku, ako znázorňuje Obr. 7. Naša bloková chyba sa však nebude nachádzať presne na pozícii, kde boli prenesené

dáta z referenčného makrobloku, ale na inom mieste, pozri Obr. 7. V našom prípade sa zamaskuje len hranica chybného bloku a zamaskovaný chybný blok ešte stále bude obsahovať chybu. Preto je možné pozorovať stopy po chybných blokoch. Niektoré blokové chyby makrobloku sa zamaskujú celé preto, že pozícia chybného bloku zodpovedala pozícií prenesenej oblasti pixelov z referenčného makrobloku.



Obr. 7 Princíp skladania snímku pomocou pohybových vektorov

IV. ZHRNUTIE

Z výsledkov simulácii môžeme konštatovať, že pri zväčšujúcom sa počte chýb v snímke klesá kvalita opraveného videa. Toto tvrdenie platí pre všetky videosekvencie. Kvalita pre rôzne videosekvencie s rovnakým počtom chýb bude odlišná. Tento rozdiel je spôsobený rozdielnou dynamikou medzi predchádzajúcou a nasledujúcou snímkom v jednotlivých videosekvenciách. Z uvedeného vyplýva, že pre videosekvenciu s najväčšou dynamikou medzi jednotlivými snímkami bude kvalita opraveného videa najhoršia. Pre efektívnejšie zmenšenie množstva viditeľných artefaktov v zamaskovanej snímke by sme navrhovali zmenšiť parameter hľadania p , s ktorým pracuje metóda TSS generujúca pohybové vektory. Toto potenciálne vylepšenie má však tú nevýhodu, že pohybové vektory budú v takom prípade vytvorené z menšej oblasti. To by malo za následok rozmazanie snímky, ktorá bude vykazovať veľkú zmenu polohy voči predchádzajúcej snímke. Pre získanie stratených pohybových vektorov sme využili Kalmanovu filtráciu. Hodnoty takto získaných pohybových vektorov boli veľmi podobné hodnotám pohybových vektorov získaných z originálnej snímky. Tento zistený poznatok potvrdzuje to, že pomocou Kalmanovej filtrácie je možné sa dopracovať k veľmi presným parametrom.

POĎAKOVANIE

Táto práca bola podporovaná Kultúrnou a edukačnou grantovou agentúrou Ministerstva školstva Slovenskej republiky VEGA na základe projektu č. 1/0045/10 (50%) a realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] I.E.G. Richardson, "H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia", John Wiley & Sons, 2003. 306 s. ISBN 0-470-84837-5
- [2] J. Škorvánek, "Algoritmy maskovania chýb pri prenose videotokov cez WiFi siete", Bakalárska práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 2008. 55 s.
- [3] B. Hrušovský, "Metódy maskovania chýb spôsobených prenosom videotokov cez bezdrôtové prostredie", Diplomová práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 2009. 87 s.
- [4] O. Rais, "H.264 (MPEG-4 part 10), 2004. Dostupné na internete : <http://www.golias.cz/index.php?modul=audio&sub=clanky&page=clanek&id=1697>.
- [5] J. Bezručka, "Využitie Kalmanovho filtra v geodézii a navigácii - Kalmanov filter pre začiatočníkov". Bratislava : Slovenská Technická univerzita, Stavebná fakulta, 14 s. Dostupné na : <http://mimmon.net/mimo/dokumenty/kalman.pdf>.
- [6] P. Trebatický, "Modifikácia Kalmanovho filtra pre tréning neurónových sietí", Bratislava: Slovenská Technická univerzita v Bratislave, Fakulta informatiky a informačných technológií. 8 s.. Dostupné na internete: <http://hilbert.chtf.stuba.sk/KUZVII/abstracts/Trebaticky.pdf>.
- [7] B. Franko, "Metódy maskovania chýb založené na filtrácii pohybových vektorov", Bakalárska práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 2010. 59 s.

Prehľad základných metód maskovania stratených paketov v priestorovej oblasti pri prenose videa zašumeným prostredím

Branislav HRUŠOVSKÝ, Stanislav MARCHEVSKÝ, Martin LIPTAJ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

branislav.hrusovsky@tuke.sk, stanislav.marchevsky@tuke.sk, martin.liptaj@tuke.sk,

Abstrakt — V súčasnej dobe sú audiovizuálne a multimediálne služby vnímané ako dôležité zdroje prenosu dát v rámci mobilných sietí. Jedno z obmedzení mobilných sietí je nízka prenosová rýchlosť, čo vyžaduje zníženie rozlíšenia videa a použitie vysoko efektívnej techniky videokompresie. Štandard H.264/AVC, ktorý je spomínaný v tomto článku, je najnovším kodekom vo videokompresii poskytujúci výrazné zlepšenie kvality v porovnaní s predošlými videoštandardmi. Prenos videotoku bezdrôtovým prostredím vedie k strate určitého množstva dát. To spôsobí vznik viditeľných artefaktov, a teda výrazné zníženie kvality obrazu. Keďže prenášanie videotoku v reálnom čase je obmedzené oneskorením prenosového kanála, nie je možné znovu prenášať všetky poruchové alebo stratené pakety. Preto je nevyhnutné tieto vzniknuté poruchy zamaskovať. V tomto článku sú popísané základné algoritmy maskovania chýb, ktoré vzniknú pri prenose videa cez zašumené prostredie.

Kľúčové slová — maskovanie chýb, videodáta, makroblok, snímka

I. ÚVOD

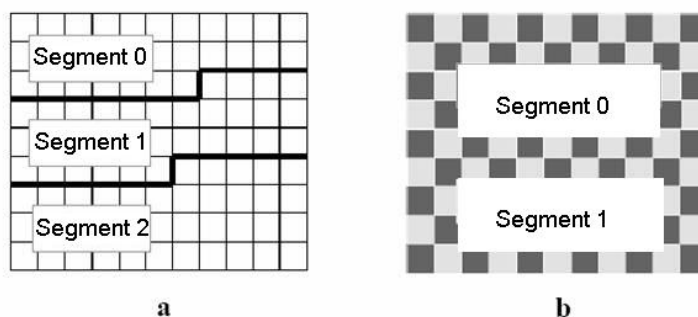
Vďaka prudkému vývoju bezdrôtových komunikácií si prenos videa bezdrôtovým prostredím vyslúžil veľkú pozornosť. Dôležitou charakteristikou bezdrôtových sietí je, že neposkytujú garantovanú kvalitu služieb QoS. Pri prenose dát tak dochádza k častým stratám paketov, rušeniu prenosu signálu, kolísaniu hodnôt pomeru signál / šum a pod. V súčasnosti existuje veľký počet služieb a prenosových médií, ako káblový modem, xDSL alebo UMTS, ktoré poskytujú nižšie prenosové rýchlosti v porovnaní s vysielacími kanálmi ako HDTV alebo DVB-T. Prihliadnuc na pokroky vo výkonoch procesora a pamäte a na pokroky v technológiách videokódovania narastá potreba efektívnej videokompresie s vysokou účinnosťou a odolnosťou kódovania v rôznych prenosových prostrediach. ITU (Medzinárodná telekomunikačná únia) spolu s ISO/IEC (Medzinárodná organizácia pre štandardizáciu / Medzinárodný elektrotechnický výbor) vytvorili skupinu expertov (JVT), ktorí definovali H.264 ako zdokonalený videokódovací štandard, ktorý rozvinul predchádzajúce štandardy H.263 a MPEG-4 časť 2. H.264 na jednej strane zabezpečuje vysokú kompresiu pre rôzne aplikácie (napríklad videokonferenciu), na druhej strane prináša flexibilitu v rôznych sieťových prostrediach.

Vo všeobecnosti sú mobilné siete charakterizované obmedzenou šírkou pásma. Podporovaná dátová prenosová rýchlosť je preto stále limitovaná. Je teda potrebné použiť vysokoúčinnú videokompresiu a malé rozlíšenie (QCIF 176x144), čo podporuje väčšina súčasných mobilných terminálov. Keďže všetky techniky videokompresie sú založené na eliminácii priestorovej a časovej redundancie v rámci videosekvencie, videotok je citlivý na prenosové poruchy. Tieto poruchy spôsobujú stratu časti videoobsahu, čo zapríčiní degradáciu výslednej kvality obrazu. Vizualne efekty týchto porúch sa šíria priestorovo do susedných oblastí v snímke a časovo do nasledujúcich snímok v dekódovanom poradí. Mechanizmus na odhalenie porúch môže byť integrovaný na stranu dekodéra, ktorý analyzuje prenosové chyby a lokalizuje chýbajúcu časť videosekvencie. Napriek znehodnoteniu celých balíkov dát a požadovaných opakovaných prenosov, dekodér dokáže obnoviť stratenú informáciu zo správne prijatých dát. Nazývame to maskovanie chýb.

II. H.264/AVC

H.264/MPEG-4 AVC (Zdokonalené video kódovanie) je najnovší štandard video kompresie, ktorý zaručuje značné výhody v porovnaní s predošlými štandardmi. Úmyslom pri návrhu H.264/AVC bolo vytvoriť štandard, ktorý by bol schopný pracovať s dobrou obrazovou kvalitou pri bitových rýchlostiach, ktoré sú podstatne nižšie ako tie, ktoré boli potrebné pri predošlých normách. Nový štandard poskytuje minimálne dvojnásobné zlepšenie kompresie oproti najlepším predošlým štandardom a značné percentuálne zlepšenie kvality v oboch prípadoch MPEG-2 a MPEG-4 [1]. Únia ITU-T pomenovala tento štandard H.264 (predtým nazývaný H.26L), zatiaľčo ISO/IEC pomenovala štandard MPEG-4 AVC (Zdokonalené video kódovanie), ktoré sa stalo časťou 10 štandardu MPEG-4 [1]. Štandard H.264/AVC dosahuje najvyššiu kompresiu dát spomedzi všetkých štandardov. Podporuje videoaplikácie vrátane bezdrôtových nízkorýchlostných aplikácií, kvalitný a vysokokvalitný obraz televízneho vysielania, prenos videa cez internet, prenos vysoko kvalitného DVD obsahu a videa najvyššej kvality pre aplikácie digitálneho filmu.

Zakódované videosekvencie v H.264/AVC pozostávajú z postupnosti kódovaných snímok (niekedy nazývaných rámce). Základné stavebné bloky štandardu sa nazývajú makrobloky, pre ktoré je stanovený kódovací a dekódovací proces. Sú organizované do segmentov (výrezov), ktoré vo všeobecnosti reprezentujú podmnožinu danej snímky. Každý segment môže byť dekódovaný samostatne. Segmenty sú skupina makroblokov, pričom každú snímku možno rozdeliť do jedného alebo viacerých segmentov, ako znázorňuje Obr.1. Poradie pri prenose makroblokov v toku bitov závisí od takzvanej Makroblokovej alokačnej (rozdeľovacej) mapy – MAM, a teda nazavisí nevyhnutne od poradia snímania, ktoré sa štandardne realizuje po riadkoch. Každý segment môže byť presne dekódovaný bez použitia dát z iných segmentov tej istej snímky. Počet makroblokov v každom segmente môže byť nastavený na konštantnú hodnotu alebo môže byť nastavený podľa stáleho počtu bajtov. Keďže každý makroblok je reprezentovaný premenlivým počtom bitov, kódér používa vysúvanie bitov na naplnenie segmentu, aby dosiahol požadovaný počet bajtov. Segment môže byť tiež vytvorený použitím techniky FMO (flexibilného makroblokového usporiadania), ako znázorňuje Obr.1 vpravo [2].



Obr. 1 Delenie videosnímkov : a, základný tvar segmentov b, FMO tvar segmentov

III. MASKOVANIE CHÝB

Ako už bolo spomenuté, strata prenášaných paketov ovplyvňuje kvalitu prijatého videa. Tento problém je spôsobený obmedzenou šírkou pásma kanála použitého v mobilných komunikačných sieťach. Keďže prenášanie videotoku v reálnom čase je obmedzené oneskorením kanála, nie je možné opätovne prenášať všetky chybné alebo stratené pakety. Vznikla potreba metód dodatočného spracovania, ktoré sa pokúšajú obmedziť vizuálne javy spôsobené poruchou bitového toku po lokalizácii chýbajúcich alebo poruchových častí videodát. Metódy maskovania chýb, ktoré budú implementované na strane prijímača, obnovia chýbajúcu a porušenú časť videa použitím predošlých dekódovaných videodát. Maskovanie chýb podporuje priestorovú a časovú koreláciu medzi videoblokmi v rámci jednej alebo viacerých snímok videosekvencie. Metódy maskovania chýb sú preto realizované v priestorovej i časovej oblasti.

Maskovanie chýb realizované v priestorovej oblasti využíva videoinformáciu zo susediacich blokov tej istej snímky na obnovenie chýbajúcich pixelov (jasových a farebných hodnôt) v rámci stanovenej oblasti (napríklad jeden alebo viac makroblokov).

Maskovanie chýb realizované v časovej oblasti využíva videoinformáciu z blokov ležiacich v predchádzajúcich a nasledujúcich snímkach na obnovenie chýbajúcich pixelov (jasových a farebných hodnôt) v rámci stanovenej oblasti (napríklad jeden alebo viac makroblokov) [2].

A. Maskovanie chýb v priestorovej oblasti

Priestorová redundancia je vlastnosťou obrazu a videesignálov. Pri maskovaní chýb v priestorovej oblasti sa využíva medzipixelová odlišnosť medzi susednými pixelmi pre pôvodnú oblasť snímky. Medzipixelová odlišnosť je definovaná ako priemerná hodnota absolútnej diferencie medzi pixelom a jeho štyrmi okolitými pixelmi. Táto vlastnosť je využívaná na realizovanie maskovania chýb [2].

Všetky metódy maskovania chýb v priestorovej oblasti sú založené na myšlienke, že hodnoty pixelov (jasových i farebných) v rámci zničeného makrobloku môžu byť obnovené vhodnou kombináciou pixelov obkolesujúcich poškodený makroblok.

1) Metóda váženého spriemerovania

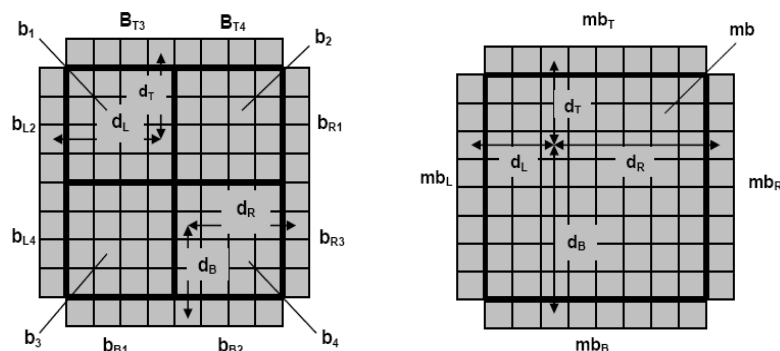
Princíp tejto metódy priestorového maskovania chýb spočíva v tom, že každý pixel v danom maskovanom makrobloku je vyrátaný ako lineárna kombinácia štyroch najbližších pixelov na hraniciach makrobloku (hornej, dolnej, ľavej a pravej hranici), zohľadňujúc vzdialenosť obnovovaného pixelu od jednotlivých hraníc makrobloku. Algoritmus sa vyznačuje pomerne jednoduchou výpočtovou metódou, preto je to najčastejšie používaná metóda na priestorové maskovanie chýb pri strate videoinformácie [1]. Tento algoritmus je vhodný len na maskovanie chýbajúcich makroblokov s homogénnym obsahom. Inak produkuje viditeľné artefakty.

V prípade, že používame makroblok s rozmermi pixelov $N \times N$, vážené spriemerovanie na úrovni makrobloku (z angl. macroblock based weighted averaging) môže byť popísané v (1), pričom $i, k = 1, 2, \dots, N$. Použité symboly sú uvedené na Obr.2.

$$mb(i, k) = \frac{1}{d_L + d_R + d_T + d_B} [d_R mb_L(i, 2N) + d_L mb_R(i, 1) + d_B mb_T(2N, k) + d_T mb_B(1, k)] \quad (1)$$

Iný spôsob ako realizovať vážené spriemerovanie sa nazýva blokovo založené vážené spriemerovanie (z angl. block based weighted averaging) [6]. Poškodený makroblok je rozdelený do štyroch samostatných blokov. Každý pixel v rámci bloku je nahradený z dvoch pixelov z jeho dvoch najbližších hraníc. V prípade, že využívame makroblok s rozmermi $2N \times 2N$, blokovo založené vážené spriemerovanie sú popísané v (2), kde $i, k = 1, 2, \dots, N$.

$$\begin{aligned} b_1(i, k) &= \frac{d_T b_{L2}(i, N) + d_L b_{T3}(N, k)}{d_L + d_T} \\ b_2(i, k) &= \frac{d_T b_{R1}(i, 1) + d_R b_{T4}(N, k)}{d_R + d_T} \\ b_3(i, k) &= \frac{d_B b_{L4}(i, N) + d_L b_{B1}(1, k)}{d_L + d_B} \\ b_4(i, k) &= \frac{d_B b_{R3}(i, 1) + d_R b_{B2}(1, k)}{d_R + d_B} \end{aligned} \quad (2)$$



Obr. 2 Vážené spriemerovanie a, blokovo založené b, makroblokovo založené

Metóda váženého spriemerovania založená na makrobloku vykazuje dobré výsledky v prípade, keď chýbajúci makroblok sa nachádza v rámci homogénnej oblasti (napríklad jasná obloha, hladké pozadie a pod.).

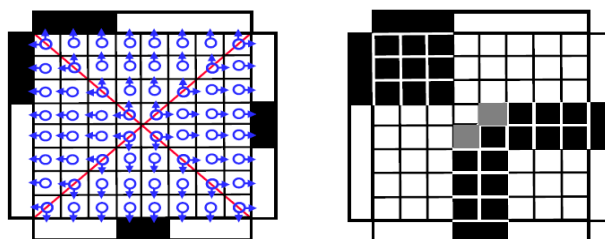
Na druhej strane metódy blokovo založeného váženého spriemerovania negarantujú hladkosť

obnoveného makrobloku a vykazujú jemný tzv. "blokový efekt". Táto blokovo založená metóda je omnoho účinnejšia ako makroblokovo založená metóda, ak chýbajúci makroblok sa skladá z dvoch alebo viacerých častí, pričom každá časť patrí k inej homogénnej oblasti. Použitie makroblokovo založeného váženého spriemerovania negarantuje homogénnosť videosignálu pozdĺž hraníc chýbajúceho makrobloku [1].

2) Metóda maximálneho vyhladzovania

Ďalšou metódou obnovy poškodeného makrobloku v priestorovej oblasti je metóda maximálneho vyhladzovania (z angl. maximal smoothing) [3].

Každý pixel pozdĺž hraníc poškodeného makrobloku musí byť hladko spojený s najbližším susedným pixelom v susediacom bloku. Taktiež aj ostatné pixely vnútri makrobloku musia byť hladko napojené na seba navzájom. Ako znázorňuje Obr.3, vnútorná hladkosť je vytvorená tak, že hodnoty pixelov z každej 1-pixel širokej hranice sú kopírované do trojuholníka daného poškodeného makrobloku. Táto metóda prináša malé zlepšenie priestorového maskovania chýb v oblasti pri hraniciach chýbajúceho makrobloku. V niektorých prípadoch ale táto metóda v porovnaní s metódou váženého spriemerovania znižuje hladkosť vnútri obnoveného makrobloku (vnútorných pixelov).



Obr. 3 Princíp metódy maximálneho vyhladzovania

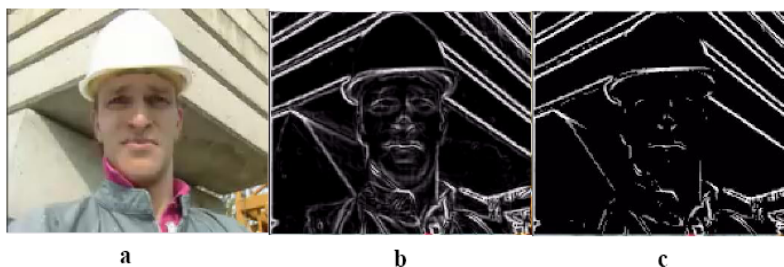
3) Metóda smerovej (direktívnej) interpolácie

Vylepšenie spomínaných metód maskovania chýb v priestorovej oblasti môže byť dosiahnuté, ak je v procese interpolácie uvažovaná spojitosť všetkých významných línií a hrán v susediacich blokoch, ktoré prechádzajú poškodeným makroblokom. To znamená, že vyhladzovanie musí byť vykonané pozdĺž všetkých významných hrán. Túto metódu nazývame smerová interpolácia (z angl. directional interpolation) [2], pretože všetky chýbajúce pixely poškodeného makrobloku sú interpolované v špecifickom smere. Je to rovnobežka so smerom obkolesujúcich línií a hrán. Táto metóda predpokladá len niekoľko jednoduchých dominantných hrán, a preto nefunguje dobre pre zložitejšie štruktúry ako sú nelineárne hrany a línie. Účinnosť metódy závisí na kvalite estimácie hrán.

Prvý krok potrebný pri detekcii hrán a línií susediacich blokov je vyšetrenie korelácie medzi dvoma a viacerými radmi pixelov obkolesujúcimi poškodený makroblok. Korelačné výpočty sú založené na hodnotách jas každého pixelu, pretože pre ľudský zrak sú zmeny jasových hodnôt viditeľnejšie ako zmena farbonosných hodnôt. Farbonosné hodnoty hraničných susediacich pixelov vykazujú väčšiu hladkosť, preto mechanizmus detekcie hrán založený na farbonosných hodnotách nie je taký účinný ako by bolo potrebné.

Rôzne obmedzenia tejto metódy znižujúce účinnosť maskovania môžu byť odstránené použitím takzvaného "Sobel" operátora, ktorý poskytuje detailnejšiu informáciu o pevnosti a smerovaní hrany [5]. Operátor sa skladá z dvoch konvolučných masiek s rozmermi 3x3. Nazývajú sa G_x a G_y a sú voči sebe pootočené o 90° . Sobelov operátor potrebuje osem susediacich pixelov na zistenie pevnosti a smerovania hrán v danej skupine pixelov. Poskytuje presnú informáciu o orientácii hrán. Sobelov operátor realizuje funkciu detektora hrán. Ak je ku detektoru hrán pridaný filter, Sobelov operátor rozpozná iba dominantné hrany. Všetky ostatné hrany budú odfiltrované. To, ktoré hrany sú dominantné, je definované nastavením prahovej úrovne. Všetky hrany s hodnotami pod touto úrovňou budú uznané ako hladká oblasť a hrany s hodnotami nad prahovou úrovňou budú uznané ako dominantné hrany.

Na Obr.4c je použitý filter s prahovou úrovňou 100. Všetky hrany s hodnotou pod 100 znázornené na obrázku Obr.4b sú odfiltrované a nahradené čiernymi pixelmi. Sobelov operátor bol použitý na každý pixel v snímke. Biele pixely znázorňujú maximálnu pevnosť hrán a čierna oblasť znázorňuje hladkú oblasť v snímke.

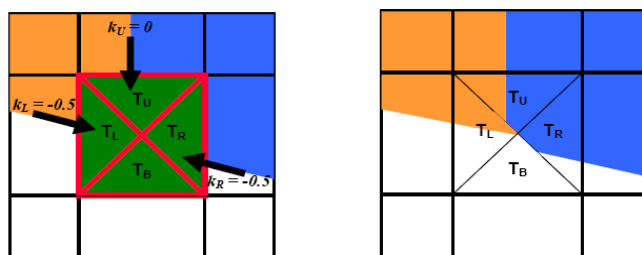


Obr. 4 Aplikovanie Sobelovho operátora na snímku a, b - bez použitia filtra, c - s použitím filtra ($|G| > 100$)

Vyšetrovanie smerového usporiadania $N \times N$ bloku aplikáciou Sobelovho operátora na každý pixel tohto bloku zvyšuje výpočtovú náročnosť procesu detekcie hrán, a to N^2 -krát. Výpočtová náročnosť môže byť znížená vynechaním niektorých pixelov pri zavedení Sobelovho operátora.

Keďže hlavnou úlohou priestorového maskovania chýb založeného na smerovej interpolácii je ochrana hrán na hraniciach chýbajúceho makrobloku, proces detekcie hrán môže byť ohrozený oblasťou blízko hraníc [3].

Pre každú dominantnú hranu priradíme časť chýbajúceho makrobloku, kde smerová interpolácia je aplikovaná adekvátne ku danému smeru hrany. Prerozdeľovanie chýbajúceho makrobloku môže byť nezávislé od pevnosti hrany a smerovania. V takomto prípade chýbajúci makroblok je rozdelený na rovnaké časti, ktoré majú rovnaké rozmery a formu. Na Obr.5 je rozdelený chýbajúci makroblok na štyri rovnaké trojuholníky. Každý trojuholník je obnovený interpoláciou chýbajúcich pixelov z pixelov v rámci 1-pixel širokých hraníc v smere detegovanej hrany susediaceho makrobloku. Napríklad chýbajúci horný trojuholník je obnovený interpoláciou chýbajúcich pixelov z pixelov v rámci 1-pixel širokej hornej hranice v smere k_U detegovanej hrany horného susedného makrobloku. Ďalšie rozdeľovanie makrobloku na vždy menšie rovnaké časti neznamena automaticky zlepšenie procedúry maskovania chýb. Závisí to na spôsobe rozdeľovania chýbajúceho makrobloku. Zvýši to iba náročnosť výpočtov.



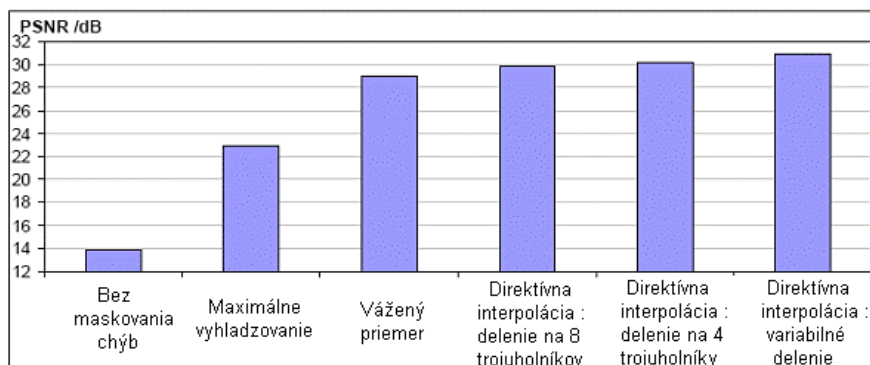
Obr. 5 Rozdelenie makrobloku na štyri rovnaké časti

V predchádzajúcom predstavovaní metód priestorového maskovania chýb sa predpokladalo, že chýbajúci makroblok je obklopený štyrmi makroblokmi. Tiež sa predpokladalo, že tieto štyri susedné makrobloky sú už dekodované bez akýchkoľvek porúch alebo akýchkoľvek strát na pôvodnej kvalite. No tento predpoklad sa môže naplniť, iba ak sa chýbajúci makroblok nenachádza na okraji snímky a ak maskovanie chýb je uskutočnené po dekodovaní všetkých makroblokov aktuálnej snímky. Posledný predpoklad môže byť splnený, ak aktuálna snímka je typu B alebo P [4].

V prípade snímky typu I tieto predpoklady nemôžu byť splnené, pretože všetky makrobloky nachádzajúce sa vedľa chýbajúceho makrobloku v dekodujúcom poradí (v smere napravo a dole) sú závislé na predtým dekodovaných makroblokoch. Tieto makrobloky sú automaticky ovplyvňované (rôznym spôsobom závisiacim na type predikcie použitej na ich zakódovanie) poruchami detegovanými v poškodenom makroblokoch, a preto nemôžu byť použité pri maskovaní chýb za účelom obnovy poškodeného makrobloku. Stratený makroblok môže byť preto obklopený maximálne dvoma makroblokmi na hornej a ľavej strane. Ak stratený makroblok je prvý makroblok v aktuálnej snímke typu I, počet dostupných susediacich makroblokov je rovný 0. Obnova takéhoto makrobloku nie je možná. Odporúča sa použiť bezpečnostné opatrenie na zabezpečenie bezchybného prenosu takýchto makroblokov. Tiež je rozumné použiť režim segmentovania snímky a obmedziť tak šírenie priestorových porúch za hranice postihnutého segmentu a zvýšiť množstvo neporušených makroblokov okolo chýbajúceho makrobloku v danom segmente.

4) Účinnosť algoritmov maskovania chýb v priestorovej oblasti

Na Obr.6 je znázornená účinnosť jednotlivých metód maskovania chýb v priestorovej oblasti porovnaním výslednej kvality videa medzi nimi. Ako kritérium výslednej kvality snímky používame odstup signálu od šumu (PSNR) [6]. Na uskutočnenie hodnoverného porovnania musia byť všetky metódy maskovania chýb realizované počas dekódovacieho procesu. To si vyžaduje použitie rovnakej postupnosti oboch inter- i intra-snímkov. Znamená to, že proces obnovy chýbajúceho makrobloku využíva iba informácie z predošlých dekódovaných makroblokov.



Obr. 6 Porovnanie priemerného PSNR medzi rozličnými metódami maskovania chýb v priestorovej oblasti

IV. ZHRNUTIE

Všetky uvedené metódy maskovania chýb v priestorovej oblasti boli založené na predpoklade, že štruktúra snímok videosignálu má hladký charakter. To znamená, že estimácia stratených pixelov dosiahne maximálnu hladkosť na hraniciach strateného makrobloku. Stratené pixely blízko neporušených hraníc môžu byť preto maskované lepšie ako tie, ktoré nie sú obklopené neporušenými hranicami. Menší počet susediacich makroblokov zase znamená menší počet zúčastnených pixelov, ktoré môžu byť použité na obnovu chýbajúceho pixelu porušeného makrobloku. Maskovanie chýb strateného makrobloku v snímke typu I smie byť realizované iba počas dekódovania snímky.

POĎAKOVANIE

Táto práca bola podporovaná Kultúrnou a edukačnou grantovou agentúrou Ministerstva školstva Slovenskej republiky VEGA na základe projektu č. 1/0045/10 (50%) a realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] LSI Logic Corporation, "H.264/MPEG-4 AVC Video Compression Tutorial", dostupný na internete : http://asicpfpga.com/site_upgrade/asicpfpga/pds/image_pds_files/H.264_MPEG4_Tutorial.pdf
- [2] O. Nemethova – A. Al-Moghrabi – M. Rupp, "Error concealment methods for video transmission over wireless networks", Diploma thesis. Wien: Institut für Nachrichtentechnik und Hochfrequenztechnik, Máj 2005 86s.
- [3] O. Nemethova – A. Al-Moghrabi – M. Rupp, "Flexible error concealment for H.264 based on directional interpolation", TU Wien, Institute of Communications and RF Engineering Gusshausstrasse 25/389, Wien, Austria, Jún 2005.
- [4] T. Wiegand – G. Sullivan, – G. Bjontegaard – A. Luthra, "Overview of the H.264/AVC Video Coding Standard." *IEEE Transactions on circuits and systems for video technology*, vol.13, no. 7, Júl 2003.
- [5] C. Weidmann–P. Kadlec–O. Nemethova–A. Al-Moghrabi, "Combined Sequential Decoding and Error Concealment of H.264", *Proceedings of IEEE MMSP'04 workshop*, 28 September – Október, Sienna, Italy.
- [6] G. Gennari–G.A. Mian –L. Celetto, "A H.264 Robust decoder for wireless environment", Department of Information Engineering, University of Padova, 35131 Padova Italy.

Pracovisko so vzdialeným prístupom pre testovanie návrhov na báze rekonfigurovateľných obvodov

Juraj TALÁN, Miloš DRUTAROVSKÝ, Ľudmila MACEKOVÁ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

juraj.talan@student.tuke.sk, milos.drutarovsky@tuke.sk, ludmila.macekova@tuke.sk

Abstrakt — Článok opisuje experimentálne meracie pracovisko, ktoré umožňuje vzdialený prístup k vývojovému kitu pre testovanie a ladenie návrhov na báze rekonfigurovateľných obvodov CPLD (Complex Programmable Logic Device) a FPGA (Field-Programmable Gate Array – programovateľné hradlové polia). Monitorovanie kitu pomocou web kamery a osciloskopu umožňuje praktické overenie funkčnosti návrhov pomocou reálnych záznamov a vizualizácie nameraných údajov.

Kľúčové slová — meracie pracovisko so vzdialeným prístupom, LabVIEW, FPGA, CPLD

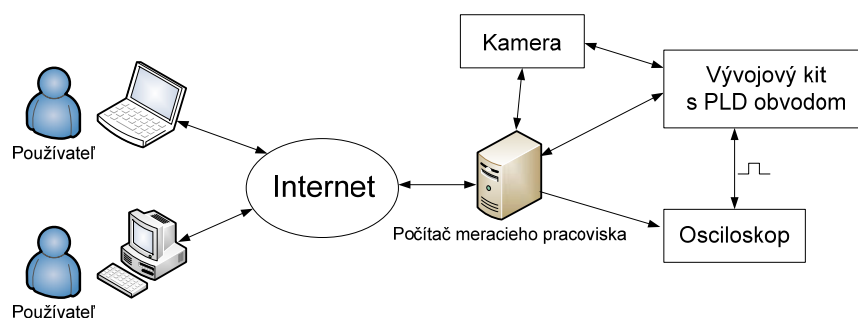
I. ÚVOD

Navrhnuté vzdialené pracovisko vzniklo modifikáciou pracoviska laboratória pre výučbu predmetu FPGA obvody [1]. Umožní pri výučbe predmetov v oblasti vložených systémov na báze programovateľných logických obvodov (PLD - Programmable Logic Device) s architektúrou CPLD (Complex Programmable Logic Device) a FPGA (Field-programmable Gate Array) praktické overenie návrhov pomocou reálneho cieľového hardvéru a vzdialeného využitia meracích prístrojov. Autorizovaní užívatelia pracoviska môžu v pohodlí domova, či internátu vytvoriť cieľovú aplikáciu, a po jej kompletnom alebo čiastočnom simulačnom overení pomocou CAE (Computer-Aided Engineering) nástrojov realizovať experimentálne overenie aj bez priameho kontaktu s vývojovým kitom. Efektívna práca so vzdialeným pracoviskom vyžaduje okrem vlastného počítača aj prístup k Internetu. Vzdialený prístup je realizovateľný prostredníctvom internetových prehliadačov Windows Internet Explorer alebo Mozilla Firefox a ďalších špeciálnych obslužných programov navrhnutých v prostredí LabVIEW. Pracovisko umožňujúce vzdialený prístup je vybavené vývojovými CPLD a FPGA kitmi firmy Altera. Jeho súčasťou je tiež digitálny pamäťový osciloskop, pomocou ktorého je možné snímať signály zodpovedajúce konkrétnemu návrhu testovaného vloženého systému. Perspektívne sa uvažuje aj s možnosťou tvorby komplexnejších zadanií v rámci takto zameraných výučbových predmetov.

V ďalších častiach tohto článku je uvedený stručný opis architektúry navrhnutého pracoviska so vzdialeným prístupom, nutná hardvérová a softvérová výbava a stručný opis experimentálnych výsledkov získaných pomocou vzdialeného pracoviska.

II. ARCHITEKTÚRA PRACOVISKA SO VZDIALENÝM PRÍSTUPOM

Architektúra pracoviska využívajúceho vzdialený internetový prístup k meracím prístrojom, vývojovému kitu a web-kamere je na obr.1. Vzdialený používateľ sa prostredníctvom svojho domáceho počítača a využitím internetového prehliadača môže pripojiť, resp. vyžiadať kontrolu nad počítačom meracieho pracoviska. K počítaču meracieho pracoviska je pripojená kamera a je vybavený softvérom pre ovládanie meracieho pracoviska s vývojovým kitom a osciloskopom. Pomocou kamery je možné sledovať napr. optickú signalizáciu stavov pomocou LED diód a podobne.



Obr. 1 Architektúra vzdialeného prístupu k vývojovému meraciemu pracovisku

III. HARDVÉROVÉ PROSTRIEDKY PRACOVISKA

Okrem osobného počítača vzdialeného užívateľa, ktorý musí byť vybavený operačným systémom MS Windows XP alebo vyšším, a počítača meracieho pracoviska so systémom MS Windows XP, sú súčasťou laboratória so vzdialeným prístupom nasledovné prístroje:

- Vývojový kit s CPLD alebo FPGA obvodom s napájacím adaptérom, príp. s káblom USB (Universal Serial Bus). Ako cieľové hardvérové platformy sú k dispozícii viaceré vývojové kity na báze CPLD a FPGA obvodov firmy Altera.
- Altera USB-Blaster [6], t.j. rozhranie pre zápis konfiguračného súboru z USB portu počítača meracieho pracoviska do konfigurovaného CPLD alebo FPGA obvodu. USB-Blaster je k riadiacemu počítaču pripojený pomocou univerzálneho USB kábla; k cieľovému vývojovému kitu je pripojený pomocou 10-vývodového konektora, ktorý je pripojený ku programovacím alebo konfiguračným vývodom CPLD, resp. FPGA obvodu. Toto rozhranie môže byť tiež použité pre JTAG (Join Test Action Group), AS (Active Serial) a PS (Passive Serial) módy konfigurácie.
- Osciloskop Tektronix TDS2024B pripojený k počítaču meracieho pracoviska pomocou USB kábla. Tektronix TDS2024B je moderný štvorkanálový osciloskop s farebným LCD displejom. Umožňuje využiť šírku pásma do 200MHz pri vzorkovacej frekvencii 2GS/s.
- Web-kamera Logitech QuickCam [7] pripojená k počítaču meracieho pracoviska pomocou USB kábla. Logitech QuickCam Messenger je bežná web kamera, postačujúca pre overenie základnej funkčnosti návrhu, napr. pre prípad rozblíkavania LED diódy a pod.

Uvedené prístroje je možné využiť pri ožiovaní a testovaní návrhov na báze CPLD a FPGA v rámci študentských zadaní ako aj v rámci rozsiahlejších projektov, čo je jedným z hlavných cieľov pracoviska v budúcnosti. Typ a počet pripojených meracích prístrojov je ľahko rozšíriteľný a v budúcnosti bude vybavenie pracoviska prispôbované testovaným cieľovým návrhom.

IV. SOFTVÉROVÉ PROSTRIEDKY PRACOVISKA

A. Softvérová výbava počítača vzdialeného užívateľa

Pre spoľahlivú prácu pri vzdialenom prístupe k meraciemu pracovisku musí mať užívateľ na svojom počítači nainštalované nasledujúce softvérové prostriedky:

- internetový prehliadač: Windows Internet Explorer alebo Mozilla Firefox,
- vývojový nástroj od firmy Altera Quartus II [2], pomocou ktorého je realizovaný návrh cieľového systému a generovanie vhodných konfiguračných súborov .jbc (alebo .jam),
- program vytvorený v LabVIEW [3], pomocou ktorého sa odošle konfiguračný súbor do riadiaceho počítača meracieho pracoviska.

B. Softvérová výbava počítača meracieho pracoviska

Počítač meracieho pracoviska je vybavený nasledujúcim softvérom:

- programom vytvoreným v LabVIEW, pomocou ktorého sa realizuje ladenie a overovanie náročných konštrukcií na báze CPLD a FPGA obvodov,
- programom Quartus II Programmer, pomocou ktorého sa realizuje konfigurácia testovaného CPLD alebo FPGA obvodu,
- programom pre vytvorenie užívateľského konta pre ftp-prístup
- programom vytvoreným v LabVIEW pre stiahnutie vygenerovaných súborov (napr. záznamov z meracích prístrojov) ako aj zobrazenie obsahu snímaného pomocou web-kamery.

V. OPIS PRÁCE PRI VZDIALENOM PRÍSTUPE K MERACIEMU PRACOVISKU

S využitím prostriedkov a prístrojov nainštalovaných resp. pripojených k jednotlivým počítačom si môže užívateľ na svojom domácom počítači vytvoriť návrh pomocou vývojového nástroja Quartus II, a následne aj konfiguračný súbor, ktorý pomocou programu vytvoreného v LabVIEW odošle a uloží na počítači meracieho pracoviska (PMP). Ďalej pomocou Internetu riadi PMP. Pomocou Quartus II Programmer-a nainštalovaného na PMP užívateľ nakonfiguruje vývojový kit s CPLD alebo FPGA obvodom. Niektoré funkcionality nového návrhu (blikanie kontrolnej LED a pod.) môže overiť aj pomocou pripojenej web-kamery. Pomocou osciloskopu je možné snímať signály z vývojového kitu a priebeh signálov uložiť do súboru, ktorý je následne možné stiahnuť do vzdialeného užívateľovho počítača (PC). Na svojom PC môže užívateľ spracovať a vykresliť namerané výsledky pomocou napr. MS-Excel-u a pod., podľa konkrétnych požiadaviek zadania.

VI. SPÔSOB KONFIGURÁCIE A OVLÁDANIA VZDIALENÉHO PRACOVISKA

A. Altera Quartus II

Prostredie Quartus II [2] je EDA (Electronic Design Automation) nástroj firmy Altera, ktorý umožňuje návrh a syntézu a simulovanie číslicových systémov s PLD obvody (FPGA a CPLD) firmy Altera. Návrh je možné vytvoriť dvoma spôsobmi: grafickým („kreslenie“ schémy), alebo textovým s využitím VHDL kódu [8] alebo Verilog jazykov pre opis hardvéru.

Po vytvorení návrhu, kompilácii a následnom úspešnom odsimulovaní je možné konfigurovať PLD obvod pomocou programovacieho nástroja Quartus II Programmer. Quartus II Programmer je súčasťou prostredia Quartus II, ale je možné ho doinštalovať aj samostatne. Keďže vývojový kit s cieľovým PLD obvodom je pripojený k počítaču meracieho pracoviska, ku ktorému vzdialený užívateľ nemá fyzický prístup, konfigurácia PLD obvodu je riešená s využitím špecializovaných softvérových nástrojov vytvorených v programovacom prostredí LabVIEW.

Súbor Jam

Nástroj Quartus II umožňuje vygenerovanie viacerých typov konfiguračných súborov [5] (napr. .pof pre CPLD, .sof pre FPGA), v tomto prípade je pre nás výhodné použitie .jbc, resp. .jam súborov.

Súbory Jam obsahujú programovacie dáta a informácie o použítom algoritme potrebné pre naprogramovanie PLD obvodu pomocou JTAG (Join Test Action Group) rozhrania. Altera podporuje dve rôzne implementácie Jam súboru: súbor Jam Byte-Code (.jbc) a súbor ASCII (American Standard Code for Information Interchange) Jam (.jam). Súbor .jbc je binárny a súbor .jam je textový. Súbor .jbc je výhodnejší vďaka menšej veľkosti a rýchlejšiemu programovaniu obvodu a bol preto využitý v navrhnutom pracovisku.

Quartus II Programmer

Je súčasťou nástroja Quartus II [2]. Po úspešnej kompilácii a vygenerovaní konfiguračných súborov umožňuje konfiguráciu alebo naprogramovanie CPLD a FPGA obvodov. Hlavnou časťou, ktorá bola využitá z Quartus II Programmer-a, bol program quartus_jli. Ten možno ovládať pomocou príkazového riadka, a podporuje všetky spôsoby konfigurácie obvodov firmy Altera [29] pomocou napríklad: ByteBlaster, ByteBlasterII a USB-Blaster, ktorý je využívaný v prípade navrhnutého pracoviska

B. Programovacie prostredie LabVIEW

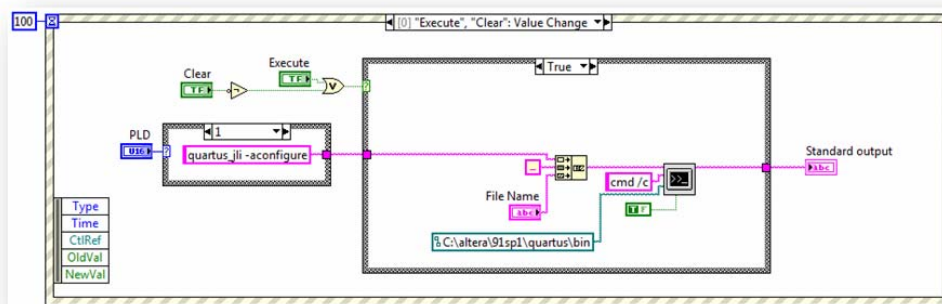
Prostredie LabVIEW [3] možno využiť pre vzdialený prístup k osciloskopu, tiež pre odoslanie konfiguračných súborov pre vývojový kit s PLD obvodom, pre zobrazenie nameraných hodnôt, priebehov signálov, ako aj pre uloženie nameraných dát do počítača meracieho pracoviska.

LabVIEW je grafické programovacie prostredie, ktoré umožňuje relatívne rýchly vývoj testovacích a meracích aplikácií, ako aj aplikácií na prezentovanie nameraných výsledkov. Je založené na množine programov v tvare ikon (Virtual Instruments – VI – virtuálne nástroje), ktoré sú poprepájané pomocou virtuálnych vodičov (wires).

Program s príponou .vi vytvorený v LabVIEW je tiež virtuálnym nástrojom. Jeho návrh prebieha v 2 grafických oknách, ktoré simulujú čelný panel prístroja (Front Panel) a jeho funkčnú schému (Block Diagram).

Front Panel (napr. na obr.4) slúži na komunikáciu s užívateľom a obsahuje vstupné tlačidlá, ovládače (Controls) a výstupné objekty ako grafy, LED diódy a pod.

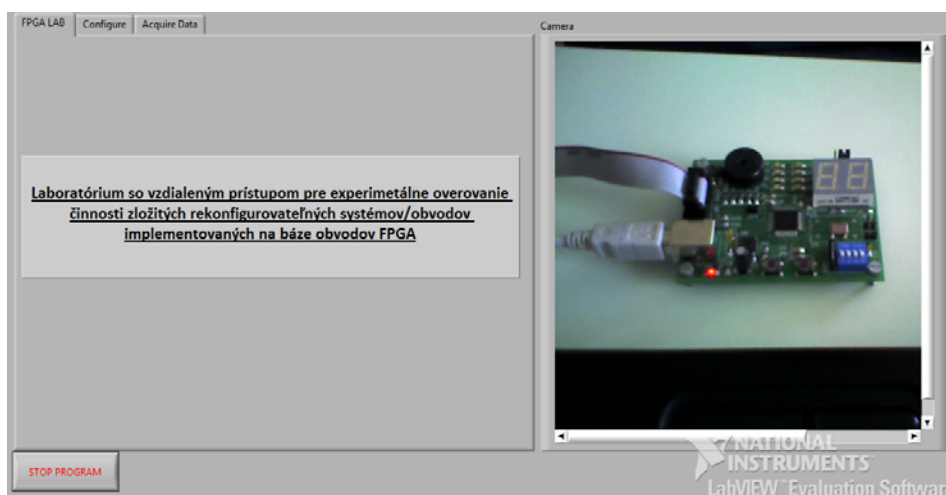
Block Diagram obsahuje samotný kód, ktorý pozostáva z rôznych podprogramov (SubVI) alebo preddefinovaných VI, ktoré sú pospájané pomocou vodičov. Príklad grafického kódu navrhnutého v LabVIEW je na Obr. 2.



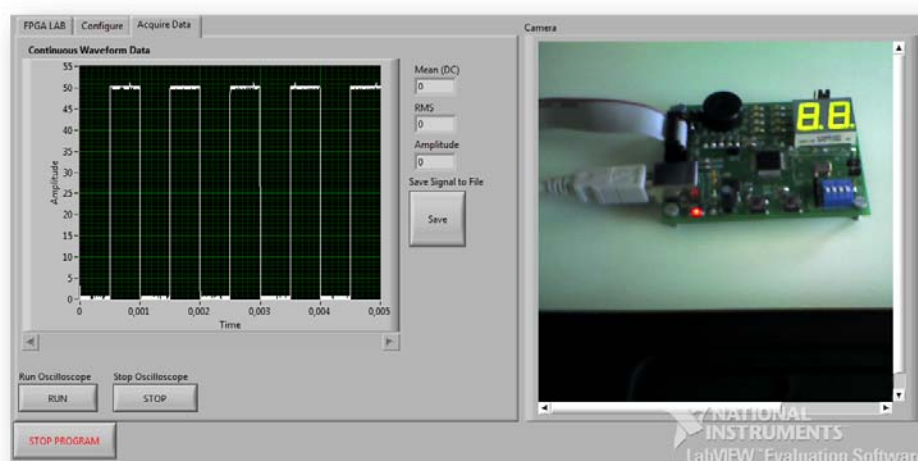
Obr. 2 LabVIEW, Block Diagram – funkčná schéma jednoduchého virtuálneho nástroja

Program pre odoslanie konfiguračného súboru

Odoslanie konfiguračného súboru vykonáva nástroj (VI) vytvorený v prostredí LabVIEW. Spomínaný VI musí mať každý užívateľ nainštalovaný vo svojom PC. Použije ho po vygenerovaní konfiguračného súboru v nástroji Quartus II na svojom počítači, a následne, pomocou FTP protokolu, súbor pošle do požadovaného priečinka na počítači meracieho pracoviska, do ktorého má prístup s automatickým overením mena a hesla.



Obr. 3 Vizualizácia vzdialeného pracoviska pomocou web-kamery zobrazenej v rámci virtuálneho prístroja



Obr. 4 LabVIEW, Front Panel - Čelný panel virtuálneho prístroja pre vzdialené osciloskopické meranie

V prostredí LabVIEW bolo navrhnuté aj vzdialené ovládanie meracieho prístroja (osciloskopu) a web-kamery cez internetovú sieť, ako vidno na obr. 3 a obr. 4. S výhodou boli využité podporné nástroje firmy National Instruments, dostupné na jej stránkach pre konfiguráciu web servera na báze prehliadača Windows Internet Explorer alebo Mozilla Firefox.

VII. ZÁVER

Opisované pracovisko umožňuje vzdialené testovanie návrhov na báze Altera CPLD a FPGA obvodov s využitím reálneho hardvéru a meracích prístrojov. Užívateľ po návrhu a odsimulovaní návrhu pomocou CAE nástrojov na svojom počítači zrealizuje vzdialenú konfiguráciu CPLD alebo FPGA obvodu v testovanom kite a môže zosnímať priebehy meraných signálov. Pomocou web-kamery môže vizuálne kontrolovať činnosť signalizačných prvkov kitov. Ďalší vývoj pracoviska bude zameraný na rozšírenie podpory pripojených meracích prístrojov a dodanie podpory pre efektívne pridelovanie strojového času rôznym užívateľom.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja, (50%).



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055, (50%).

LITERATÚRA

- [1] Stránka laboratória FPGA Virtual Lab. Dostupné na <http://www.kemt.feit.tuke.sk/fpga/kega/index.html>
- [2] Altera Quartus II. Dostupné na <http://www.altera.com/products/software/quartus-ii/subscription-edition/qts-se-index.html>
- [3] Labview. Dostupné na <http://www.ni.com/labview/>
- [4] Mentor Graphics. Dostupné na <http://www.mentor.com/>
- [5] JAM súbory. Dostupné na ftp://ftp.altera.com/pub/lit_req/document/an/an088.pdf
- [6] USB-Blaster. Dostupné na http://www.altera.com/literature/ug/ug_usb_blstr.pdf
- [7] Logitech QuickCam. Dostupné na <http://www.logitech.com/en-us/home/>
- [8] FISHER, Viktor: VHDL Language - Introduction to the design and synthesis of digital systems, 2008. Dostupné na http://www.kemt.feit.tuke.sk/Predmety/KEMT436_FPGA/_materialy/prednasky/VHDL_Fischer.pdf

Klasifikácia metód digitálnej vodotlače

*Peter GOČ – MATIS, Tomáš KANÓCZ,
Patrik GALLO, Zita KLENOVIČOVÁ, Dušan LEVICKÝ*

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

peter.goc-matis@tuke.sk, tomas.kanocz@tuke.sk, patrik.gallo@tuke.sk,
zita.klenovicova@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Tento článok stručne popisuje techniku digitálnej vodotlače a taktiež sa zaoberá klasifikáciou metód digitálnej vodotlače z dvoch rôznych pohľadov, a to na základe charakteristík digitálnych vodoznakov a na základe ich použitia v jednotlivých aplikáciách.

Kľúčové slová — autorské práva, digitálna vodotlač, frekvenčná oblasť, identifikačné vodoznaky, perceptívna transparentnosť, robustné vodoznaky, útoky

I. ÚVOD

Dnešná doba prináša so sebou masívne rozšírenie spracovania a prenosu multimediálnych dát, ktoré sú zaznamenávané prevažne v digitálnej forme. Digitálna forma dát umožňuje jednoduchý záznam, spracovanie a prenos multimediálnych dát cez komunikačné prostredie. Na druhej strane, ale stoja nevýhody digitálnej formy dát, a to hlavne možnosť jednoduchého kopírovania týchto dát bez pozorovateľnej zmeny kvality vytvorenej kópie. Takto vytvorené kópie môžu byť vďaka vysokým prenosovým rýchlostiam súčasných sietí ilegálne distribuované do celého sveta.

Ochranu multimediálnych dát je možné zabezpečiť použitím techník digitálnej vodotlače.

II. PRINCÍP DIGITÁLNEJ VODOTLAČE

Metódy digitálnej vodotlače spočívajú vo vložení prídavnej informácie (vodoznaku) do multimediálnych dát. Dôležitou vlastnosťou vkladanej informácie je jej zmyslová nepostrehnuteľnosť a odolnosť voči potenciálnemu poškodeniu.

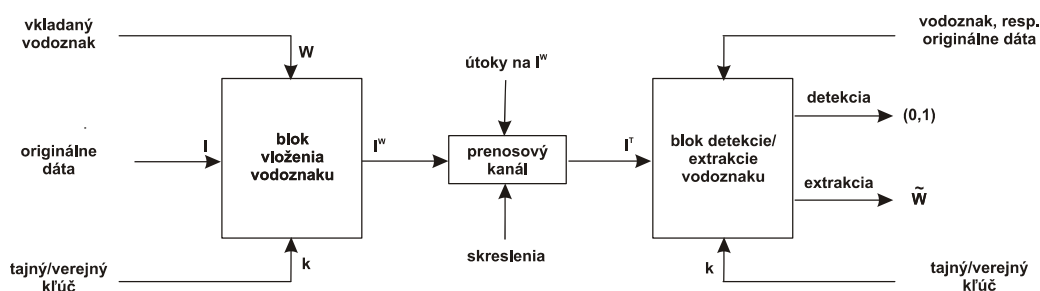
Prídavná informácia v sebe môže niesť údaje o vlastníčkovi autorských práv, resp. o distribútorovi a taktiež môže plniť funkciu zabezpečenia integrity dát.

Vodoznak môže mať formu konečnej postupnosti symbolov, resp. číslíc, prídavnej obrazovej informácie (loga), segmentu rečového signálu, poprípade môže mať formu 1-bitovej informácie, ktorá indikuje, či originálne dáta sú alebo nie sú označené vodoznakom.

A. Blokový model digitálnej vodotlače

Blokový model digitálnej vodotlače je znázornený na Obr. 1. Tento model je možné rozdeliť na dva základné bloky, a to [1]:

- o blok vloženia vodoznaku,
- o blok detekcie, resp. extrakcie vodoznaku.



Obr. 1 Blokový model a princíp digitálnej vodotlače

B. Požiadavky kladené na digitálne vodoznaky

Metódy digitálnej vodotlače musia v závislosti od svojho určenia spĺňať päť základných požiadaviek [2]:

- o robustnosť,
- o perceptívna transparentnosť,
- o nedetekovateľnosť,
- o bezpečnosť,
- o kapacita.

C. Útoky na digitálne vodoznaky

Útokom rozumieme akúkoľvek cieľenú operáciu, ktorú vykonáva nepovolaná osoba (útočník) s označenými dátami, cieľom ktorej je buď úplne odstrániť vodoznak s označených dát, resp. poškodiť ho natoľko, aby sa znemožnila jeho detekcia popřípade extrakcia. Podľa [1][3] možno útoky na vodoznaky zdeliť do štyroch nasledujúcich skupín:

- o útoky odstraňujúce vodoznak,
- o geometrické útoky,
- o kryptografické útoky,
- o protokolové útoky.

III. KLASIFIKÁCIA METÓD DIGITÁLNEJ VODOTLAČE NA ZÁKLADE CHARAKTERISTÍK

Podľa charakteristík môžeme vkladané vodoznaky rozdeliť do piatich skupín, a to na základe [4]:

- o závislosti od originálnych dát,
- o postrehnuteľnosti v označených dátach,
- o vzťahu k označeným dátam,
- o robustnosti,
- o oblasti vkladania.

A. Delenie na základe závislosti od originálnych dát

Techniky digitálnej vodotlače môžeme nazvať **nezávislé** (*blind*) ak v procese extrakcie nie je nutná prítomnosť originálnych neoznačených dát I .

Na rozdiel od toho, techniky digitálnej vodotlače môžeme nazvať **závislé** (*nonblind*) ak v procese extrakcie vodoznaku je nutná prítomnosť originálnych dát I .

Po extrakcii vodoznaku následne prebieha porovnávanie extrahovaného vodoznaku $\tilde{W}$ a originálneho vloženého vodoznaku W , pričom najčastejšou mierou podobnosti je ich vzájomná korelácia.

Vo všeobecnosti sú závislé techniky pre vkladanie vodoznakov robustnejšie ako nezávislé techniky, pretože je zrejmé že vodoznak môže byť jednoduchšie extrahovaný s využitím originálnych neoznačených dát. Vo viacerých aplikáciách však neoznačené originálne dáta nie sú k dispozícii v procese detekcie, resp. extrakcie, a tak sú **nezávislé** techniky užitočnejšie práve kvôli tejto absencii nutnosti prítomnosti originálnych dát.

B. Delenie na základe postrehnuteľnosti v označených dátach

Vodoznak môžeme označiť ako **postrehnuteľný** (*perceptible*) ak je zámerom aby tento vložený vodoznak bol viditeľný v označených dátach, napr. logo vložené do rohu obrazu. Tieto vodoznaky si môžeme predstaviť ako analógiu k vodoznakom na peňažných bankovkách. Viditeľné vodoznaky vkladané do obrazov slúžia na zamedzenie použitia týchto dát pre komerčné účely. Tento typ vodoznakov musí spĺňať dve presne definované podmienky. Prvou je, že pre osobu, ktorá nevlastní autorské práva by malo byť technicky náročné odstrániť vložený vodoznak z označených dát. Druhou podmienkou je odolnosť voči ich falšovaniu.

Na druhej strane **nepostrehnuteľný** (*imperceptible*) vodoznak je vložený do multimediálnych dát bez toho aby došlo k podstatnému narušeniu informačného obsahu, ktorý je bežne vnímateľný ľudským okom. Táto vlastnosť sa označuje ako **perceptívna transparentnosť** (*perceptual transparency*). Skryté vodoznaky sa do originálnych dát vkladajú za použitia kľúča, ktorý je potrebný aj pri ich extrakcii. Takto označené dáta sa môžu zverejniť, pričom pri kopírovaní je každá kópia označená vodoznakom. Pri spore o legálnosti ktorejkoľvek kópie, iba povolaná osoba vie na základe kľúča extrahovať vodoznak a takýmto spôsobom potvrdiť svoje autorské práva k daným dátam.

C. Delenie na základe vzťahu k označeným dátam

Vodoznak môžeme nazývať **súkromný** (*private*), ak iba autorizovaná osoba je schopná detekovať tento vodoznak v multimediálnych dátach. Inak povedané, privátne techniky digitálnej vodotlače vynakladajú všetko úsilie za účelom znemožnenia extrakcie vodoznaku z označených dát neautorizovanou osobou, a to napríklad použitím privátneho (súkromného) pseudonáhodného kľúča. Tento privátny kľúč popisuje oblasť pre vloženie digitálneho vodoznaku do originálnych dát a taktiež aj pre inverzný proces jeho extrakcie. Podstatou súkromných vodoznakov je teda ich využitie pre možnosti autentizácie a ochrany autorských práv.

Na druhej strane **verejný** (*public*) techniky digitálnej vodotlače povoľujú ľubovoľnej osobe čítať vodoznak. Sú vkladané do známych oblastí takže ich detekcia, resp. extrakcia môže byť jednoducho vykonaná. Verejné vodoznaky slúžia teda na prenos informácie v originálnych dátach. Z toho vyplýva, že nejde o vodoznak ako taký, pretože kľúč použitý na vloženie prídavnej informácie je verejný, a teda prístupný širokej verejnosti.

Vo všeobecnosti sú teda privátne techniky digitálnej vodotlače robustnejšie ako verejné techniky, v prípade ktorých môže útočník pomerne jednoducho odstrániť alebo poškodiť vložený vodoznak.

Poznáme taktiež asymetrickú formu verejných techník digitálnej vodotlače, ktorá je špecifická tým, že každý používateľ môže vodoznak čítať ale nie je schopný odstrániť ho z označených dát. Táto forma je označovaná ako **asymetrický kryptosystém** (*asymmetric cryptosystem*). V tomto prípade, je použitá dvojica kľúčov:

- o *privátny kľúč - pre vloženie,*
- o *verejný kľúč - pre verifikáciu.*

D. Delenie na základe robustnosti

Jednou zo základných požiadaviek na digitálne vodoznaky je robustnosť, ktorá vyjadruje schopnosť vodoznaku odolávať úmyselným aj neúmyselným útokom. Úmyselné útoky majú za cieľ zámerne poškodiť alebo úplne odstrániť vodoznak vložený do multimediálnych dát. Na druhej strane neúmyselné útoky nie sú vykonávané za účelom zámernej modifikácie, resp. odstránenia ukrytej informácie. Podľa účelu vkladania, môžeme vodoznaky rozdeliť do troch skupín:

- o robustné,
- o polokrehké,
- o krehké.

Robustné vodoznaky (*robust watermarks*) sú navrhnuté tak aby odolali úmyselným a neúmyselným modifikáciám označených dát [5]. K úmyselným modifikáciám patria odstránenie alebo pozmenenie vloženého vodoznaku a následné neautorizované vloženie iného vodoznaku do týchto upravených dát. K neúmyselným útokom a modifikáciám patria najmä operácie spracovania obrazu, medzi ktoré patria zmena priestorovej rozlišovacej schopnosti, orezanie, filtrovanie a stratová kompresia. Robustné vodoznaky sú využívané prevažne pre účely ochrany autorských práv, a teda dôkazu oprávneného vlastníctva multimediálneho obsahu.

Polokrehké vodoznaky (*semifragile watermarks*) sú navrhnuté tak aby dokázali zistiť každú neautorizovanú modifikáciu, s prihliadnutím na určité operácie spracovania obrazu [6]. Sú používané pre selektívnu autorizáciu, ktorá je schopná detekovať neoprávnené narušenie dát s prihliadnutím na výskyt bežného skreslenia. Inými slovami polokrehké vodoznaky sú schopné rozlíšiť bežný šum a skreslenia produkované metódami spracovania obrazu, ako sú stratové kompresie, od šumu a skreslenia produkovaných úmyselnými modifikáciami.

Krehké vodoznaky (*fragile watermarks*) sú svojou zámerne zníženou robustnosťou prispôbené, tomu aby dokázali detekovať akúkoľvek neautorizovanú modifikáciu dát pre účely autentizácie [7]. Požadovanou vlastnosťou u tohto typu vodoznakov je teda schopnosť identifikovať časti zdrojových dát, ktoré boli narušené. Techniky digitálnej vodotlače s krehkými vodoznakmi sú schopné kompletnej kontroly integrity dát. Podstatou je že aj tá najmenšia modifikácia obrazových dát spôsobí pozmenenie alebo zničenie vloženého krehkého vodoznaku.

E. Delenie na základe oblastí vkladania

Pre účely vkladania vodoznakov môžeme využiť dve typické oblasti, a to priestorovú oblasť alebo frekvenčnú oblasť.

V **priestorovej oblasti** (*spatial domain*) môžeme jednoducho koeficientami vodoznaku modifikovať koeficienty obrazu na vybraných pozíciách [8]. Táto skupina metód digitálnej vodotlače patrí k úplne prvým, ktoré začali realizovať vkladanie vodoznaku do originálnych dát, a to buď modifikáciou časových vzoriek alebo obrazových prvkov zdrojových dát. Medzi najčastejšie používané metódy tejto skupiny patria metódy založené na modifikácii bitových rovín a jasových úrovní statického obrazu.

Výhodami týchto metód sú nízka výpočtová náročnosť a jednoduchá implementácia, ale na druhej strane vložená informácia je jednoducho detekovateľná použitím metód štatistickej analýzy a taktiež sú tieto techniky málo robustné voči útokom stratovej kompresie.

Vo **frekvenčnej oblasti** (*frequency domain*) môžeme vodoznak vkladať modifikovaním transformačných (spektrálnych) koeficientov, ktoré získame po aplikovaní niektorej z diskretných ortogonálnych transformácií [9]. Najčastejšie využívanými transformáciami sú DCT, DFT, DWT a DST. Výhodou týchto metód na vkladanie vodoznaku je, že do originálnych dát môžeme vložiť väčšie množstvo prídavnej informácie ako je to pri metódach vkladania v priestorovej oblasti, a to pri súčasnom zachovaní dostatočnej úrovne transparentnosti. Vloženie väčšieho množstva informácie má zase vplyv na zvýšenie robustnosti tejto informácie. Hlavnými nevýhodami týchto metód je hlavne vyššia výpočtová náročnosť ako pri metódach vkladania v priestorovej oblasti a taktiež, že pri vložení väčšieho množstva informácie značne degradujeme kvalitu obrazu.

IV. KLASIFIKÁCIA METÓD DIGITÁLNEJ VODOTLAČE NA ZÁKLADE POUŽITIA V APLIKÁCIÁCH

Techniky digitálnej vodotlače vkladajú skrytú informáciu priamo do multimediálnych dát. Okrem kryptografických schém, predstavuje vodotlač účinnú technológiu pre zabezpečenie integrity dát ako aj overenie pôvodu dát, autorských práv a autorizovaného príjemcu. Z pohľadu použitia v aplikáciách môžeme vodoznaky rozdeliť do týchto piatich skupín [4]:

- o *vodoznaky zabezpečujúce ochranu autorských práv,*
- o *vodoznaky zabezpečujúce autorizáciu dát,*
- o *identifikačné vodoznaky,*
- o *vodoznaky zabezpečujúce ochranu pred kopírovaním,*
- o *vodoznaky využívajúce overovacie zariadenia.*

A. Vodoznaky zabezpečujúce ochranu autorských práv

Asymetrické šifrovacie algoritmy ako napr. RSA [10] neposkytujú dostatočnú úroveň zabezpečenia vytvárania neautorizovaných kópií multimediálnych dát. Všetky šifrované dáta je totiž nutné pred ich používaním dešifrovať, a tak sa stávajú čitateľnými a prinášajú možnosť pre jednoduché kopírovanie a šírenie nelegálnych kópií. Myšlienka vloženie neviditeľnej prídavnej informácie, ktorá slúži ako identifikácia vlastníctva a príjemcu sa stáva atraktívnou pre mnoho vydavateľstiev a spoločností zaoberajúcich sa distribúciou multimediálnych dát.

Princíp spočíva vo vložení transparentného vodoznaku do multimediálnych dát, pričom extrakcia tohto vodoznaku prebieha porovnaním označených dát s originálnymi. Vodoznaky zabezpečujúce ochranu autorských práv sú navrhované tak aby dokázali identifikovať zdroj multimediálnych dát ako aj ich autorizovaného používateľa.

Jedna z možností zabezpečenia obsahu a ochrany autorských práv je vloženie vodoznaku do toku videa. Tento vodoznak nesie informáciu o odosielateľovi (distribútor) ako aj o príjemcovi (konečný zákazník) spomínaného digitálneho obsahu. Týmto spôsobom dokáže vložený digitálny vodoznak vystopovať rôzne kópie digitálneho videa. Táto metóda môže byť aplikovaná vo video distribúcii prostredníctvom Internetu, televízneho vysielania a taktiež pri označovaní videodiskov.

B. Vodoznaky zabezpečujúce autorizáciu dát

Tradičný význam zabezpečenia pravosti pôvodu dát (autorizácie dát) je v spojení s vlastnoručne podpísanými dokumentmi. Keďže podpis je uvedený na určitom dokumente je ťažké tento podpis modifikovať alebo presunúť na iný dokument. Porovnaním dvoch vlastnoručných podpisov je možné posúdiť, či patria jednej a tej istej osobe.

Digitálne podpisy môžeme vnímať ako analógiu k vlastnoručným podpisom. Digitálne

podpisy nám však ponúkajú oveľa silnejší stupeň autorizácie ako je to u vlastnoručných podpisov. Využívajú princípy asymetrickej kryptografie, a teda prítomnosti dvoch kľúčov, ktoré sú matematicky závislé. Používateľ podpíše vytvorený digitálny dokument zašifrovaním pomocou svojho súkromného kľúča. Dešifrovanie je možné iba verejným kľúčom používateľa, ktorý dokument podpísal. Tento verejný kľúč je verejne dostupný ľubovoľnej osobe a v prípade, že je ním možné dešifrovať podpísaný dokument, dokazuje to, že bol skutočne vytvorený a podpísaný spomínaným používateľom.

Vysoký stupeň ochrany spočíva vo fakte, že je veľmi obtiažné zo známeho verejného kľúča vypočítať súkromný kľúč.

Redundancia dát v obraze vytvára schopnosť verifikácie vloženého vodoznaku bez nutnosti prítomnosti originálnych dát. Aj keď konverzia medzi rôznymi obrazovými formátmi vedie k rôznej reprezentácii, nezmení nám vizuálny vzhľad ani overovací stav. Avšak overený obraz bude nevyhnutne skreslený malým množstvom šumu, ktorý nie je prípustný v presných lekárskejších obrazoch alebo u strategicky dôležitých obrazov vo vojenských aplikáciách.

C. Identifikačné vodoznaky

Biometrické metódy ako rozpoznávanie tváří, odtlačkov prstov alebo rozpoznávanie očných dúhoviek zohrávajú veľmi dôležitú úlohu v súčasnosti vyvíjaných a používaných systémov pre identifikáciu jednotlivcov. Digitálna vodotlač multimediálnych dát označených biometrickými odtlačkami môže byť aplikovaná s cieľom zabezpečenia týchto dát voči zlomyseľným útokom. Taktiež je schopná detekcie podvodne vytvorených kópií ako aj zabezpečenia bezpečného prenosu.

Identifikačné vodoznaky sú pre potreby digitálnej vodotlače zvyčajne uvažované ako vloženie informácie o totožnosti do multimediálnych dát takým spôsobom aby tento vodoznak bolo zložité odstrániť. V prípade, že sa nelegálna kópia multimediálnych dát dostane na čierny trh, podľa jedinečného vodoznaku je možné zistiť kto je porušiteľom autorských práv a následne vyvodíť dôsledky. Ide teda o jednoznačnú identifikáciu distribútora

Ak je vodoznak použitý touto metódou informáciu, ktorú nesie nazývame aj **payload**. Vodoznak teda nesie plnohodnotnú informáciu o vlastníkovi obsahu multimediálnych dát alebo povolenia pre jeho používanie. Forma takéhoto vodoznaku je identifikačné číslo, ktoré je unikátne pre každého z legálnych príjemcov multimediálneho obsahu.

D. Vodoznaky zabezpečujúce ochranu pred kopírovaním

V roku 1996 výskumné laboratórium firmy IBM v Tokiu ako prvé navrhlo použitie technológie digitálnej vodotlače pre ochranu DVD diskov pred kopírovaním a ukázalo sa, že navrhovaná technológia je dokonca schopná detekovať vložený vodoznak v dátovom toku kompresie MPEG-2. Táto technológia bola založená na vložení prídavnej informácie o ochrane proti kopírovaniu do video obsahu. Vložená informácia bola následne schopná automaticky zabrániť zariadeniam prehrávanie neoprávnených DVD kópií alebo vytváranie ďalších neoprávnených kópií. Digitálne dáta sú vysielané prístrojom zo strany vysielateľa k prijímaču cez rozhrania, ktoré povoľujú odosielanie a prijímanie digitálneho obsahu iba medzi oprávnenými prístrojmi. Dáta uložené na disku sa skladajú z dvoch častí. Prvá časť obsahuje obraz a zvuk a druhá časť dát obsahuje informáciu o ochrane proti kopírovaniu, a teda vkladajú vodoznak.

E. Vodoznaky využívajúce overovacie zariadenia

Táto skupina vodoznakov je používaná pre kontrolu prístupu k prostriedkom, za použitia overovacích zariadení [11]. Podstatou je vloženie autorizačného kódu do signálu a tento signál (napr. televízny alebo rádio) je následne vysielaný do overovacích zariadení. V overovacom zariadení je autorizačný kód extrahovaný z označeného signálu a na základe povolení, ktoré tento kód nesie sú potom vykonávané povolené operácie a prístup k prostriedkom.

Vodoznaky tejto skupiny môžu byť taktiež využívané pri vzdialenom ovládaní zariadení ako sú napríklad počítače, hračky a spotrebiče. Podstata spočíva vo vložení vodoznaku do audio signálu. Prijímacie zariadenie je vybavené detektorom, ktorý je schopný identifikovať skrytý signál a na základe neho spustiť preddefinovanú operáciu, resp. zmeniť status zariadenia. Táto skupina vodoznakov je taktiež využívaná v zariadeniach s nastaveným časovým obmedzením, ktoré si po detekcii vodoznaku nastaví určitý preddefinovaný interval, počas ktorého je používateľ oprávnený vykonávať operácie na danom zariadení.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] D. Levický, "Multimediálne telekomunikácie. 1.vyd. Košice : ELFA s.r.o., 2002. 240 s. ISBN 80-89066-58-5.
- [2] M. WU, B. Liu, "Multimedia data hiding", Springer-verlag New York, Inc. 2003.
- [3] R. Ridzoň, "Útoky na digitálne vodoznaky v statických obrazoch. Diplomová práca, FEI TU Košice, 2003.
- [4] F. Y. Shih, "Digital Watermarking and Steganography: Fundamentals and Techniques, CRC Press, 2007. 200 p. ISBN 978-1-4200-4757-8.
- [5] N. Nikolaidis, S. Pitas, "Robust Robust Image Watermarking in th Spatial Domain. Signal Processing 66 (1998): 385.
- [6] Q. Sun, S. F. Chang, "Semifragile Image Authentication Using Generic Wavelet Domain Features and ECC. In Proc IEEE Int. Conf. on Image Processing. 2002.
- [7] M. U. Celik, "Hierarchical Watermaking for Secure Image Authentication with Localization. IEEE Trans. Image Processing 11 (2002): 585.
- [8] H. Berghell, L. O'Gorman, " Protecting Ownership Rights through Digital Watermarking. IEEE Computer Mag. 29 (1996): 101.
- [9] I. J. Cox, "Secure Spread Spectrum Watermarking for Multimedia. IEEE Trans. Inage Processing 6 (1997): 1637.
- [10] W. G. Chambers, "Basics of Communications and Coding. Oxford Science Publications. Oxford: Clarendon Press, 1985, ISBN 978-0-1985-3195-1.
- [11] S. E. Balci, "Robust Watermarking of Images. M.Sc thesis, Ankara, Turkey, 2003.

SúčasnÉ metódy digitálnej vodotlače vo videu v reálnom čase

Peter GOČ - MATIS, Tomáš KANÓCZ, Patrik GALLO, Dušan LEVICKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

peter.goc-matis@tuke.sk, tomas.kanocz@tuke.sk,
patrik.gallo@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Tento článok stručne popisuje princíp techniky digitálnych vodoznakov ako aj súčasné metódy digitálnej vodotlače vo videu s prihliadnutím na potreby použitia pre video v reálnom čase.

Kľúčové slová — digitálna vodotlač vo videu, neúmyselné útoky, singulárny rozklad hodnôt, úmyselné útoky, video v reálnom čase, vlc kód

I. ÚVOD

Dnešná doba prináša so sebou masívne rozšírenie spracovania a prenosu multimediálnych dát, ktoré sú zaznamenávané prevažne v digitálnej forme. Digitálna forma dát umožňuje jednoduchý záznam, spracovanie a prenos multimediálnych dát cez komunikačný prostredie. Na druhej strane, ale stoja nevýhody digitálnej formy dát, a to hlavne možnosť jednoduchého kopírovania týchto dát bez pozorovateľnej zmeny kvality vytvorenej kópie. Takto vytvorené kópie môžu byť vďaka vysokým prenosovým rýchlostiam súčasných sietí nelegálne distribuované prakticky do celého sveta. Prístupy pre ochranu multimediálnych dát môžeme rozdeliť do dvoch základných skupín [1][2]:

- o ochrana multimediálneho obsahu počas prenosu,
- o ochrana multimediálneho obsahu po prenose.

Prístupy pre ochranu multimediálnych dát počas prenosu sú v prevažnej miere zabezpečované použitím symetrickej, resp. asymetrickej kryptografie.

Ochranu multimediálnych dát po prenose je možné zabezpečiť použitím techník digitálnej vodotlače.

II. DIGITÁLNA VODOTLAČ

Metódy digitálnej vodotlače spočívajú vo vložení prídavnej informácie (vodoznaku) do multimediálnych dát. Dôležitou vlastnosťou vkladanej informácie je jej zmyslová nepostrehnuteľnosť a odolnosť voči potenciálnemu poškodeniu. Prídavná informácia v sebe môže niesť údaje o vlastníkovi autorských práv, resp. o distribútorovi a taktiež môže plniť funkciu zabezpečenia integrity dát.

Vodoznak môže mať formu konečnej postupnosti symbolov, resp. číslíc, prídavnej obrazovej informácie (loga), segmentu rečového signálu, poprípade môže mať formu 1-bitovej informácie, ktorá indikuje, či originálne dáta sú alebo nie sú označené vodoznakom.

Medzi základné požiadavky kladené na digitálne vodoznaky patria [3]:

- o robustnosť,
- o perceptívna transparentnosť,
- o nedetekovateľnosť,
- o bezpečnosť,
- o kapacita.

A. Digitálna vodotlač vo videu

Metódy digitálnej vodotlače vo videu môžeme rozdeliť do troch základných skupín [4]:

- o metódy na báze využitia vodotlače v statických obrazoch,
- o metódy na báze využitia časovej dimenzie videa,
- o metódy na báze využitia formátu komprimovaného videa.

B. Útoky na digitálne vodoznaky špecifické pre video

Útoky na digitálnu vodotlač vo videu predstavujú všetky úmyselné a neúmyselné operácie, ktoré sú vykonávané útočníkom s cieľom odstrániť vodoznak z označených dát, resp. poškodiť ho natoľko aby ho nebolo možné detekovať v označených dátach. Podľa úmyslu vykonávaných operácií môžeme útoky na digitálne vodoznaky vo videu rozdeliť do dvoch základných skupín:

- o úmyselné útoky,
- o neúmyselné útoky.

Medzi úmyselné útoky patria útoky výmeny a vynechania snímok, útok spriemernenia snímok a útoky na báze štatistickej analýzy. K neúmyselným útokom môžeme zaradiť operácie súvisiace so spracovaním obrazu (rôzne filtrácie, gamma korekcia ...) a kompresiu videa z jedného video formátu na iný.

III. PREHEAD SÚČASNÝCH METÓD DIGITÁLNEJ VODOTLAČE VO VIDEU V REÁLNOM ČASE

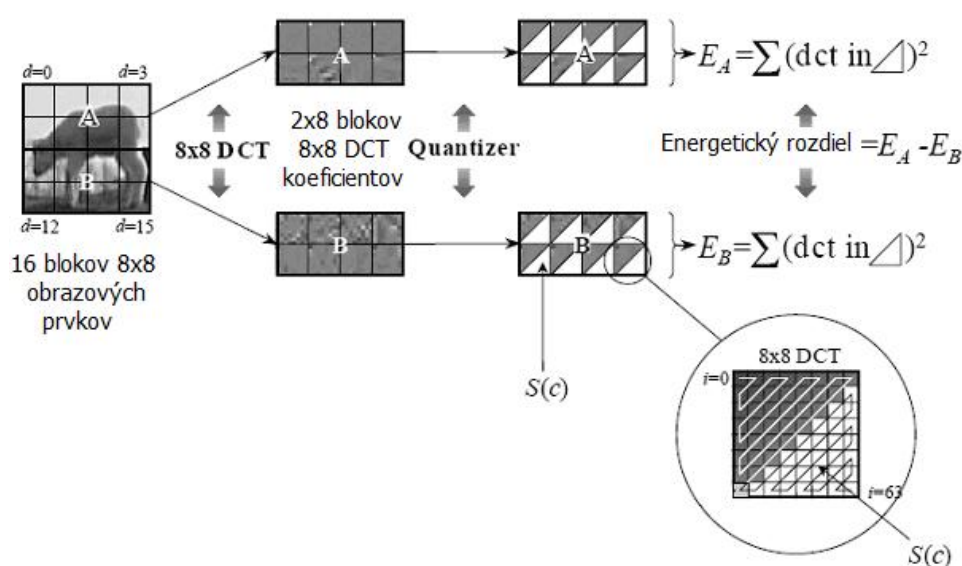
Metódy vkladania vodoznakov na báze využitia formátu komprimovaného videa sa stávajú čoraz viac populárnymi a žiadanými hlavne z dôvodu efektívnej ochrany on-line videa proti pirátstvu. Výhodou týchto metód je zníženie časovej náročnosti pri označovaní videa.

V nasledujúcich častiach článku bude uvedené stručné zhrnutie niekoľkých metód používaných v algoritmoch pre vkladanie digitálnych vodoznakov do videa v reálnom čase.

A. Metódy založené na nepriamej modifikácii transformačných koeficientov

Najznámejšou metódou digitálnej vodotlače vo videu založenou na modifikácii transformačných koeficientov je DEW (Differential Energy Watermarks). Pôvodne bola navrhnutá pre vkladanie vodoznakov do statických obrazov. Uvedený algoritmus je však možné aplikovať aj pre vkladanie vodoznakov do videa. Vkladanie vodoznaku sa realizuje do I snímok MPEG videa [5].

Podstatou tejto metódy je zanedbanie vysokofrekvenčných DCT koeficientov v bitovom toku komprimovaných dát. Proces vloženia vodoznaku je zobrazený na Obr. 1.



Obr. 1 Proces vloženia vodoznaku v algoritme DEW

Prvým krokom v procese vkladania je rozklad snímky na neprekrývajúce sa bloky 8x8 obrazových prvkov. Tieto bloky sú permutované pomocou tajného kľúča, čo zabezpečí dekorreláciu susedných blokov 8x8 obrazových prvkov. Získaná permutovaná snímka sa ďalej

rozdeli na n väčších neprekrývateľných blokov (v prípade Obr. 1, $n=16$). Každý väčší blok možno rozdeliť na dve polovice, t.j. horná časť A a dolná časť B, ktoré hrajú dôležitú úlohu pri vkladaní vodoznaku, ktorého bity sú postupne vkladané do každého väčšieho bloku, a to zavedením rozdielu energie vysokofrekvenčných DCT koeficientov medzi spomínanou hornou polovicou A a dolnou polovicou B väčšieho bloku.

Z dôvodu získania energie hornej a dolnej časti je na každý väčší blok aplikovaná 2D DCT transformácia po blokoch obrazových prvkov 8×8 . Získané koeficienty sú následne kvantované za použitia kvantizačnej matice pre štandard JPEG. Kvantované koeficienty sú potom usporiadané do cik-cak sekvencie a rozdelené na dve časti použitím indexu c hraničného DCT koeficientu (v našom prípade je $c=27$).

Výpočet energie ako aj podrobnejší popis tohto algoritmu je opísaný v [6]:

V poslednom kroku je aplikovaná 2D IDCT po blokoch 8×8 DCT koeficientov a následne inverzná permutácia za účelom získania označenej snímky.

Je potrebné poznamenať, že celý algoritmus možno efektívne využívať priamo v bitovom toku komprimovaného videa posunutím symbolu EOB zodpovedajúcich DCT blokov smerom ku DC koeficientu až po hraničný DCT koeficient s indexom c .

B. Metódy založené na modifikácii informácie o pohybe

Metódy tejto skupiny realizujú vkladanie vodoznaku modifikáciou vektorov pohybu. Je ich teda možné použiť iba pre tie formáty videa, ktoré potláčajú časovú redundanciu pomocou referenčného kódovania snímok. Keďže snímky videa sú navzájom veľmi podobné, je možné jednu snímku z druhej estimovať.

Spomínaný štandard MPEG-2 má tri typy snímok (I, P a B). Snímky I sú referenčné snímky a je na nich aplikovaná vnútrosnímková predikcia. Snímky P a B sú kódované referenčne na jednu I (alebo predchádzajúcu P) snímku, resp. referenčne na dve snímky (I a P) využitím informácie o pohybe. Táto informácia je vyjadrená skupinou vektorov pohybu určujúcich relatívny posun blokov obrazových prvkov pohyblivých oblastí na snímke. Tieto vektory pohybu môžeme rozložiť na dve zložky: horizontálnu a vertikálnu zložku a vloženie vodoznaku sa realizuje práve modifikáciou týchto vektorov pohybu získaných predikčným kódovaním.

Jednou z možností ako vložiť vodoznak do týchto vektorov je modifikovať ich tak, aby bola dodržaná parita na zložky vektora pohybu. Napríklad ak je bit vodoznaku 0, potom horizontálna zložka vektora pohybu je kvantovaná na párnú hodnotu, ak je bit 1 na nepárnu [7]. Nevýhodou týchto metód je nemožnosť predpovedať, aký vplyv bude mať modifikácia vektorov pohybu na video a na imperceptibilitu vodoznaku.

C. Metódy založené na priamej modifikácii kódových slov VLC kódu

Algoritmy tejto skupiny sú postavené na fakte, že nedokonalosť kompresných štandardov nedokáže odstrániť celú redundanciu z videa v procese kompresie. Táto redundancia je následne využívaná pre vkladanie vodoznaku modifikáciou kódových slov.

Ako príklad si môžeme opäť uviesť štandard MPEG-2. Po aplikovaní DCT sú získané transformačné koeficienty najprv kvantované, a následne sú zoradené do tzv. „cik-cak“ sekvencie. Zoradené kvantované koeficienty z tejto sekvencie sa následne transformujú na dvojice symbolov *run* a *level*, pričom symbol *run* sa rovná počtu nulových koeficientov predchádzajúcich nenulovému kvantovanému DCT koeficientu *level*. Symbol *level* nám teda udáva hodnotu (veľkosť) daného nenulového kvantovaného DCT koeficientu.

V poslednom kroku je potrebné tieto symboly podrobiť entropickému kódovaniu. V entropickom kódovaní sa dvojiciam symbolov (*run*, *level*) priradujú kódové slová s premennou dĺžkou z kódovacej tabuľky VLC kódu. Výsledkom entropického kódovania je výstupný bitový tok MPEG-2 videa, ktorý je možné modifikovať vkladáním vodoznakom.

Vloženie vodoznaku je možné realizovať nahradením LSB bitu symbolu *level* bitom vodoznaku [6]. Vzhľadom na to, že modifikácia kódových slov nesmie spôsobiť výskyt viditeľných artefaktov v snímkach, resp. výrazné zníženie vizuálnej kvality snímok, nie sú všetky kódové slová VLC vhodné na vloženie vodoznaku.

Z tohto dôvodu sa na označenie videa vyberá len také kódové slovo VLC, ktoré sa dá nahradiť kódovým slovom s nasledujúcimi vlastnosťami [4][6]:

- o rovnaká dĺžka kódového slova,
- o rovnaký symbol „run“,
- o symbol „level“ sa líši o 1.

Po splnení týchto podmienok teda dostaneme súbor kódových slov VLC vhodných pre vloženie vodoznaku, ktoré sa označujú ako *lc-VLC (label bit carrying VLC)*.

Tieto metódy sú podrobnejšie popísané v [4][6].

V procese extrakcie vodoznaku sa prehľadávajú všetky VLC kódové slová v každom makrobloku. Ak sa nájde kódové slovo *lc-VLC*, tak sa LSB bit jeho symbolu *level* presadí do bitového toku prúdového vodoznaku. Postup sa opakuje, kým sa nenájdu ďalšie kódové slová *lc-VLC*.

Výhodami metód modifikácie bitového toku je jednoduchý proces vkladania a extrakcie vodoznaku vo videu, a tak nízka časová a výpočtová náročnosť, ktorá je predpokladom použitia metódy pre označenie videa v reálnom čase. Nevýhodami týchto metód je ich citlivosť na operácie súvisiace so spracovaním videa a na geometrické útoky.

D. Metódy založené na singulárnom rozklade hodnôt

Singulárny rozklad matice (ďalej len SVD) je užitočnou metódou, ktorá nachádza využitie pri mnohých maticových výpočtoch a pri analýze dát. Poskytuje informácie o geometrickej štruktúre matice a realizuje maticové rozklady, ktorých hlavným cieľom je zjednodušiť výpočty pri riešení sústav lineárnych rovníc. V poslednej dobe je SVD používaná k efektívnemu vyhľadávaniu vo veľkých objemoch dát, pri porovnávaní signálov alebo obrazov z dôvodu realizácie vhodnej kompresie týchto dát.

Metóda SVD je ortogonalizačný proces, ktorý vedie k dekompozícii matice A na diagonálnu maticu S a na jej ľavé a pravé singulárne matice U a V pričom platí [8][9]:

$$A = USV^T \quad (1)$$

Za predpokladu, že matica A je rozmeru $m \times n$ môžeme vzťah (6.5) zapísať takto:

$$SVD(A_{m \times n}) = [U_{m \times m} S_{m \times n} V_{n \times n}] \quad (2)$$

Vynásobením U, S a V^T (kde V^T je transponovaná matica) získame maticu:

$$A_{m \times n} = U_{m \times m} \cdot S_{m \times n} \cdot V_{n \times n}^T = \sum_{i=1}^{\min(m,n)} \sigma_i \cdot u_i \cdot v_i^T \quad (3)$$

kde $\sigma_i = \{R_+\}$, $i=1, \dots, \min(m,n)$ sú singulárne hodnoty (singular values SV) t.j. diagonálne prvky matice S , ktoré sú zoradené v zostupnom poradí. u_i sú ľavé singulárne vektory t.j. stĺpce matice U (alebo riadky matice U) a v_i sú pravé singulárne vektory t.j. riadky matice V^T (alebo stĺpce matice V). U a V sú unitárne matice, čo znamená že:

$$U \cdot U^T = I_{m \times m}, V \cdot V^T = I_{n \times n} \quad (4)$$

kde $I_{m \times m}$ a $I_{n \times n}$ sú jednotkové matice.

Hodnoty matice S sú úzko späté s jasom obrazu, zatiaľ čo skutočná geometria obrazu závisí na unitárnych maticiach U a V , ktoré predstavujú horizontálne a vertikálne okraje obrazu.

Taktiež je zrejmé, že singulárne hodnoty t.j. diagonálne prvky S matice, ktoré sú zoradené v zostupnom poradí, sú násobené príslušným ľavým a pravým singulárnym vektorom. Toto násobenie vytvára rôzne vrstvy obrazu, ktoré sú nakoniec spočítané a tvoria konečný obraz. Tak napríklad: násobenie prvej singulárnej hodnoty s príslušným ľavým a pravým singulárnym vektorom tvorí tzv. obrazový profil, ktorý sústreďuje veľké množstvo energie obsiahnutej vo výslednom obraze. Ľavé a pravé singulárne vektory, ktoré zodpovedajú najväčším singulárnym hodnotám predstavujú tvar (výrazné okraje obrazu), zatiaľ čo zvyšok singulárných vektorov predstavuje oblasti hrán a textúr [10].

Jednou z veľkých výhod použitia SVD je invariantnosť singulárných hodnôt k bežným operáciám spracovania obrazu ako aj odolnosť voči geometrickým zmenám, napr. rotácia a taktiež odolnosť voči škálovaniu. Ak vezmeme do úvahy taktiež jednoduchosť jej aplikácie, táto metóda sa javí ako vhodný kandidát pre vodoťlač v reálnom čase.

Algoritmy digitálnej vodoťlače s využitím SVD môžeme rozdeliť do nasledujúcich skupín:

- o algoritmy vodoťlače, založené na modifikácii singulárnej hodnoty,
- o algoritmy vodoťlače, založené na modifikácii singulárneho vektora,
- o algoritmy kombinujúce viaceré metódy vkladania.

Algoritmy vodoťlače založené na modifikácii singulárnej hodnoty

Pri týchto algoritmoch máme možnosť návrhu schémy vkladania vo viacerých variantoch.

Najjednoduchšia schéma vkladania je založená na pridaní vodoznaku prostredníctvom modifikácie singulárnych hodnôt celého originálneho obrazu, kde vložený vodoznak je binárny obraz alebo pseudonáhodná postupnosť. Ďalšou možnosťou je modifikácia singulárnych hodnôt originálneho obrazu singulárnymi hodnotami vodoznaku, na ktorý bola aplikovaná SVD. Výhodou týchto metód je hlavne jednoduchá implementácia. Ďalším zjednodušením výpočtovej a časovej náročnosti môže byť redukcia počtu snímok z videosekvencie, ktoré chceme označiť.

Algoritmy vodoťlače založené na modifikácii singulárneho vektora

Podstatou týchto algoritmov vodoťlače je modifikácia ľavého alebo pravého singulárneho vektora (matice U alebo V). Príklad takéhoto algoritmu je bližšie opísaný v [11].

Algoritmus používa pri vkladani binárny vodoznak. Originálny obraz je rozdelený do neprekrývajúcich sa blokov o veľkosti 8×8 . Bloky určené pre vkladanie vodoznaku sú vybrané pomocou tajného kľúča. Každý bit vodoznaku je vložený tak, že sa modifikuje závislosť medzi druhým a tretím koeficientom prvého stĺpca matice U. Analogický je možné použiť tento algoritmus aj modifikovaním matice V.

Výhodou tohto algoritmu je jeho robustnosť voči rôznym druhom útokov a taktiež absencia nutnosti prítomnosti originálnych dát v procese extrakcie. Na druhej strane, však použitím tohto algoritmu nedosiahneme presne extrahovaný vodoznak keďže pracujeme s unitárnymi maticami, ktoré predstavujú horizontálne a vertikálne okraje obrazu.

Algoritmy kombinujúce viaceré metódy vkladania

Na zvýšenie robustnosti algoritmov digitálnej vodoťlače využívajúcich SVD transformáciu môžeme tieto algoritmy kombinovať s použitím iných DOT, ako napr. DCT, DWT, atď.

Jednou z mnohých metód je transformácia pôvodných farebných zložiek obrazu RGB do zložiek YUV a následná aplikácia DWT na luminančnú zložku Y. Ďalšou aplikáciou SVD transformácie na aproximačný obraz získavame singulárnu maticu, ktorú následne modifikujeme vodoznakom. Podobný popis vkladania vodoznaku patriaceho do tejto skupiny je opísaný v [12].

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] W. Stallings, "Cryptography Cryptography and Network Security": Principles and Practise. 3rd ed, Prentice Hall, 2002. 696 p. ISBN 978-0130914293.
- [2] R. Ridzoň, D. Levický, "Usage of different color models in robust digital watermarking". In *Radioelektronika 2009: 19th International Czech - Slovak scientific Conference*, April 2009, Bratislava, Slovak Republic.
- [3] M. WU, B. Liu, "Multimedia data hiding", Springer-verlag New York, Inc. 2003.
- [4] G. Doerr, "Security Issue and Collusion Attacks in Video Watermarking". PhD thesis, Universite de Nice Sophia-Antipolis, 2005.
- [5] G. Langelaar, R. Lagendijk, "Optimal Differential Energy Watermarking of DCT Encoded Images and Video". *IEEE Transactions on Image Processing*, 10(1): 148-158, January 2001.
- [6] G. Langelaar, R. Lagendijk, R. Biemond, "Real-time Labelling of MPEG-2 Compressed Video". *Journal of Visual Communication and Image Representation*, 9(4):256-270, August 1998.
- [7] F. Jordan, M. Kutter, T. Ebrahimi, "Proposal of Watermarking Technique for Hiding/Retrieving Data in Compressed and Decompressed Video". In *JTC1/SC29/WG11 MPEG97/M2281. ISO/IEC*, July 1997.
- [8] A. Basso, F. Bergadano, D. Cavagnino, "A n A Novel Block-Based Watermarking Scheme Using the SVD Transform", January 2009.
- [9] L. Rajab, T. Al-Khatib, A. Al-Haj, "Video Watermarking Algorithms Using the SVD Transform", 2009.
- [10] M. Andráš, "Vodoznaky vo videu v reálnom čase", Diplomová práca, FEI TU Košice, 2010.
- [11] B. C. Mohan, S. A. Kumar, "Robust Digital Image Watermarking Scheme using Singular Value Decomposition", 2008.
- [12] A. Al-Haj, T. Al-Khatib, L. Rajab, "A Robust Video Watermarking Algorithm", 2008.

Násobička frekvencie troma pre UWB radarový systém v technológií SiGe BiCMOS

Martin LIPTAJ, Pavol GALAJDA, Branislav HRUŠOVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

martin.liptaj@tuke.sk, pavol.galajda@tuke.sk, branislav.hrusovsky@tuke.sk

Abstrakt – Článok opisuje návrh násobičky frekvencie troma pre UWB radarový systém. Opisovaná topológia je založená na diferenciálnom zosilňovači s induktívnou záťažou v kolektorovom obvode. Realizovaná násobička bude súčasťou niekoľkých ASIC obvodov implementovaných v novom UWB radarovom systéme. V článku je prezentovaná koncepčná simulácia a tiež experimentálna realizácia násobičky. Simulácia je vykonaná pomocou návrhových prostriedkov od firmy Mentor Graphics s využitím reálnych technologických dát. Pre experimentálnu realizáciu sa využili obvody vyrobené v technológií 0,25 μm SiGe BiCMOS od IHP, Nemecko. Pasívna časť je realizovaná s vysokokvalitných SMD súčiastok. Cieľom je úplná implementácia prezentovanej násobičky frekvencie troma - triplera do technológie 0,35 μm SiGe BiCMOS od Austriamicrosystems, Rakúsko.

Kľúčové slová – frekvenčná násobička troma, tripler, UWB, širokopásmové aplikácie, ASIC, SiGe BiCMOS

I. ÚVOD

Širokopásmové aplikácie sa v súčasnosti rozširujú v každej oblasti ľudského pôsobenia. Veľmi významné je najmä rozšírenie v oblasti určovania polohy či v senzorovej technike. V týchto oblastiach je dôležité zlepšovanie súčasných alebo vývoj lepších nových riešení. Medzi takéto riešenia patrí aj UWB radarový systém. V tomto UWB radarovom systéme a tiež v mnohých ďalších systémoch sú používané frekvenčné násobičky na konverziu stabilnej frekvencie oscilátora na vyššie frekvencie. Násobičky frekvencie troma (tzv. tripler-y), založené na FET technológií, sa často používajú v jednoducho zapojenej štruktúre [1]. Tento článok však opisuje násobičku frekvencie troma - tripler na báze bipolárnej SiGe BiCMOS technológie, ktorá konvertuje vstupný signál s frekvenciou 2 GHz na výstupný signál s frekvenciou 6 GHz. Tieto frekvencie vychádzajú zo súčasných štandardov ECC pre UWB radarové systémy [2][3][4]. Inšpiráciou pre takéto riešenie boli podobné projekty opísané v [5] a [1]. Tripler opísaný v článku je podobne ako tripler v [1] založený na diferenciálnom zosilňovači s induktívnou záťažou v kolektorovom obvode. Tripler opísaný v [5] je tiež založený na báze diferenciálneho zosilňovača avšak nevyužíva induktívnu záťaž v kolektorovom obvode. Zo [5] pochádza inšpirácia zapojenia filtra na výstup triplera.

Prvá časť článku opisuje návrh a simuláciu navrhovaného triplera v technológií 0,35 μm SiGe BiCMOS od Austriamicrosystems, Rakúsko. Ďalej je porovnanie technológií pre výrobu ASIC obvodov a hlavne výhody a nevýhody SiGe BiCMOS technológie všeobecne. Článok pokračuje opisom experimentálnej implementácie triplera realizovaného pomocou dvoch diferenciálnych zosilňovačov v jednom púzde, ktoré boli pôvodne navrhnuté ako RX zosilňovač pre rozšírenú UWB radarovú architektúru. Nakoniec sú opísané výsledky meraní.

II. NÁVRH OBVODU

Diferenciálny zosilňovač reprezentuje základný stavebný blok v mnohých lineárnych a nelineárnych moduloch integrovaných obvodov.

Ak si rozdielové napätie na vstupe diferenciálneho zosilňovača vyjadríme ako

$$\nu_1 - \nu_2 = V_1 \cos \omega t \quad (1)$$

a zdefinujeme $x = qV_1/kT$, potom vieme prúdy tečúce kolektormi tranzistorov vyjadriť ako

$$i_1 = \frac{I_k}{2} + i \quad \text{a} \quad i_2 = \frac{I_k}{2} - i \quad (2)$$

kde

$$i = \frac{I_k}{2} \tanh\left(\frac{x}{2} \cos \omega t\right) \quad (3)$$

Tento prúd kolektorom i môžeme vyjadriť pomocou Fourierovho radu:

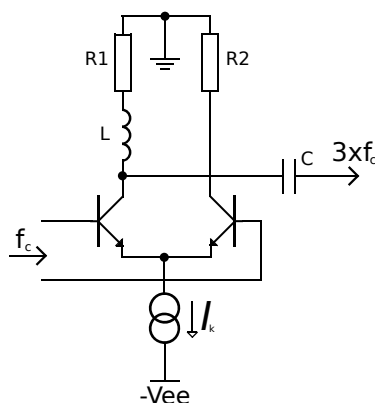
$$\begin{aligned} i &= I_1 \cos \omega t + I_3 \cos 3\omega t + I_5 \cos 5\omega t + \dots \\ &= I_k \sum_{n=2}^{\infty} a_{2n-1}(x) \cos(2n-1)\omega t \end{aligned} \quad (4)$$

kde

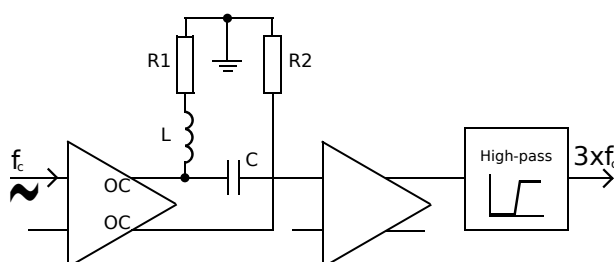
$$a_n(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} \left[\frac{1}{2} \tanh\left(\frac{x}{2} \cos \theta\right) \right] \cos n\theta d\theta = \frac{I_n}{I_k} \quad (5)$$

Z rovnice 4 je na prvý pohľad zrejmé, že aj pri malých signáloch diferenciálny zosilňovač nevytvára párne harmonické zložky [6]. Tento fakt využijeme pre vytvorenie násobičky frekvencie troma.

V prípade, že na výstup takého zosilňovača pripojíme hornopriepustný filter, ktorý odfiltruje základnú harmonickú, najvyšší výkon bude mať tretia harmonická. Ak do obvodu jedného kolektora ešte zapojíme ladený obvod s rezonančnou frekvenciou tretej harmonickkej zložky vstupného signálu, zvýšime konverziu násobičky troma na maximum.



Obr. 1 Násobička frekvencie troma založená na diferenciálnom zosilňovači

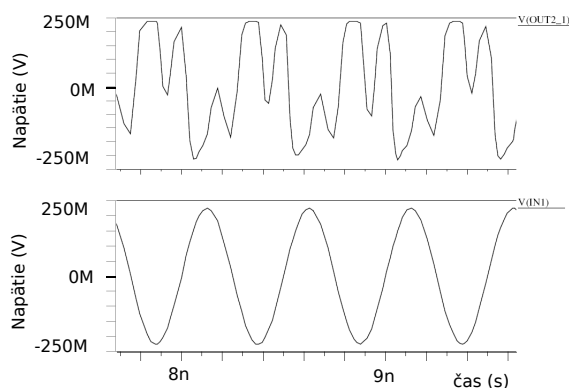


Obr. 2 Bloková schéma násobičky frekvencie troma založenej na diferenciálnom zosilňovači

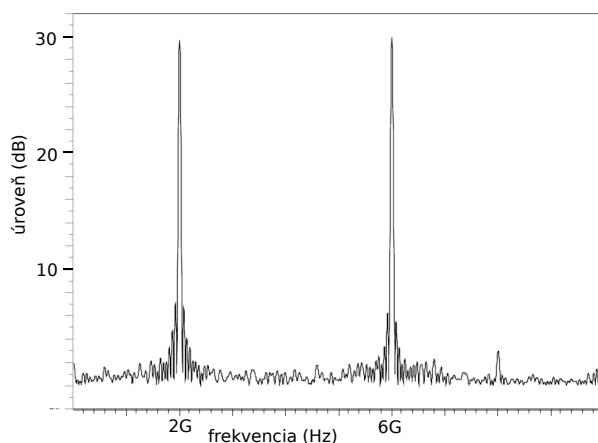
V našej realizácii testovacieho návrhu pozostáva násobička z dvoch diferenciálnych zosilňovačov (Obr. 2). Prvý zosilňovač obsahuje vo svojom kolektore zapojenú kvalitnú cievku. Táto cievka spolu s parazitnými kapacitami tranzistorov a pripojeným odporom vytvára RLC ladený obvod (Obr. 1). Tento obvod je naladený približne na tretiu harmonickú zložku základnej harmonickej. Druhý zosilňovač slúži ako výstupný buffer, charakteristický veľmi vysokou vstupnou impedanciou. Medzi zosilňovače je zaradený kondenzátor kvôli oddeleniu jednosmernej zložky z obvodov nastavenia pracovného bodu. Pre lepšie oddelenie tretej harmonickkej od základnej harmonickej je na výstup zaradený ešte hornopriepustný filter (Obr. 2).

Násobička frekvencie troma bola odsimulovaná a optimalizovaná bez pripojenia hornopriepustného filtra. Simulované výsledky sú platné pre SiGe BiCMOS 0,35 μm technológiu od Austriamicrosystems, Rakúsko (AMS). Simulácie boli realizované pre zapojenie s jednoduchým výstupom s výkonom tretej harmonickkej na úrovni -6 dBm. Vstupný výkon základnej harmonickej s frekvenciou 2 GHz bol na úrovni 0 dBm. Takto zapojený obvod pri simulácii odoberal prúd o veľkosti 40 mA pri napájacom napätí 3,3 V. To znamená, že straty násobičky frekvencie troma

zapríčinené konverziou sú na úrovni 6 dB. Potlačenie druhej harmonickej je na úrovni -26 dBc, to znamená, že je takmer na úrovni šumu. Úroveň základnej harmonickej na výstupe je -2dBm. Simulované výsledky sú zobrazené na Obr. 3 a 4.



Obr. 3 Simulovaný vstupný a výstupný signál z násobičky frekvencie troma



Obr. 4 Simulované výstupné frekvenčné FFT-spektrum násobičky frekvencie troma

III. TECHNOLÓGIA

Prvé pokusy so SiGe ako polovodičovým materiálom sa uskutočnili už v roku 1960 no prvý tranzistor založený na tejto technológii bol demonštrovaný až v roku 1987. Do celosvetovej pozornosti sa táto technológia dostala hlavne na začiatku 90-tých rokov minulého storočia, kedy sa vyrobili prvé tranzistory s hraničnou frekvenciou 75 GHz [7]. V tej dobe to bola približne dvojnásobná frekvencia, akú dosahovali najlepšie štandardné kremíkové tranzistory. Postupne sa technológia vylepšovala a dnes sa SiGe BiCMOS tranzistory vyrábajú s maximálnymi frekvenciami presahujúcimi 300 GHz a s veľmi nízkym šumovým číslom. Keďže BiCMOS technológia vychádza a je kompatibilná s CMOS technológiou, SiGe BiCMOS technológia sa využíva hlavne v oblastiach, kde je potrebné spojiť analógovú aj digitálnu časť do jedného čipu. Výsledkom sú čipy, ktoré sú vo veľkých sériach lacné, vykazujú relatívne vysoký analógový aj výpočtový výkon a pritom sú zároveň energeticky úsporné.

Hlavnými konkurentmi SiGe BiCMOS technológie sú technológie založené na materiáloch GaAs a InP. V priemysle sa preferujú výrobné procesy na báze kremíka hlavne kvôli možnosti vyrobiť veľmi veľké kusy čistého kremíku a tým zefektívniť a teda aj zlacniť výrobu polovodičových ASIC obvodov. Výhodnou vlastnosťou kremíku je tiež pomerne dobrá tepelná vodivosť. GaAs aj InP sa naproti tomu vyrábajú na menších platniach s vyššou chybovosťou v kryštalickej mriežke a zároveň majú horšie tepelné vlastnosti ako kremík. Tabuľka 1 porovnáva tieto tri technológie v niekoľkých aspektoch. Pre porovnanie je v tabuľke uvedená aj CMOS technológia, ktorá ale ako vieme dominuje hlavne digitálnemu svetu elektroniky. Ako je možné vidieť, SiGe BiCMOS technológia je vo všetkých aspektoch uspokojivá. Ak však potrebujeme vyššie spínacie rýchlosti, je vhodnejšie využiť GaAs alebo InP technológiu, či v prípade požiadavky na vysokú linearitu systému môže byť vhodnejšie využitie GaAs technológie [8].

Tabuľka 1
 RELATÍVNE POROVNANIE TECHNOLOGIÍ CMOS, SiGe BiCMOS, GaAs A InP [8][9][10][11][12]

Parameter	CMOS	SiGe BiCMOS	GaAs	InP
Linearita	Lepšia	Dobrá	Najlepšia	Dobrá
Pomer SNR	Zlý	Dobrý	Zlý	Dobrý
irokopásmový šum	Zlý	Dobrý	Zlý	Zlý
Pracovná frekvencia	Dobrá	Dobrá	Lepšia	Lepšia
Earlyho napätie	Zlé	Dobré	Dobré	Dobré
Transkonduktancia	Zlá	Dobrá	Dobrá	Dobrá
Cenová dostupnosť	Najlepšia	Dobrá	Zlá	Zlá

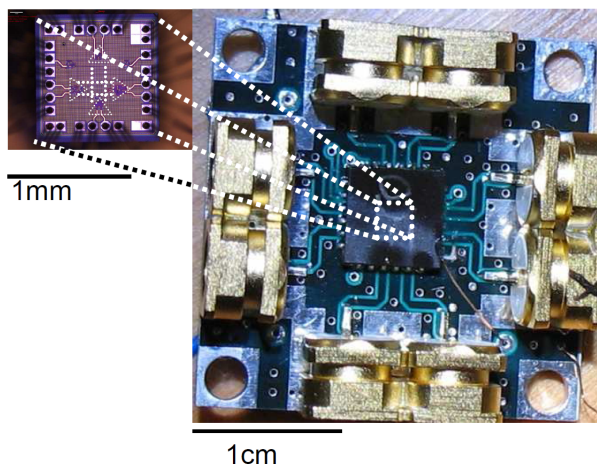
Vo svete existuje veľké množstvo firiem, ktoré sa zaoberajú vývojom a výrobou elektroniky založenej na báze SiGe BiCMOS obvodov. Medzi najznámejších a najväčších výrobcov patrí napríklad IBM, ktoré je aj zodpovedné za vznik a rozvoj tejto technológie. Medzi ďalšie väčšie a známe firmy môžeme zaradiť Jazz semiconductor, Freescale semiconductor, nemecký IHP GmbH, alebo rakúsku firmu Austriamicrosystems (AMS). Všetky tieto firmy majú vo svojom portfóliu služieb zahrnutú technológiu SiGe BiCMOS s technologickým procesom od 800 nm až po 130 nm. Je samozrejmé, že čím je tento základný technologický rozmer menší, tým je technológia drahšia no zároveň aj výkonnejšia. Pri výbere použitej technológie je preto potrebné si určiť, čo od navrhovaného obvodu potrebujeme a ktorá technológia bude z ekonomického hľadiska pre tento účel najvýhodnejšia.

Tento článok je opisuje tripler, ktorý pracuje do 6 GHz. Pre tento účel je preto vhodná 0,35 μm SiGe BiCMOS technológia od AMS. Výrobca pre túto technológiu uvádza medznú frekvenciu $F_t = 60$ GHz (frekvencia, pri ktorej absolútna hodnota prúdového zosilňovacieho činiteľa v zapojení so spoločným emitorom klesne na hodnotu $\beta = 1$) a maximálnu frekvenciu $F_{max} = 70$ GHz (frekvencia, pri ktorej je tranzistor schopný vybudenia vlastných kmitov). Je vhodná pre výrobu širokopásmových ASIC obvodov pre aplikácie do 20 Gb/s resp. s prevádzkovou frekvenciou do 10 GHz.[13].

IV. IMPLEMENTÁCIA

Overenie myšlienky realizácie triplera bolo uskutočnené podľa zapojenia na Obr. 2. Tripler je vytvorený z dvoch diferenčných zosilňovačov, pôvodne navrhnutých ako zosilňovače RX stupňa UWB radarového systému. Tieto zosilňovače sú mierne odlišné ako simulované zosilňovače, avšak tento rozdiel nie je natoľko podstatný, aby sa nedali použiť na overenie správnosti simulácií.

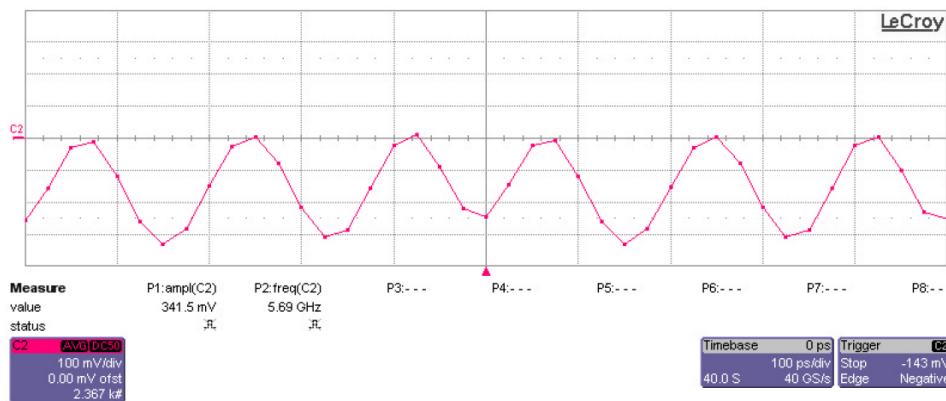
Prvý zosilňovač bol realizovaný ako diferenčný zosilňovač s otvorenými kolektormi, z ktorých na jeden sa pripojila externá indukčná záťaž v podobe seriového zapojenia cievky a odporu určených pre povrchovú montáž (SMD). Na druhý kolektor sa pripojil len odpor, tak isto v SMD prevedení. Výstup s indukčnou záťažou bol cez kondenzátor pripojený na vstup druhého zosilňovača. Ten bol v prevedení s vysokou vstupnou impedanciou (asi 5 k Ω). Oba zosilňovače boli navrhnuté na jednom čipe v technológii 0,25 μm SiGe BiCMOS od IHP, Nemecko a zapuzdrené v púzdre QFN24. Takto vytvorená súčiastka bola priletovaná na kvalitnejšiu dosku plošných spojov s označením Rogers 4003c (Obr. 5). Na tejto doske je tiež priletovaná externá cievka a odpory pripojené k otvoreným kolektorom prvého diferenčného zosilňovača. Výstup druhého zosilňovača bol pripojený na hornopriepustný filter s hraničnou frekvenciou 3,8 GHz.



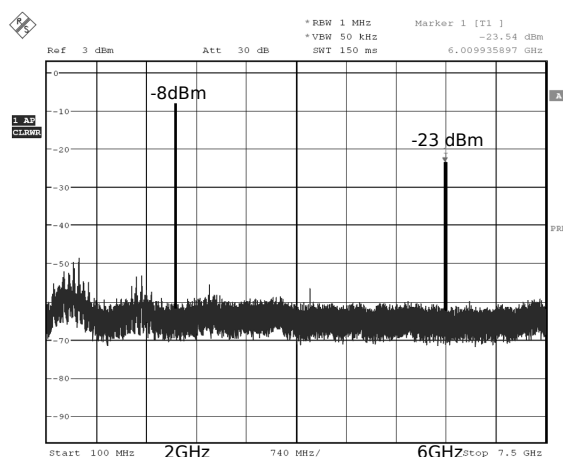
Obr. 5 Fotografia nezapuzdreného čipu a doska plošných spojov s priletovaným čipom

V. NAMERANÉ VÝSLEDKY

Pri napájaní napätím -3,3 V tripler spotrebuje menej ako 70 mA. Na Obr. 7 je zobrazené frekvenčné spektrum výstupného signálu z triplera. Vstupný signál vstupujúci do triplera bol s frekvenciou 2 GHz a s výkonovou úrovňou 0 dBm. Použité signálové káble prispeli stratou približne 5 dB @ 2 GHz. To znamená, že výkonová úroveň signálu na vstupe triplera bola znížená na -5 dBm. Pri týchto podmienkach pre výstupný signál s frekvenciou 6 GHz bola nameraná výkonová úroveň -23 dBm. Nameraná výkonová úroveň základnej harmonickej bola -8 dBm a druhej harmonickej -55 dBm. Použitím hornopriepustného filtra pripojeného na výstup tejto testovacej štruktúry sa selektivita výrazne zvýšila. Na Obr. 6 a 8 je zobrazený výstupný signál z triplera s pripojeným filtrom meraný vo frekvenčnej aj časovej oblasti.



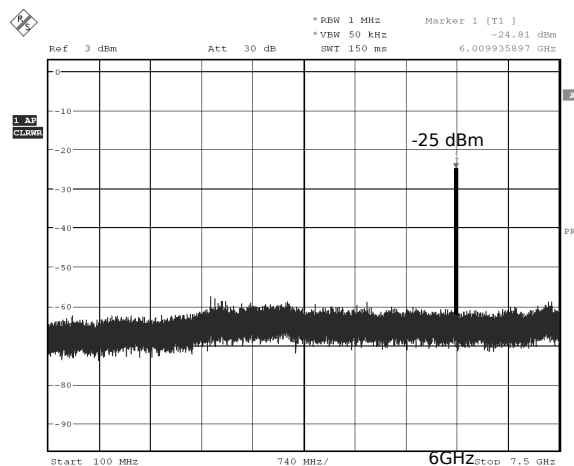
Obr. 6 Výstupný signál násobičky frekvencie troja so zapojeným filtrom v časovej oblasti



Obr. 7 Namerané výstupné spektrum násobičky frekvencie troja

VI. ZÁVER

V tomto článku je prezentovaná násobička frekvencie troja, tzv. tripler v technológií 0,35 μm SiGe BiCMOS od Austriamicrosystems, Rakúsko, výsledky simulácií a meraní testovacej štruktúry. Myšlienka bola overená na reálnej testovacej štruktúre s využitím čipu pozostávajúceho z dvojice diferenciálnych zosilňovačov. Napriek tomu, že použitá štruktúra bola pôvodne navrhnutá pre iný účel a v inej technológii, simulované a namerané výsledky sú veľmi podobné. V blízkej budúcnosti sa všetky súčasti triplera okrem indukčnej cievky navrhnu ako jeden čip a budú vyrobené v spomínanej technológii od AMS. Pre konečnú realizáciu sa bude používať len jedna externá cievka, pomocou ktorej sa bude dať prednastaviť požadovaná rezonančná frekvencia RLC ladeného obvodu. Táto integrovaná štruktúra bude súčasťou nového UWB radarového systému vyvíjaného na TU Ilmenau. V rámci spolupráce medzi TU Ilmenau a TU Košice sa na Katedre elektroniky a multimediálnych telekomunikácií začal a plánuje ďalší vývoj ASIC obvodov v technológii 0,35 μm SiGe BiCMOS od AMS, Rakúsko.



Obr. 8 Namerané výstupné spektrum násobičky frekvencie tromu so zapojeným filtrom

POĎAKOVANIE

Autori ďakujú Petrovi Rauschenbachovi (Meodat GmbH) a Martinovi Kmecovi (TU Ilmenau) za pomoc. Táto práca vznikla s podporou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] M. Danesh, F. Gruson, P. Abele, and H. Schumacher, "Differential VCO and Frequency Tripler using SiGe HBTs for the 24GHz ISM Band," *RFIC*, 2003, Philadelphia, PA, USA.
- [2] M. Kmec, R. Herrmann, J. Sachs, P. Peyerl, and P. Rauschenbach, "Extended Approaches for M-Sequence based UWB Systems," *Ultra-Wideband Short-Pulse Electromagnetics 8*, 2007, edited by Tyo, J. Scott; Baum, Carl E.; Stone, Alexander P.
- [3] J. Sachs, R. Herrmann, M. Kmec, M. Helbig, and K. Schilling, "Recent Advances and Applications of M-Sequence based Ultra-Wideband Sensors," *ICUWB*, September 2007, Singapore.
- [4] R. Herrmann, J. Sachs, K. Schilling, and F. Bonitz, "New extended M-sequence ultra wideband radar and its application to the disaggregation zone in salt rock," *12th International Conference on Ground Penetrating Radar*, 15-19 June 2008, Birmingham, UK.
- [5] Y. Zeng and C.E.Saavedra, "A bipolar MMIC Frequency Tripler," *Proceedings of the 7th International Caribbean Conference on Devices*, 2008, Circuits and Systems, Mexico.
- [6] K. K. Clarke and D. T. Hess, "Communication Circuits: Analysis and Design," *Addison-Wesley Publishing Company, Inc.*, 1971.
- [7] J. Cressler, "SiGe HBT Technology: A New Contender for Si-Based RF and Microwave Circuit Applications," *IEEE Transactions on microwave theory and techniques*, vol. 46, May 1998.
- [8] M. Liptaj, M. Kmec, P. Galajda, and J. Sachs, "Recent SiGe frequency tripler development for a new UWB system architectures," *Research in Telecommunication Technologies 2010 : the 12th International Conference*, September 2010, Velké Losiny, Czech Republic. Ostrava : VŠB-TU, ISBN 978-80-248-2261-7.
- [9] S. Dunn and et al., "Foundation of RF CMOS and SiGe BiCMOS technologies," *IBM Journal of Research and Development*, pp. 101–138, March/May 2003.
- [10] S. Voinigescu and et al., "Comparison of Si CMOS, SiGe BiCMOS, and InP HBT Technologies for High-Speed and Millimeter-Wave ICs," *SiRF*.
- [11] N. Guofu, "SiGe Technology for Revolution of Personal Communications," *dostupne na internete*., <http://www.eng.auburn.edu/~guofu>.
- [12] A. Pawlikiewicz, "RF CMOS or SiGe BiCMOS in RF and Mixed Signal Circuit Design," *Mixed Design of Integrated Circuits and Systems, 14th International Conference*, pp. 333 – 338, 2007.
- [13] www.austriamicrosystems.com, "stránka spoločnosti," *dostupne na internete*., <http://www.austriamicrosystems.com>.

Parallel Processing Applied on Elliptic Curve Arithmetic in Affine Coordinates

Patrik GALLO, Peter GOČ-MATIS, Dušan LEVICKÝ

Dept. of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

patrik.gallo@tuke.sk, peter.goc-matis@tuke.sk, dusan.levicky@tuke.sk

Abstract — in this paper we present a view over elliptic curve arithmetic with regard to parallel processing on multiprocessor units in general. We provide an analysis of how optimal or not it can be using parallel processors instead of single processor units. Amdahl's law is considered as a first step in this analysis. Analysis includes theoretical introduction to elliptic curve cryptosystem, especially to elliptic curve arithmetic with regard to affine coordinate system. Analysis is concluded at the end of the paper.

Keywords — elliptic curve arithmetic, parallel processing, affine coordinates

I. INTRODUCTION

The theory of doing computational work in parallel has some fundamental laws that place limits on the benefits one can derive from parallelizing a computation. To understand these laws, we have to first define the objective. In general, the goal in large scale computation is to get as much work done as possible in the shortest possible time. In a perfect world, a computational job that is split up among N processors would complete in $1/N$ time, leading to an N -fold increase in power. However, any given piece of parallelized work to be done will contain parts of the work that *must* be done serially, one task after another, by a single processor. This part does *not* run any faster on a parallel collection of processors (and might even run more slowly). Only the part that can be parallelized runs as much as N -fold faster.

In case of this research paper, the object of parallelism is a part of computation over elliptic curve arithmetic. By elliptic curve arithmetic we mean operations on elliptic curve defined over some finite field. Thus, algebraically it means performing operations over several coordinate system, for example affine, projective, Jacobian, modified Jacobian, Chudnovsky, and so on. The aim of this paper is to analyze a parallelism in affine coordinate system, but the theoretical basis holds for all the coordinate systems.

II. ELLIPTIC CURVE DISCRETE LOGARITHM PROBLEM

Given an elliptic curve E defined over a finite field F_q , and two points $P, Q \in E(F_q)$, find an integer l such that $lP = Q$ in E , provided that such an integer exists. The algebraic objects in the ECDLP (elliptic curves over finite fields) are equipped with only one basic operation: addition of elliptic curve points. On the other hand, objects in DLP are equipped with two basic operations: addition and multiplication of field elements. The additional structure present in the DLP has led to the discovery of the index-calculus methods, which have a subexponential running time. Elliptic curves do not possess this additional structure, and for this reason no one has been able to apply the index-calculus methods to the ECDLP (except in very special and wellunderstood cases). This absence of subexponential-time algorithms for the ECDLP, together with efficient implementation of the elliptic curve arithmetic, is precisely the reason that elliptic curve cryptosystems have proven so attractive for practical use.

III. ELLIPTIC CURVE ARITHMETIC

A. Definition of an elliptic curve

An elliptic curve is a set of points satisfying an equation of the form $y^2 = x^3 + ax + b$ where the coefficients a and b are elements of a field F with the characteristic of F is not equal to 2 or 3. Detailed explanation why are we using an equation of this form and why the characteristic of

a field should not to be equal to 2 or 3 is shown in [2]. We write E/F to indicate that the elliptic curve is over the field. Such a curve is said to be in *Weierstrass normal form*. We can think of the points on an elliptic curve as being either points in a set or as rational functions of x and y , and can freely change between the two points of a view as needed. The requirement that the characteristic of F be greater than 3 is not strictly required for an elliptic curve, but this restriction limits us to the elliptic curves of interest in the discussion of secure cryptographic systems. If the characteristic of the field over which an elliptic curve is defined is equal to 2 or 3, alternative forms other than Weierstrass normal form need to be used. In particular, the algorithm for adding points on an elliptic curve in Weierstrass normal form uses the constants 2 and 3, which makes it behave badly over a field of characteristic 2 or 3, so alternate forms are needed in these cases.

Elliptic curves for which the cubic has repeated roots turn out to have undesirable properties from the cryptographic point of view, so it is useful to be able to identify which curves these are. To determine this, we can find the roots of the cubic part of an elliptic curve in Weierstrass normal form by factoring $x^3 + ax + b = 0$ into $(x - x_1)(x - x_2)(x - x_3)$ and finding values of x_1 , x_2 and x_3 . This process is quite difficult so I'm not showing it at this place. You can find it in literature [2]. Using these values for roots of the cubic we find that

$$(x_1 - x_2)^2(x_1 - x_3)^2(x_2 - x_3)^2 = -(4a^3 + 27b^2)$$

The right side of the equation we is called the *discriminant* of an elliptic curve. So that we have a repeated root whenever $-(4a^3 + 27b^2) = 0$.

B. Algebraic structure of elliptic curves

Points on an elliptic curve provide a structure that can define in the terminology of abstract algebra.

If E is an elliptic curve over field F then we write $E(F)$ to indicate the set of points on E along with the operation of adding points described later in algorithm.

If F is a field and E is an elliptic curve then $E(F)$ is a group. The point at infinity O acts as identity element for this group. Note that there is only operation defined for $E(F)$, which we are thinking of as addition, so it is impossible to multiply or divide elements of $E(F)$. Thus, $E(F)$ cannot be a field, which requires two operations that we think of being addition and multiplication. Multiplication of a point P on an elliptic curve by an integer n is the result of adding a point to itself n times, so that

$$nP = P + P + \dots + P \quad (n \text{ times})$$

C. Elliptic curve encryption and decryption processes

As stated in [4], several approaches to encryption/decryption using elliptic curves are analyzed in the literature. A basic encryption process requires that a plaintext message m is encoded as a point P_m on the elliptic curve. The corresponding ciphertext C_m consists of a pair of points given by: $C_m = \{lG, P_m + kP_B\}$, where l is a random positive integer generated by the source A, G is the base point and $P_B = n_B G$ the point representing the public key of the destination B and n_B is the private key of the destination B. The decryption process is given by:

$$C_m \{P_m + lP_B\} - n_B(lG) = P_m + n_B lG - n_B lG = P_m$$

Most of the processing time consumed by the encryption/decryption processes is consumed by the computations related to lG and lP_B . Therefore, increasing the encryption/decryption rates requires to speed-up the computations related to point-multiplication.

IV. POINT ADDITION AND POINT DOUBLING IN AFFINE COORDINATES

A. Mathematical background

Additions on elliptic curve defined above, in section III.B., are controlled by the following rules:

$$O = -O \quad P(x, y) + O = P(x, y) \quad P(x, y) + P(x, -y) = O$$

The addition of two different points on the elliptic curve is computed as shown below:

$$P_1(x_1, y_1) + P_2(x_2, y_2) = P_3(x_3, y_3) \quad \text{where } x_1 \neq x_2$$

$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} \quad x_3 = m^2 - x_1 - x_2 \quad y_3 = m(x_1 - x_3) - y_1$$

The addition of a point to itself (point doubling) on the elliptic curve is computed as shown below:

$$P_1(x_1, y_1) + P_1(x_1, y_1) = P_3(x_3, y_3)$$

$$m = \frac{3x_1^2 + a}{2y_1} \quad x_3 = m^2 - 2x_1 \quad y_3 = m(x_1 - x_3) - y_1$$

m is the slope of the line intersecting the points on elliptic curve. Because of space shortage I decided not to derive the equations shown above, so you can find it in literature [2].

B. Algorithm of elliptic curve point addition

Input: $P_1(x_1, y_1)$, $P_2(x_2, y_2)$

Output: $P_3(x_3, y_3)$

- | | | |
|--|---|--------------------------------------|
| <ol style="list-style-type: none"> 1. If $x_1 = x_2$ return O 2. If $P_1 = P_2$ then 3. If $y_1 = 0$ return O 4. Else $m \leftarrow \frac{3x_1^2 + a}{2y_1}$ 5. Else $m \leftarrow \frac{y_2 - y_1}{x_2 - x_1}$ 6. $x_3 \leftarrow m^2 - x_1 - x_2$ 7. $y_3 \leftarrow m(x_1 - x_3) - y_1$ 8. Return $P_3(x_3, y_3)$ | } | Sequential part of algorithm |
| <ol style="list-style-type: none"> 6. $x_3 \leftarrow m^2 - x_1 - x_2$ 7. $y_3 \leftarrow m(x_1 - x_3) - y_1$ | } | Part that can be partly parallelized |

V. PARALLEL PROCESSING

A. Introduction

Traditionally, the vast majority of software applications are written as sequential programs, as an example we introduced an algorithm above. The execution of these programs can be understood by a human sequentially stepping through the code. Historically, computer users have become accustomed to the expectation that these programs run faster with each new generation of microprocessors. Such expectation is no longer strictly valid from this day onward. A sequential program will only run on one of the processor cores, which will not become significantly faster than those in use today. Without performance improvement, application developers will no longer be able to introduce new features and capabilities into their software as new microprocessors are introduced, thus reducing the growth opportunities of the entire computer industry.

In contrast, the many-core device focuses more on the execution throughput of parallel applications. The many-cores began as a large number of much smaller cores, and, once again, the number of cores doubles with each generation. By term of many-core devices we refer to general-purpose microprocessors or central processor units (CPUs), graphic processor units (GPUs), and other devices using more than one single core of processor. In this paper however we are dealing with GPUs and all of the operations that can be parallelized are performed on GPU.

The reason why we are trying to parallelize this algorithm is really simple. There is a really large performance gap between many-core GPUs and general-purpose multicore CPUs and the answer lies in the differences in the fundamental design philosophies between the two types of processors. The design of a CPU is optimized for sequential code performance. It makes use of sophisticated control logic to allow instructions from a single thread of execution to execute in parallel or even out of their sequential order while maintaining the appearance of sequential execution. More importantly, large cache memories are provided to reduce the instruction and data access latencies of large complex applications. Memory bandwidth is another important issue. Graphics chips have been operating at approximately 10 times the bandwidth of contemporaneously available CPU chips. On the other hand GPU takes advantage of a large number of execution threads to find work to do when some of them are waiting for long-latency

memory accesses, thus minimizing the control logic required for each execution thread. Small cache memories are provided to help control the bandwidth requirements of these applications. As a result, much more chip area is dedicated to the floating-point calculations.

B. Amdahl's law

One should expect that most applications will use both CPUs and GPUs, executing the sequential parts on the CPU and numerically intensive parts on the GPUs. In what manner and what range we can parallelize our algorithm gives us answer an Amdahl's law. It considers that in order to improve performance (time to complete the process), only a part of the entire process is usually improved. It shows that the improvement is limited by the fraction of time the improved part is used. It can be stated in terms of two parameters: f , the fraction of time spent in the improved part, and the speedup S , how many times faster the improved part performs. Amdahl's Law gives how many times faster the entire process will be as:

$$S_{overall} = \frac{1}{(1-f) + \frac{f}{S}}$$

Using absolute execution times, Amdahl's law is in terms of T_{after} , the execution time after an improvement; $T_{improved}$, the execution time affected by the improvement S , how many times faster the improved part runs, or its *speedup*; and $T_{unaffected}$, the execution time unaffected by the improvement. In these terms, Amdahl's Law states that:

$$T_{after} = \frac{T_{improved}}{S} + T_{unaffected}$$

VI. PARALLEL IMPLEMENTATION

Let's discuss how we can parallelize the algorithm of point addition shown in upper section. As a first step we have to decide which parts of computations can be divided into parts by means of cores in a device (GPU). Breaking the whole computation can be done according to adding point algorithm:

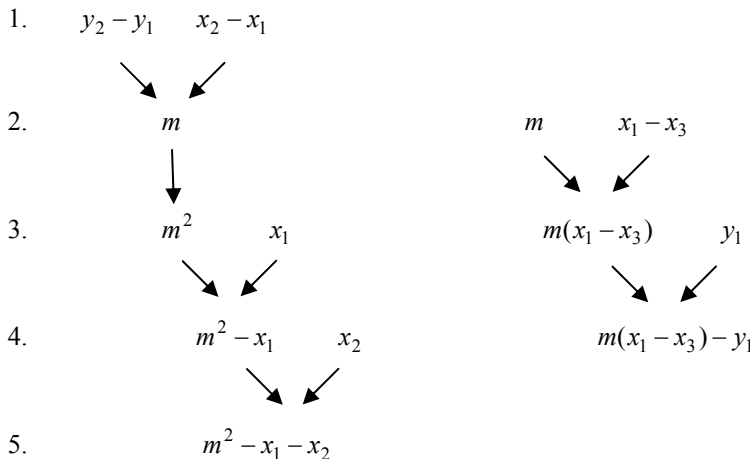
Steps 1. through 3. are fully sequential, because of deciding parts of algorithm. Thus, all operations are done on CPU.

Steps 4. through 7. are numerically intensive parts so they can be partly parallelized. The next table shows a cost of operations in affine coordinates.

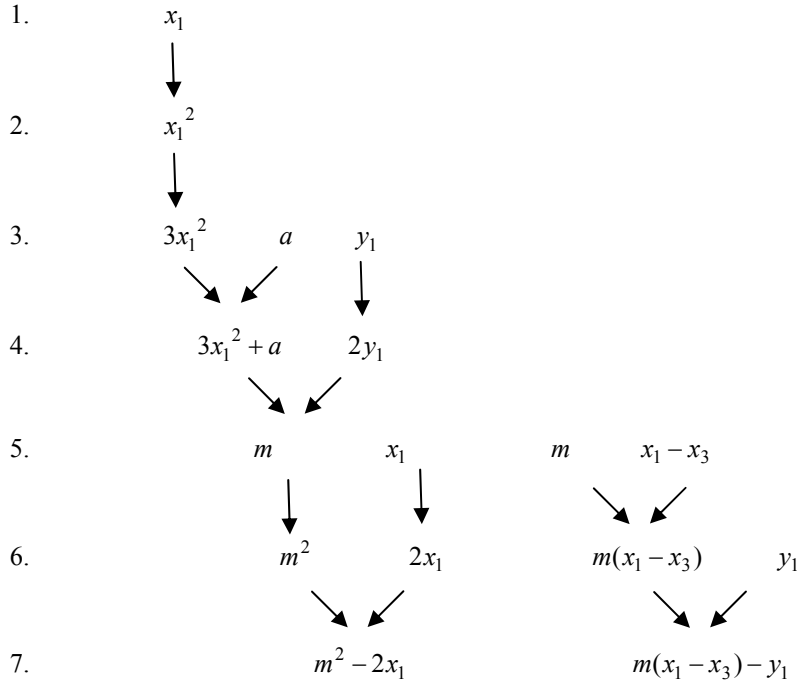
Table 1
Cost of ECC operations

Operations	Point addition	Point doubling
Additions	6	4
Multiplications	2	4
Inversions	1	1
Total	6A+3M+1I	4A+4M+1I

A. Point adding dependency graph



B. Point doubling dependency graph



The thing that is very important to say is that the operations are not equal in terms of computational time and performance. So before we compute the exact speedup and time consumption of parallel computation we have to evaluate the particular operations. Three operations are mentioned at the beginning of this section. According to [5] the ratios between these operations are very simple. When we consider the computation cost of adding and multiplying as 1, because there is no exact notation about the ratio between them in any literature, than the ratio between adding (or multiplying) and inversion is estimated by ratio of about 80. At this point we can see that this value is quite large in comparison to the number of required operations.

In point adding graph there are exactly three sets of operations that are parallelized, so instead of 9 operations in sequential algorithm we have 5 steps that multicore device needs to run. Let's evaluate the whole process. In first step we make three operations in one cycle instead of doing three separate operations in three different cycles so reached 3 times faster algorithm at this point. However calculation of the slope includes one inversion that cannot be parallelized with other operations because the calculation of point coordinates are dependent with the slope m . In the other two steps (3., 4.) we get twice 2 times faster algorithm in comparison of sequential variant. The final step is not parallelized.

In terms of performance we reached 7 from 9 operations in parallel. That means using 3 cores at the moment at most, from which we get:

f - 77.77 % (7 from 9 operations parallelized)

S - 3 (cores used at one cycle)

$$S_{overall} = \frac{1}{(1-f) + \frac{f}{S}} = \frac{1}{(1-0.7777) + \frac{0.7777}{3}} = 2.0470$$

Speedup overall system in adding point algorithm is 2.047 times faster than it would be in sequential execution of program. It is necessary to say that speedup equation doesn't regard the fact that inversion cost 80 times more than simple addition or multiplying. However, this fact is very important in means of parallel computing. Let's see what is the real time needed to perform this algorithm:

$$T_{after} = \frac{T_{improved}}{S} + T_{unaffected} = \frac{7}{3} + 81 = 83.3333 \text{ cycles}$$

From 88 cycles (80 for inversion 8 for addition and multiplying) we get 83.333 which is just 5.303 % improvement compared to sequential executing program.

Because of shortage of this paper point doubling is not considered, but the speedup improvement is computed exactly the same way.

VII. CONCLUSION

As we can see at the example of the parallelism above, addition algorithm over elliptic curve can be easily parallelized in affine coordinates. Though, before applying parallel algorithms on existing sequential ones, it is necessary to consider what speedups and time improvements we can get from this rearranging. First of all, affine coordinates are not really good in terms of fast computation on sequential devices. As I mentioned before, this is because of inversion operation applied to evaluate the slope of the line intersecting points. However this fact is limiting in parallel computation too and it cannot be bypassed by any other operation in affine coordinates. From these results it could be, in most of the cases, contra-productive to use massively parallel devices to evaluate points on elliptic curves in affine coordinates.

ACKNOWLEDGMENT

This work is the result of the project implementation: Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120020) supported by the Research & Development Operational Program funded by the ERDF. (EN) - príklad kombinovaného pod'akovania This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0080-09 (10%); the Slovak Cultural and Educational Grant Agency of Ministry of Education of Slovak Republic under the contract No. 3/7523/09 (20%). This work is also the result of the project implementation Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120020) supported by the Research & Development Operational Program funded by the ERDF (70%).

REFERENCES

- [1] Certicom Research, "Certicom ECC Challenge," (Published work style), Latest update November 2009 pp. 6–8.
- [2] L.Martin, *Introduction to identity-based encryption* (Book style), Norwood, MA: Artech house, 2008.
- [3] H.Khali, A. Farah, *Cost-Effective Implementations of GF (p) Elliptic Curve Cryptography Computations*. (Published work style), International Journal of Computer Science and Network Security. Vol.7 No.8, August 2007.
- [4] W. Stallings, *Cryptography and network security* (Book style) Prentice Hall; 5 edition, January 2010.
- [5] D. Kirk, W. Hwu, *Programming massively parallel processors* (Book style), Elsevier, Burlington, MA, 2005.
- [6] J. Sanders, E. Kandrot, *CUDA by example* (Book style), Addison-Wesley, Boston, MA, July 2010.

Parallel Processing Applied on Elliptic Curve Arithmetic in Jacobian Projective Coordinates

Patrik GALLO, Peter GOČ-MATIS, Dušan LEVICKÝ

Dept. of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

patrik.gallo@tuke.sk, peter.goc-matis@tuke.sk, dusan.levicky@tuke.sk

Abstract — In this paper we present a view over elliptic curve arithmetic with regard to parallel processing on multiprocessor units in general. We provide an analysis of how optimal or not it can be using parallel processors instead of single processor units. Amdahl's law is considered as a first step in this analysis. Analysis includes theoretical introduction to elliptic curve cryptosystem, especially to elliptic curve arithmetic with regard to Jacobian projective coordinate system. Analysis is concluded at the end of the paper.

Keywords — elliptic curve arithmetic, parallel processing, Jacobian projective coordinates

I. INTRODUCTION

The theory of doing computational work in parallel has some fundamental laws that place limits on the benefits one can derive from parallelizing a computation. To understand these laws, we have to first define the objective. In general, the goal in large scale computation is to get as much work done as possible in the shortest possible time. In a perfect world, a computational job that is split up among N processors would complete in $1/N$ time, leading to an N -fold increase in power. However, any given piece of parallelized work to be done will contain parts of the work that *must* be done serially, one task after another, by a single processor. This part does *not* run any faster on a parallel collection of processors (and might even run more slowly). Only the part that can be parallelized runs as much as N -fold faster.

In case of this research paper, the object of parallelism is a part of computation over elliptic curve arithmetic. By elliptic curve arithmetic we mean operations on elliptic curve defined over some finite field. Thus, algebraically it means performing operations over several coordinate system, for example affine, projective, Jacobian projective, modified Jacobian, Chudnovsky, and so on. The aim of this paper is to analyze a parallelism in Jacobian projective coordinate system, but the theoretical basis holds for all the coordinate systems.

II. ELLIPTIC CURVE ARITHMETIC

A. Definition of an elliptic curve

An elliptic curve is a set of points satisfying an equation of the form $y^2 = x^3 + ax + b$ where the coefficients a and b are elements of a field F with the characteristic of F is not equal to 2 or 3. Detailed explanation why are we using an equation of this form and why the characteristic of a field should not be equal to 2 or 3 is shown in [2]. We write E/F to indicate that the elliptic curve is over the field. Such a curve is said to be in *Weierstrass normal form*. We can think of the points on an elliptic curve as being either points in a set or as rational functions of x and y , and can freely change between the two points of a view as needed. The requirement that the characteristic of F be greater than 3 is not strictly required for an elliptic curve, but this restriction limits us to the elliptic curves of interest in the discussion of secure cryptographic systems. If the characteristic of the field over which an elliptic curve is defined is equal to 2 or 3, alternative forms other than Weierstrass normal form need to be used. In particular, the algorithm for adding points on an elliptic curve in Weierstrass normal form uses the constants 2 and 3, which makes it behave badly over a field of characteristic 2 or 3, so alternate forms are needed in these cases.

Elliptic curves for which the cubic has repeated roots turn out to have undesirable properties from the cryptographic point of view, so it is useful to be able to identify which curves these are. To determine this, we can find the roots of the cubic part of an elliptic curve in Weierstrass normal form by factoring $x^3 + ax + b = 0$ into $(x - x_1)(x - x_2)(x - x_3)$ and finding values of x_1 , x_2 and x_3 . This process is quite difficult and it's shown in [2]. Using these values for roots of the cubic we find that

$$(x_1 - x_2)^2(x_1 - x_3)^2(x_2 - x_3)^2 = -(4a^3 + 27b^2)$$

The right side of the equation is called the *discriminant* of an elliptic curve. So that we have a repeated root whenever $-(4a^3 + 27b^2) = 0$.

B. Algebraic structure of elliptic curves

Points on an elliptic curve provide a structure that can define in the terminology of abstract algebra.

If E is an elliptic curve over field F then we write $E(F)$ to indicate the set of points on E along with the operation of adding points described later in algorithm.

If F is a field and E is an elliptic curve then $E(F)$ is a group. The point at infinity O acts as identity element for this group. Note that there is only operation defined for $E(F)$, which we are thinking of as addition, so it is impossible to multiply or divide elements of $E(F)$. Thus, $E(F)$ cannot be a field, which requires two operations that we think of being addition and multiplication. Multiplication of a point P on an elliptic curve by an integer n is the result of adding a point to itself n times, so that

$$nP = P + P + \dots + P \quad (n \text{ times})$$

C. Elliptic curve encryption and decryption processes

As stated in [4], several approaches to encryption/decryption using elliptic curves are analyzed in the literature. A basic encryption process requires that a plaintext message m is encoded as a point P_m on the elliptic curve. The corresponding ciphertext C_m consists of a pair of points given by: $C_m = \{lG, P_m + kP_B\}$, where l is a random positive integer generated by the source A, G is the base point and $P_B = n_B G$ the point representing the public key of the destination B and n_B is the private key of the destination B. The decryption process is given by:

$$C_m \{P_m + lP_B\} - n_B(lG) = P_m + n_B lG - n_B lG = P_m$$

Most of the processing time consumed by the encryption/decryption processes is consumed by the computations related to lG and lP_B . Therefore, increasing the encryption/decryption rates requires to speed-up the computations related to point-multiplication.

III. POINT ADDITION AND POINT DOUBLING IN JACOBIAN PROJECTIVE COORDINATES

A. Projective coordinates

Dealing with O , the point at infinity on an elliptic curve can be troublesome using affine coordinates, the usual (x, y) coordinates that we use to define the Weierstrass normal form of an elliptic curve. One easy way to handle this point is through the use of projective coordinates. Projective coordinates encode a point (x, y) with two coordinates in three coordinates (x, y, z) where (x, y, z) represent any point of the form $(x/z, y/z)$. Such projective coordinates are called *standard projective coordinates*. In particular, we can represent a point on an elliptic curve $P=(x, y)$ as $(x, y, 1)$ and the point at infinity can be represented by the coordinates $(0, 1, 0)$. We can also easily convert from projective coordinates (x, y, z) where $z \neq 0$ into affine coordinates $(x/z, y/z)$.

In addition to being an easy way to handle the point at infinity, projective coordinates are often useful in performing computations on elliptic curves because it is possible to add two points on an elliptic curve using projective coordinates without performing any divisions, which are typically very expensive computationally in finite fields. In cryptographic applications where we may want to perform operations in F_q for large values of q , determining the inverse of an element of finite field can be fairly expensive relative to multiplication and using projective coordinates will often provide a performance advantage over using affine coordinates. *Jacobian projective coordinates* encode an affine point $(x/z^2, y/z^3)$ as the

projective point (x, y, z) . In general, each type of projective coordinates requires a different number of field operations to add or double points which is summarized in the table 1. The choice of the most efficient projective coordinate system will depend on the application.

Table 1
Cost of ECC operations

Coordinate system	Point addition	Point doubling
Jacobian	12M+4S	4M+6S
Standard	12M+2S	7M+5S
Chudnovsky	11M+3S	5M+6S

B. Adding points in Jacobian projective coordinates

Additions on elliptic curve defined above in section III.B. are controlled by the following rules:

$$O = -O \quad P(x, y) + O = P(x, y) \quad P(x, y) + P(x, -y) = O$$

The addition of two different points on the elliptic curve is computed as shown below:

$$P_1(x_1, y_1, z_1) + P_2(x_2, y_2, z_2) = P_3(x_3, y_3, z_3) \text{ where } x_1 \neq x_2 \text{ and } z_3 \neq 0$$

$$\begin{aligned} u_1 &= x_1 z_2^2 & u_2 &= x_2 z_1^2 \\ s_1 &= y_1 z_2^3 & s_2 &= y_2 z_1^3 \\ h &= u_2 - u_1 & r &= s_2 - s_1 \\ x_3 &= r^2 - h^3 - 2u_1 h^2 & y_3 &= r(u_1 h^2 - x_3) - s_1 h^3 & z_3 &= h z_1 z_2 \end{aligned}$$

The addition of a point to itself (point doubling) on the elliptic curve is computed as shown below:

$$\begin{aligned} P_1(x_1, y_1, z_1) + P_1(x_1, y_1, z_1) &= P_3(x_3, y_3, z_3) \\ s &= 4x_1 y_1^2 & m &= 3x_1^2 + a z_1^4 \\ x_2 &= m^2 - 2s & y_2 &= m(s - x_2) - 8y_1^4 & z_2 &= 2y_1 z_1 \end{aligned}$$

Because of space shortage I decided not to derive the equations shown above, so you can find it in literature [2].

C. Algorithm of elliptic curve point addition

Input: $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$

Output: $P_3(x_3, y_3, z_3)$

- | | | |
|---|---|--------------------------------------|
| <ol style="list-style-type: none"> 1. $u_1 \leftarrow x_1 z_2^2$ 2. $u_2 \leftarrow x_2 z_1^2$ 3. $s_1 \leftarrow y_1 z_2^3$ 4. $s_2 \leftarrow y_2 z_1^3$ | } | Part that can be partly parallelized |
| <ol style="list-style-type: none"> 5. If $u_1 = u_2$ 6. If $s_2 \neq s_1$ 7. Return (0,1,0) 8. Else 9. Return Jacobian Double (x_1, y_1, z_1) | } | Sequential part of algorithm |
| <ol style="list-style-type: none"> 10. $h \leftarrow u_2 - u_1$ 11. $r \leftarrow s_2 - s_1$ 12. $x_3 \leftarrow r^2 - h^3 - 2u_1 h^2$ 13. $y_3 \leftarrow r(u_1 h^2 - x_3) - s_1 h^3$ 14. $z_3 \leftarrow h z_1 z_2$ 15. Return $P_3(x_3, y_3, z_3)$ | } | Part that can be partly parallelized |

IV. PARALLEL PROCESSING

A. Introduction

Traditionally, the vast majority of software applications are written as sequential programs, as an example we introduced an algorithm above. The execution of these programs can be understood by a human sequentially stepping through the code. Historically, computer users have become accustomed to the expectation that these programs run faster with each new generation of microprocessors. Such expectation is no longer strictly valid from this day onward. A sequential program will only run on one of the processor cores, which will not become significantly faster than those in use today. Without performance improvement, application developers will no longer be able to introduce new features and capabilities into their software as new microprocessors are introduced, thus reducing the growth opportunities of the entire computer industry.

In contrast, the many-core device focuses more on the execution throughput of parallel applications. The many-cores began as a large number of much smaller cores, and, once again, the number of cores doubles with each generation. By term of many-core devices we refer to general-purpose microprocessors or central processor units (CPUs), graphic processor units (GPUs), and other devices using more than one single core of processor. In this paper however we are dealing with GPUs and all of the operations that can be parallelized are performed on GPU.

The reason why we are trying to parallelize this algorithm is really simple. There is a really large performance gap between many-core GPUs and general-purpose multicore CPUs and the answer lies in the differences in the fundamental design philosophies between the two types of processors. The design of a CPU is optimized for sequential code performance. It makes use of sophisticated control logic to allow instructions from a single thread of execution to execute in parallel or even out of their sequential order while maintaining the appearance of sequential execution. More importantly, large cache memories are provided to reduce the instruction and data access latencies of large complex applications. Memory bandwidth is another important issue. Graphics chips have been operating at approximately 10 times the bandwidth of contemporaneously available CPU chips. On the other hand GPU takes advantage of a large number of execution threads to find work to do when some of them are waiting for long-latency memory accesses, thus minimizing the control logic required for each execution thread. Small cache memories are provided to help control the bandwidth requirements of these applications. As a result, much more chip area is dedicated to the floating-point calculations.

B. Amdahl's law

One should expect that most applications will use both CPUs and GPUs, executing the sequential parts on the CPU and numerically intensive parts on the GPUs. In what manner and what range we can parallelize our algorithm gives us answer an Amdahl's law. It considers that in order to improve performance (time to complete the process), only a part of the entire process is usually improved. It shows that the improvement is limited by the fraction of time the improved part is used. It can be stated in terms of two parameters: f , the fraction of time spent in the improved part, and the speedup S , how many times faster the improved part performs. Amdahl's Law gives how many times faster the entire process will be as:

$$S_{\text{overall}} = \frac{1}{(1-f) + \frac{f}{S}}$$

Using absolute execution times, Amdahl's law is in terms of T_{after} , the execution time after an improvement; T_{improved} , the execution time affected by the improvement S , how many times faster the improved part runs, or its *speedup*; and $T_{\text{unaffected}}$, the execution time unaffected by the improvement. In these terms, Amdahl's Law states that:

$$T_{\text{after}} = \frac{T_{\text{improved}}}{S} + T_{\text{unaffected}}$$

V. PARALLEL IMPLEMENTATION

Let's discuss how we can parallelize the algorithm of point addition shown in upper section. As a first step we have to decide which parts of computations can be divided into parts by

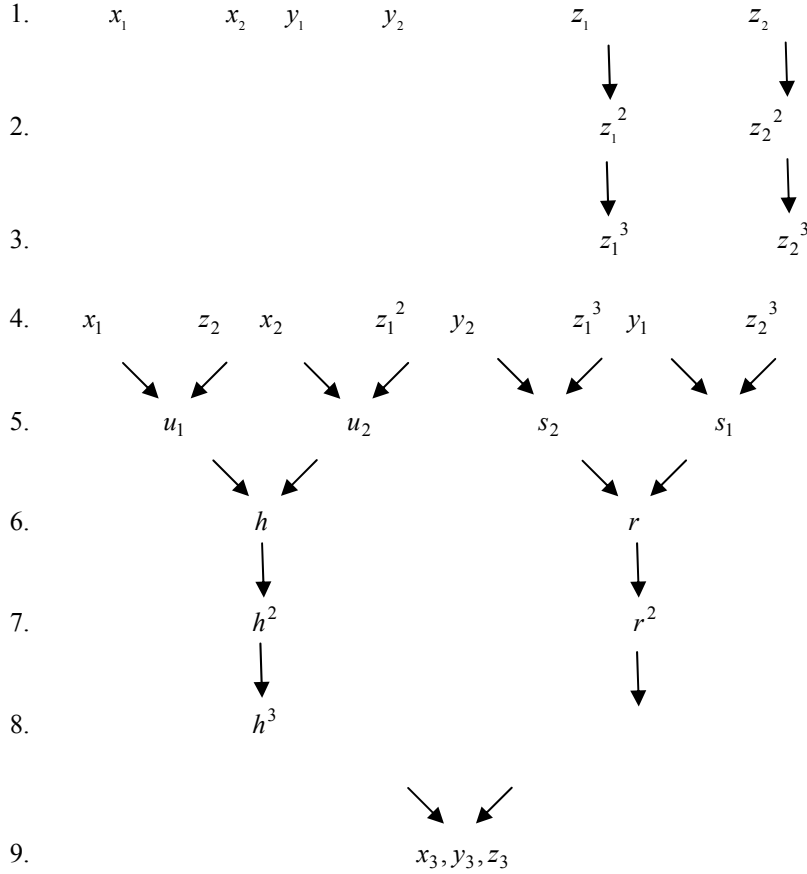
means of cores in a device (GPU). Breaking the whole computation can be done according to adding point algorithm:

Steps 1. through 4. are numerically intensive parts so they can be partly parallelized.

Steps 5. through 9. are fully sequential, because of deciding parts of algorithm. Thus, all operations are done on CPU.

Steps 10. through 14. are numerically intensive parts so they can be partly parallelized.

A. Point adding dependency graph



The thing that is very important to say is that the operations are equal in terms of computational time and performance. Two operations are mentioned at the beginning of this section. According to [5] the ratios between these operations are equal to 1. At this point we can see that there is not any operation that is quite large in comparison to two other operations. This is one of the big differences between affine and projective coordinate systems.

In point adding graph there are exactly two sets of operations that are parallelized, so instead of 14 operations in sequential algorithm we have 7 steps that multicore device needs to run. Let's evaluate the whole process. In first step we make squaring of two coordinates which is done in one cycle. Another cycle is similar and does the same with the same couple of coordinates. In another step parallel algorithm makes four operations in one cycle, which is the biggest speedup in this program. Four other steps are quite simple and are parallelized as it is shown on the graph above.

In terms of performance we reached 7 from 14 operations in parallel. That means using 3 cores at the moment at most, from which we get:

f - 85.72 % (12 from 14 operations parallelized)

S - 4 (cores used at one cycle)

$$S_{overall} = \frac{1}{(1-f) + \frac{f}{S}} = \frac{1}{(1-0.8572) + \frac{0.8572}{4}} = 2.8$$

Speedup overall system in adding point algorithm is 2.8 times faster than it would be in sequential execution of program. Let's see what is the real time needed to perform this algorithm, because this is really the thing that interested us:

$$T_{after} = \frac{T_{improved}}{S} + T_{unaffected} = \frac{8}{2} + \frac{4}{4} + 2 = 7 \text{ cycles}$$

↑

(8 cycles with 2 cores and 4 cycles with 4 cores)

From 14 cycles we get 7 which is 50 % improvement compared to sequential executing program.

Because of shortage of this paper point doubling is not considered, but the speedup improvement is computed exactly the same way.

VI. CONCLUSION

As we can see at the example of the parallelism above, addition algorithm over elliptic curve can be easily parallelized in Jacobian projective coordinates. Though, before applying parallel algorithms on existing sequential ones, it is necessary to consider what speedups and time improvements we can get from this rearranging. The big problem of affine coordinate system which is inversion operation is fully bypassed by introducing Jacobian coordinates which uses just additions and multiplication, that are incomparable faster operations. At that point, the parallelism of addition formula over elliptic curve became really effective in terms of speedups and time improvement. The time improvement of this algorithm is equal to 50% just because of parallelism. From the viewpoint of security implementation, elliptic curve cryptosystems are not just considered as most powerful and secure public key cryptosystem, but can get more and more effective regarding to using massively parallel devices.

ACKNOWLEDGMENT

This work is the result of the project implementation: Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120020) supported by the Research & Development Operational Program funded by the ERDF. (EN) - príklad kombinovaného pod'akovania This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0080-09 (10%); the Slovak Cultural and Educational Grant Agency of Ministry of Education of Slovak Republic under the contract No. 3/7523/09 (20%). This work is also the result of the project implementation Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120020) supported by the Research & Development Operational Program funded by the ERDF (70%).

REFERENCES

- [1] Certicom Research, "Certicom ECC Challenge," (Published work style), Latest update November 2009 pp. 6–8.
- [2] L.Martin, *Introduction to identity-based encryption* (Book style), Norwood, MA: Artech house, 2008.
- [3] H.Khali, A. Farah, *Cost-Effective Implementations of GF(p) Elliptic Curve Cryptography Computations*. (Published work style), International Journal of Computer Science and Network Security. VOL.7 No.8, August 2007.
- [4] W. Stallings, *Cryptography and network security* (Book style) Prentice Hall; 5 edition, January 2010.
- [5] D. Kirk, W. Hwu, *Programming massively parallel processors* (Book style), Elsevier, Burlington, MA, 2005.
- [6] J. Sanders, E. Kandrot, *CUDA by example* (Book style), Adison-Wesley, Boston, MA, July 2010.

Obrazová steganografia v komunikačných systémoch

*Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ,
Zita KLENOVIČOVÁ, Dušan LEVICKÝ*

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

vladimir.banoci@tuke.sk, gabriel.bugar@tuke.sk, tomas.kanocz@tuke.sk,
zita.klenovicova@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Rozmach dátových sietí akým je napríklad Internet, rýchly pokrok IT technológií sa stáva každodennou súčasťou života bežného človeka. Priama závislosť na týchto technológiách a ich globálneho nasadenie v reálnych podmienkach vyvstáva otázku potenciálu ich zneužitia na sekundárne účely akou je možnosť podprahovej tajnej komunikácie. Spôsob tejto utajenej komunikácie nie je zaujímavý iba z hľadiska legálneho charakteru akým je napríklad oblasť digitálnych vodoznakov, ktorej predmetom je prioritne ochrana duševného vlastníctva, ale taktiež z hľadiska nelegálnej komunikácie.

Predmetom tohto príspevku je oblasť steganografie, ktorá pokrýva problematiku utajovania informácií v komunikačných systémoch. Jednotlivé časti príspevku poukazujú na princípy a požiadavky kladené na steganografické systémy. V závere je uvedený stručný prehľad aktuálnych a najviac používaných steganografických metód.

Kľúčové slová — steganografia, ukrývanie informácií, tajná správa, tajný kľúč, podprahový kanál

I. ÚVOD

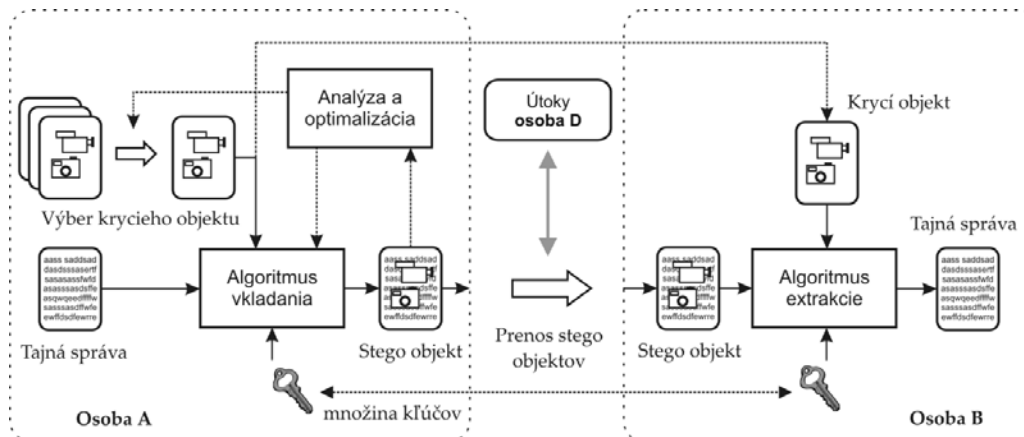
Jednou z najväčších technologických udalostí posledných dvoch desaťročí bola nepochybne „invázia“ digitálnych médií takmer do každej oblasti nášho života. Digitálne dáta (aké sú napr. textové dokumenty, statické obrazy, dynamické obrazy resp. video, audio, atď.) spolu s multimediálnymi dokumentmi sa v súčasnosti stávajú dominantnými v oblasti zábavy, umenia, výučby, atď. Na tejto skutočnosti sa do značnej miery podieľajú ich vlastnosti, akými sú napr. efektívne uloženie a ich jednoduchá manipulácia pomocou výpočtovej techniky, ako aj ich rýchly a jednoduchý prenos pomocou dátových komunikačných sietí. Ako výsledok tohto trendu vzniklo množstvo nových aplikácií (video a audio na požiadanie, služby *Triple Play* a množstvo ďalších sieťových multimediálnych aplikácií), ktoré sa neustále rozrastajú a veľkou mierou prispievajú k rastu elektronického obchodu [1].

Steganografia predstavuje vysoký stupeň ochrany informácií a definuje metódy ukrývania dát v iných dátach, ktoré sa takto stávajú pre náhodného pozorovateľa nevnímateľné, teda ide o samotné utajenie existencie komunikácie (na rozdiel od kryptografie, kde ide primárne o utajenie obsahu informácie, pričom skutočnosť, že prebieha komunikácia je všeobecne známa). Steganografiu spolu s digitálnymi vodoznakmi je možné zaradiť pod termín ukrývanie informácií.

V priebehu posledného desaťročia boli navrhnuté a overené viaceré metódy vkladania tajných informácií do multimediálnych dát a vyčlenili sa viaceré typy techník, ktoré sú využívané mnohými steganografickými systémami. V súčasnosti sú dostupné mnohé publikácie [3][4][5][7], v ktorých sú uvedené základné pojmy, princípy a požiadavky kladené na steganografické systémy. Postupným vývojom a vylepšovaním jednotlivých techník utajenia komunikácie sa navrhujú dômyselnejšie steganografické metódy. Tieto metódy využívajú rozličné transformácie krycích dát v procese vkladania informácií [3]. Steganografické metódy využívajúce transformačnú oblasť sú stále predmetom početných výskumov a sú neustále vo vývoji. Využívajú sa rôzne typy transformácií a medzi najviac používané patria DCT, DFT a DWT. Pravdepodobne, najviac steganografických metód v transformačnej oblasti bolo navrhnutých na báze DCT. Tento jav zrejme ovplyvnil aj fakt, že DCT má rýchly algoritmus výpočtu a transformačné koeficienty sú reálne čísla.

II. STEGANOGRAFICKÝ MODEL

Základný model utajenej (neviditeľnej) komunikácie bol prvý krát navrhnutý *Simmonsom* a bol označený ako „*Väzenský problém*“. Základná myšlienka je založená na predpoklade, že dvaja väzni osoba A a osoba B sú uväznení vo väzení, pričom navzájom na seba nevidia a už vôbec sa nepočujú. Každá komunikácia medzi nimi je monitorovaná a následne analyzovaná dozorcami *D*. Ten im nedovolí vymieňať šifrované správy a taktiež hrozí, že pri akomkoľvek podozrení im bude zakázaná vzájomná komunikácia [9].



Obr. 1 Základný model steganografického systému

Väzni si teda musia vymieňať informácie bez vzbudenia podozrenia a musia vytvoriť tzv. „podprahový kanál“. Najjednoduchší spôsob ako to dosiahnuť je ukrytie tajnej správy v inej, podozrenie nevzbudzujúcej správe. Základný model steganografického systému je uvedený na Obr. 1.

Vo všeobecnosti rozoznávame dve základné skupiny systémov steganografie v závislosti od množstva informácie potrebnej pre úspešnú extrakciu utajenej informácie:

1. *systémy s prenosom dodatočnej informácie* - pri ich použití je nutná prítomnosť identickej kópie originálnych krycích dát v procese extrakcie utajenej informácie na strane prijímača.
2. *systémy bez prenosu dodatočnej informácie* - tieto systémy nevyžadujú prítomnosť originálnych krycích dát na strane prijímača a označujú sa ako tzv. „nezávislé“, resp. „slepé“ systémy. Tieto môžeme ďalej rozdeliť na:
 - *systémy s estimáciou krycích dát* – originálne krycie dáta sa získavajú estimáciou prijatých označených dát. Ich použitie je typické v prípade aplikácie techník s využitím systémov s rozprestretým spektrom.
 - *systémy bez estimácie krycích dát* – v ich prípade sa extrakcia tajnej informácie uskutočňuje iba z prijatých označených dát.

Procesy a úpravy na krycích objektoch sa musia vykonávať tak, aby v žiadnom prípade nevznikalo podozrenie, že označený (stego) objekt obsahuje skrytú informáciu. V dokonalom steganografickom systéme nesmie byť badateľný rozdiel medzi krycím a označeným objektom postrehnuteľný ľudským vnímaním alebo odhalený štatistickými metódami [8].

A. Hľadiská hodnotenia steganografických systémov

Steganografické systémy možno hodnotiť z viacerých hľadísk. Sú to najmä:

- Hľadisko vnímateľnosti
 - *vnímateľný steganografický systém* – vloženie tajnej správy je možné vnímať zmyslami, napr. vložené symboly, emblémy alebo objekty, ktoré dotvárajú zmysel krycieho objektu,
 - *nevnímateľný steganografický systém* – vložená tajná správa v stego objekte nie je vnímateľná zmyslami. Na jej úspešnú extrakciu je potrebné ju digitálne spracovať.
- Hľadisko odolnosti
 - *robustný steganografický systém* – v ktorom tajná správa po vložení do krycieho objektu by mala odolávať rôznym modifikáciám, ktoré sú spôsobené bežnou manipuláciou, prípadne útokmi snažiacimi sa o odhalenie, resp. prípadné zamedzenie komunikácie,
 - *krhký steganografický systém* – v ktorom vložená tajná správa podlieha zmenám tak isto ako stego objekt a je poškodená každým útokom. V niektorých prípadoch možno tento systém využiť na detekciu integrity dát. V prípade, že sa zo stego objektu podarí vyextrahovať tajnú správu, potom integrita dát nebola po vložení porušená, a v opačnom prípade bola porušená integrita dát a je možné konštatovať útok na stego objekt.

- Hľadisko závislosti
 - *závislý steganografický systém* – v ktorom na zistenie prítomnosti, resp. extrakcie tajnej správy sú potrebné originálne krycie alebo referenčné dáta,
 - *nezávislý (slepý) steganografický systém* – v ktorom na zistenie prítomnosti, alebo extrakcie tajnej správy nie sú potrebné originálne krycie objekty a ani referenčné dáta.
- Hľadisko dostupnosti
 - *čistý steganografický systém* – v ktorom pri vkladani alebo extrakcii tajnej správy sa nepoužíva žiadny kľúč,
 - *symetrický steganografický systém* – v ktorom na vkladanie a aj extrakciu tajnej správy sa používa ten istý tajný kľúč,
 - *asymetrický steganografický systém* – používa dva kľúče, jeden kľúč na vkladanie a druhý na extrakciu, resp. detekciu správy.
- Hľadisko domény
 - steganografické systémy pracujúce v *priestorovej oblasti (spatial-domain)*, v ktorej vkladanie tajnej správy je realizované priamo do krycieho objektu v obrazovom priestore,
 - steganografické systémy pracujúce v *transformačnej oblasti (transform-domain)*, v ktorých sa krycí objekt najprv pretransformuje do vhodnejšej bázy resp. transformačnej oblasti, kde sa realizuje samotný proces vkladania tajnej správy.
- Hľadisko druhu krycích objektov
 - bitmapové obrazy, video sekvencie, hudba obsahujú veľa redundantných dát, a z toho vyplýva, že je pomerne jednoduché do takýchto krycích objektov ukryť tajné správy,
 - vektorové obrazy obsahujú oveľa menej redundantných dát ako porovnateľné bitmapové a preto je veľmi ťažké do nich ukrývať tajné správy. Do tejto oblasti patria aj vektorové mapy,
 - textové dáta v ktorých modifikáciou umiestnenia písmen, tvaru, medzery medzi slovami, riadkami, resp. zmiešavaním slov za ich synonymá prebieha samotné vkladanie a následná extrakcia tajných správ [4][8].

III. ZÁKLADNÉ DEFINÍCIE SYSTÉMOV STEGANOGRRAFIE

V oblasti steganografie možno definovať tri základné typy systémov. Sú to:

- čistá steganografia (*pure steganography*),
- symetrická steganografia so súkromným kľúčom (*secret key steganography*),
- asymetrická steganografia s verejným kľúčom (*public key steganography*).

A. Čistá steganografia

Je steganografia, ktorá ukrýva tajnú správu m (*secret message*) bez pomoci kľúča k (*stego key*). Je definovaná ako množina prvkov $\gamma = \{C, M, C_m, E, D\}$, kde C predstavuje množinu všetkých krycích objektov, M predstavuje množinu všetkých tajných správ, ktoré musia spĺňať podmienku $|C| \geq |M|$, C_m predstavuje množinu všetkých stego objektov, E predstavuje zobrazenie $E: C \times M \rightarrow C_m$ (*funkcia vkladania*) a D predstavuje zobrazenie $D: C_m \rightarrow M$ (*funkcia extrakcie*) s vlastnosťou $D(E(c, m)) = m$ pre všetky $m \in M$ a $c \in C$.

Odosielateľ resp. príjemca musí mať prístup k algoritmu vkladania resp. extrakcie. Tieto algoritmy by ale nemali byť verejne dostupné, keďže bezpečnosť tohto systému závisí jedine od utajenia týchto algoritmov.

B. Symetrická steganografia s tajným kľúčom

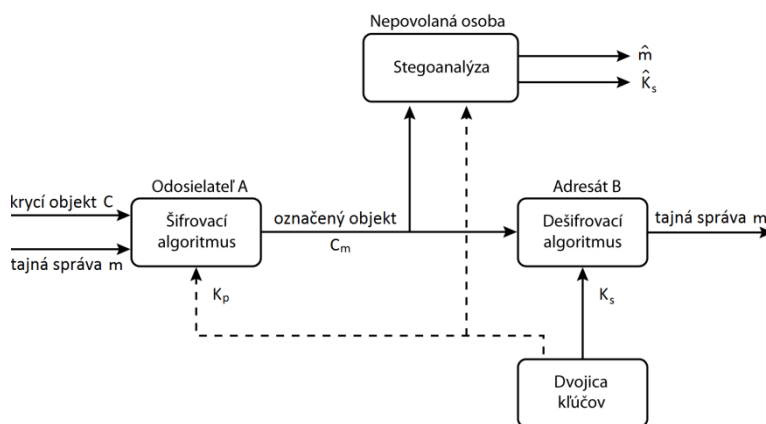
Tento typ steganografického systému je veľmi podobný princípu symetrickej šifry. Je definovaný ako množina prvkov $\gamma = \{C, M, K, C_m, E_K, D_K\}$ kde C predstavuje množinu všetkých možných krycích objektov, M predstavuje množinu všetkých tajných správ, pre ktoré musí byť opäť splnená podmienka $|C| \geq |M|$, K je súbor tajných kľúčov, C_m predstavuje množinu všetkých stego objektov a E_K, D_K predstavujú zobrazenia $E_K: C \times M \times K \rightarrow C_m$ a $D_K: C_m \times K \rightarrow M$ s vlastnosťou $D_K(E_K(c, m, k), k) = m$ pre všetky $m \in M$ a $c \in C$ a $k \in K$. Odosielateľ v procese vkladania tajnej informácie využíva okrem krycieho objektu c z C a tajnej správy m z M aj tajný kľúč k z K . Tento kľúč, ktorý predstavuje ďalší parameter úspešného vkladania tajnej informácie, musí vlastniť aj prijímateľ k úspešnej extrakcii. Bezpečnosť tohto systému spočíva práve v utajení a dĺžke kľúča k . Slabou stránkou tejto koncepcie je problém výmeny kľúča medzi komunikujúcimi stranami. Jednou z možností riešenia tohto problému je výmena tajného kľúča pred začatím samotnej komunikácie prostredníctvom iného kanála, ktorá znižuje riziko odhalenia samotnej komunikácie. Ak proces

vkladania nemení vlastnosti krycieho objektu, potom ďalšou možnosťou získania tajného kľúča je, že si ho prijímacia strana odvodí zo stego objektu prostredníctvom hašovacej funkcie [3].

C. Asymetrická steganografia s verejným kľúčom

Je steganografia, ktorej základná myšlienka je veľmi podobná princípu kryptografii s verejným kľúčom. Steganografia s verejným kľúčom využíva dva kľúče. Jeden z nich je označený ako verejný kľúč a je uložený vo verejnej databáze kľúčov, ku ktorej majú všetci účastníci komunikácie prístup. Druhý je označený ako súkromný kľúč a vlastníť by ho mal iba príjemca tajnej správy. Verejný kľúč sa používa v procese vkladania tajnej správy a súkromný kľúč v procese extrakcie tejto správy. Systému je definovaný množinou prvkov $\gamma = \{C, M, K_p, K_s, C_m, E_K, D_K\}$ kde C je množina krycích objektov, M predstavuje množinu všetkých tajných správ, pre ktoré platí podmienka $|C| \geq |M|$. K_p je množina verejných kľúčov, K_s je množina súkromných kľúčov, C_m predstavuje množinu stego objektov a E_K, D_K predstavujú zobrazenia $E_K : C \times M \times K_p \rightarrow C_m$ a $D_K : C_m \times K_s \rightarrow M$ s vlastnosťami $D_K(E_K(c, m, k_p), k_s) = m$ a $D_K(E_K(c, m, k_s), k_p) \neq m$.

Tento systém tiež využíva jednocestné funkcie, ktoré zabezpečujú, že verejný kľúč je odvodený od súkromného, ale opačne to nie je možné. Steganografia s verejným kľúčom využíva fakt, že dekódovacia jednocestná funkcia sa môže aplikovať na krycí objekt bez ohľadu na to, či obsahuje alebo neobsahuje tajnú správu. Prijemca nemá informáciu o tom, či prijatý stego objekt obsahuje tajnú správu, preto po obdržaní takéhoto stego objektu sa ju pokúsi extrahovať použitím svojho súkromného kľúča. Uvedený model steganografického systému je naznačený na Obr. 2.



Obr. 2 Model steganografického systému s verejným kľúčom

Väčšina aplikácií však využíva protokol čistej steganografie, pretože je jednoduchší a komunikujúce strany nemusia používať žiaden kľúč a to aj napriek tomu, že existuje väčšie riziko odhalenia tajnej komunikácie v prípade, že útočník pozná steganografické algoritmy [9].

IV. STEGANOGRAFIA AKO PRENOSOVÝ KANÁL

Na rozdiel od možnosti modelovania prenosových kanálov s rôznym rozdelením pravdepodobnosti šumu pomocou modelu kanála s bielym gaussovým šumom (napr. v prípade kanála s Laplaceovým rozdelením pravdepodobnosti šumu platí pre odhad jeho prenosovej kapacity vzťah $C_g \leq C_{\text{Laplace}} \leq C_g + 0.1044$) je modelovanie prirodzených digitalizovaných obrazov vyznačujúcich sa silnou koreláciou medzi susednými obrazovými bodmi pomocou takéhoto prístupu veľmi obtiažne. Avšak aj v tomto prípade je kapacita steganografických kanálov zdola ohraničená kapacitou gaussovho prenosového kanála. Z tohto pohľadu je výhodnejšie modelovať kapacitu steganografického kanála pomocou kanálov s ľubovoľným rozdelením pravdepodobnosti šumu. Shannonova teória vraví, že „náhodnosť“ šumu s ľubovoľným rozdelením pravdepodobnosti môže byť popísaná gaussovým bielym šumom pri použití merania výkonu entropie, známym aj ako efektívny výkon šumu. Ak má šum Z s ľubovoľným rozdelením pravdepodobnosti entropiu $H(Z)$, je priemerný výkon šumu s gaussovým rozdelením pravdepodobnosti majúceho rovnakú entropiu ako šum Z

$$N_e(Z) = \frac{1}{2\pi e} e^{2H(Z)}. \quad (1)$$

Na základe uvedených úvah platí, že kapacita steganografického kanála je ohraničená

pomocou vzťahu

$$C_g \leq C_{\text{stego}} \leq \frac{1}{2} \log_2 \frac{S+N}{N_e}, \quad (2)$$

kde N_e predstavuje výkon entropie šumu krycích dát. Keďže N_e je vždy nižšia ako N pre všetky prenosové kanály s iným ako gaussovým rozdelením pravdepodobnosti šumu, môžeme spresniť odhad kapacity steganografického prenosového kanála pomocou vzťahu (Marvel and Charles 1999)

$$C_g \leq \frac{1}{2} \log_2 \frac{S+N_e}{N_e} \leq C_{\text{stego}} \leq \frac{1}{2} \log_2 \frac{S+N}{N_e}. \quad (3)$$

Pre výkon entropie krycích dát na základe vzťahu (7) platí

$$N_e(\text{krycie dáta}) = \frac{1}{2\pi e} e^{2H(\text{krycie dáta})}. \quad (4)$$

Posledným krokom pred odhadom prenosovej kapacity steganografického systému ostáva výpočet entropie krycích dát. Tento výpočet je možné uskutočniť pomocou vzťahu

$$H(N) = -\frac{1}{n} \int \dots \int p(x_1, \dots, x_n) \log(p(x_1, \dots, x_n)) dx_1 \dots dx_n, \quad (5)$$

kde

$$H_{\text{krycie dáta}} = \lim H_n. \quad (6)$$

Na výpočet entropie teda potrebujeme poznať funkciu rozdelenia pravdepodobnosti krycích dát. Aj keď v súčasnosti existuje viacero modelov obrazových dát, neexistuje žiaden konzistentný, všeobecne použiteľný model a bez vhodného modelu je výpočet entropie krycích dát nerealizovateľný. Preto sa v tejto oblasti využívajú poznatky získané z teórie kanálového kódovania, kde platí, že entropia zdroja je približne rovná priemernej dĺžke kódového slova. Túto hodnotu je možné získať napríklad aplikáciou použitím bezstratového kompresného algoritmu *CALIC* na krycie dáta, ktorého výstupná hodnota prenosovej rýchlosti bude predstavovať horný odhad hodnoty entropie krycích dát. Použitím týchto hodnôt môžeme po dosadení do vzťahu (6) získať horný a dolný odhad kapacity steganografického prenosového kanála pre konkrétne krycie dáta. Teória odhadu kapacity steganografických kanálov sa v súčasnosti dostáva čoraz viac do popredia a výskum v tejto oblasti významne napomáha pri optimalizácii algoritmov steganografie.

V. ZÁVER

Predložený článok je venovaný problematike ochrany informácií prenášaných cez verejné dátové siete so zameraním na oblasť metód určených hlavne na utajovanie komunikácie, t.j. metódy steganografie. Táto oblasť, na rozdiel od oblasti kryptografie, predstavuje nový pohľad na spôsob riešenia problému ochrany dát s dôrazom na utajenie ich prenosu a nie na utajenie ich obsahu.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] KATZENBEISSER, S. and PETICOLAS, F.: Information Hiding Techniques for Steganography and Digital Watermarking. *Artech House*, INC, 2000. ISBN1-58053-035-4.
- [2] PROAKIS, J. G.: Digital Communications. *McGraw Hill*, USA 1989.
- [3] Agrawal, N. – Gupta, A. 2009. DCT Domain Message Embedding in Spread-Spectrum Steganography System. In *Data Compression Conference*. ISSN: 1068-0314, 2009, p. 433-433.
- [4] HOPPER, N. J. – LANGFORD, J. – AHN, L. 2009. Provably Secure Steganography. In *Computers, IEEE Transactions*. ISSN: 0018-9340, 2009, vol. 58, no. 5, p. 662-676.
- [5] BRUNDICK, F. S. – MARVEL, L. M. 2001. Implementation of Spread Spectrum Image Steganography. *Army Research Laboratory*, Aberdeen Proving Ground, 2001, MD 210055067.
- [6] KER, A., D. 2005. Steganalysis of LSB matching in grayscale images. In *Signal Processing Letters, IEEE*. ISSN: 1070-9908, 2005, vol. 12, no. 6, p. 441-444.
- [7] SULEIMAN M. A. R. 2009. Colored Image-In-Image Hiding. In *CAD Systems in Microelectronics*. Lviv-Polyana: CADSM, 2009. Print ISBN: 978-966-2191-05-9 p. 445-450.
- [8] BUGÁR, G. – BÁNOCI, V. – LEVICKÝ, D. 2010. Information hiding using pseudo-random number sequences with error correction. In *Radioelektronika 2010: proceedings of 20th international conference : April 19-21, 2010, Brno, Czech Republic. - Brno: Brno University of Technology*. 2010. ISBN: 978-1-4244-6320-6, p. 1-4.
- [9] BÁNOCI, V. – BUGÁR, G. – LEVICKÝ, D. 2008. Moderná steganografia v dátových komunikačných systémoch. In *Manažérstvo bezpečnosti : zborník vedeckých prác*. Košice : Vysoká škola bezpečnostného manažérstva v Košiciach, 2008. ISBN: 978-80-89282-29-6, s. 40-51.

Slepá steganografia na báze DWT

Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ,
Zita KLENOVIČOVÁ, Dušan LEVICKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

vladimir.banoci@tuke.sk, gabriel.bugar@tuke.sk, tomas.kanocz@tuke.sk,
zita.klenovicova@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Článok je zameraný na problematiku slepej, resp. nezávislej steganografie na báze diskkrétnej waveletovej transformácie (DWT) v statických farebných obrazoch. Jednotlivé časti príspevku opisujú vlastnosti steganografických systémov a ich konflikty. V ďalšej kapitole je rozpracovaná základná myšlienka DWT a jej reprezentácie v statických obrazoch. Štvrtá kapitola opisuje základnú myšlienku navrhutej metódy a v závere sú uvedené najlepšie dosiahnuté výsledky.

Kľúčové slová — steganografia, ukrývanie informácií, DWT, podprahový kanál

I. ÚVOD

Už od počiatku civilizácie existovali spôsoby, ako chrániť informácie. Najznámejším spôsobom je kryptografia, ktorá zamedzuje prístup k dátam pomocou rôznych techník šifrovania, avšak šifrovanie informácií im neposkytuje úplnú ochranu. Preto sa výskum v tejto oblasti zamerával na vývoj vhodnejších techník utajovania informácie medzi ktoré patrí *steganografia*. Steganografia predstavuje ďalší a silnejší stupeň ochrany informácií a zahŕňa metódy, pomocou ktorých je možná realizácia ukrývania dát v iných dátach. Ukryté dáta sa takto stávajú pre náhodného pozorovateľa nevnímateľné, a teda ide o samotné utajenie prebiehajúcej komunikácie. V kryptografii je primárne utajovaný obsah informácie a skutočnosť, že prebieha komunikácia, je všeobecne známa.

II. VLASTNOSTI STEGANOGRAFICKÝCH SYSTÉMOV

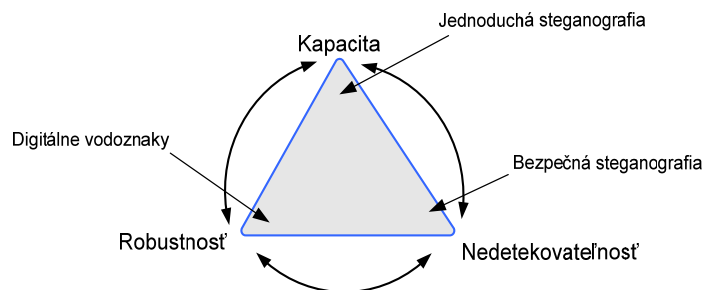
Steganografia ako aj digitálna vodotlač využívajú na skrytý prenos dát techniky vkladania pomocou ktorých je možné ukryť tajné dáta do krycích objektov. Rozdiel medzi týmito prístupmi spočíva v účele ich použitia. Kým steganografia je primárne určená na utajenie komunikácie medzi dvoma skupinami (osobami), digitálna vodotlač sa využíva na autentizáciu a ochranu duševného vlastníctva spôsobom „jeden odosielateľ – neobmedzené množstvo prijímateľov“. Steganografické systémy môžu byť charakterizované nasledujúcimi vlastnosťami:

- Nedetekovateľnosť (*undetectability*)
 - analýzy steganografických systémov ukazujú, že nedetekovateľnosť pre teoreticky bezpečný steganografický systém možno zabezpečiť, ak operácia vkladania tajnej správy má náhodný charakter a vložená tajná správa má nezávislý charakter od krycieho a tiež aj od stego objektu. Tieto podmienky zabezpečujú a zaručujú nezistiteľnosť tajnej správy nepovolenou osobou, ktorá má prístup k stego objektu, ale nemá žiadne dostupné informácie o nedeterministickom vkladaní tajnej správy.
- Perceptívna transparentnosť a nevnímateľnosť (*perceptual transparency and imperceptibility*)
 - tieto požiadavky sú založené na vlastnostiach ľudského vizuálneho alebo ľudského sluchového systému. Vložená tajná informácia je perceptívne transparentná, ak priemerný ľudský subjekt nie je schopný rozlíšiť originálne, resp. krycie objekty od stego objektov, teda od dát obsahujúcich ukrytú tajnú správu. Táto vlastnosť sa experimentálne zisťuje na veľkom počte krycích a stego objektoch, v ktorých buď je alebo nie je vložená tajná informácia tak, že skupina ľudí porovnáva a zisťuje podobnosť medzi týmito objektmi.
- Bezpečnosť a robustnosť (*security and Robustness*)

- algoritmus steganografického systému možno označiť ako „bezpečný“ ak vloženie tajnej informácie nie je možné extrahovať zo stego objektu bez znalosti algoritmov vkladania alebo extrakcie, resp. čiastočného poznania týchto algoritmov,
- algoritmus steganografického systému je „robustný“, ak extrakcia, resp. detekcia tajnej správy zo stego objektu, ktorý bol predtým poškodený útokmi, prebehla úspešne.
- Kapacita (*capacity*)
 - kapacita steganografického systému je definovaná ako množstvo informácie, ktoré je odosielateľ schopný vložiť a ktoré prijímateľ bude schopný úspešne extrahovať pomocou dostupných algoritmov.

A. Konflikt medzi vlastnosťami

Vyššie opísané vlastnosti steganografických systémov sú vo vzájomnom konflikte, čo vylučuje ich súčasné splnenie. Znamená to, že nie je možné súčasne splniť všetky požiadavky tak, aby súčasne dosahovali dostatočne uspokojujúce hodnoty. V prípade, že je predmetom záujmu dosiahnuť vloženie veľkého množstva informácie (t.j. dôraz je na kapacite vložených dát), nie je možné súčasne požadovať to, aby vložená tajná správa vykazovala vysokú robustnosť alebo nedetekovateľnosť. Na druhej strane, ak sú vysoké nároky na robustnosť vložených dát dochádza k ich ľahšej detekovateľnosti. Na obr. 1 je schematicky zobrazená táto situácia a tiež požiadavky, ktoré sú uprednostňované v jednotlivých oblastiach ukrývania informácie.



Obr. 1 Konfrontácia vlastností steganografického systému

Vo všeobecnosti rozoznávame dve základné skupiny systémov steganografie v závislosti od množstva informácie potrebnej pre úspešnú extrakciu utajenej informácie:

1. *systémy s prenosom dodatočnej informácie*
 - pri ich použití je nutná prítomnosť identickej kópie originálnych krycích dát v procese extrakcie utajenej informácie na strane prijímateľa,
2. *systémy bez prenosu dodatočnej informácie*
 - tieto systémy nevyžadujú prítomnosť originálnych krycích dát na strane prijímateľa a označujú sa ako tzv. „nezávislé“, resp. „slepé“ systémy. Tieto môžeme ďalej rozdeliť na:
 - *systémy s estimáciou krycích dát*
 - originálne krycie dáta sa získavajú estimáciou prijatých označených dát. Ich použitie je typické v prípade aplikácie techník s využitím systémov s rozprestretým spektrom,
 - *systémy bez estimácie krycích dát*
 - v takomto prípade sa extrakcia tajnej informácie uskutočňuje iba z prijatých označených dát.

Procesy a úpravy na krycích objektoch sa musia vykonávať tak, aby v žiadnom prípade nevznikalo podozrenie, že stego objekt obsahuje skrytú informáciu. V dokonalom steganografickom systéme nesmie byť badateľný rozdiel medzi krycím a označeným objektom postrehnuteľný ľudským vnímaním alebo odhalený štatistickými metódami.

III. DISKRÉTNÁ WAVELETOVÁ TRANSFORMÁCIA - WAVELETOVÁ REPREZENTÁCIA OBRAZU

Teória waveletov pokrýva veľmi rozsiahlu oblasť aplikácií a možno ju aplikovať na spojité aj diskrétne signály. Zvlášť zaujímavou je jej aplikácia na analýzu nestacionárnych signálov, pretože poskytuje alternatívu ku klasickej krátkodobej Fourierovej transformácii alebo ku Gáborovej transformácii.

Waveletová reprezentácia obrazu patrí spolu s pyramídovou a subpásmovou reprezentáciou medzi hierarchické reprezentácie obrazu. Hierarchické reprezentácie umožňujú reprezentovať obraz vo viacerých úrovniach s rozdielnym priestorovým rozlíšením. Základom efektívnych

algoritmov výpočtu DWT je teória dekompozície signálov s viacnásobným rozlíšením. Princíp reprezentácie signálov s viacnásobným rozlíšením je založený na tom, že funkcia $f(t)$ môže byť vyjadrená aproximáciami s rôznym množstvom detailov, pričom každá aproximácia korešponduje rôznemu rozlíšeniu a je vyhladenou verziou funkcie $f(t)$. Jednotlivé aproximácie sa získajú konvolúciou s dolnopriepustným filtrom a funkcia, ktorú filter realizuje sa označuje ako mierková funkcia $\phi(t)$ (*scaling function*).

Reprezentácia signálov s viacnásobným rozlíšením je štruktúrou veľmi podobná pyramidálnym reprezentáciám. Efektívnym spôsobom získania diskkrétnej aproximácie signálu sa javí postup, pri ktorom sa diskrétna aproximácia signálu pri rozlíšení m získa z jemnejšej diskkrétnej aproximácie pri rozlíšení $(m-1)$. Uvedený postup možno realizovať použitím FIR filtra, ktorý označíme symbolom G . Impulzovú odpoveď filtra G možno vyjadriť v tvare:

$$g_0(n) = \langle \phi(t/2), \phi(t-n) \rangle \quad (1)$$

Filter označený symbolom H , ktorý je zrkadlovým filtrom ku G , má impulzovú odozvu

$$h_0(n) = g_0(-n) \quad (2)$$

Ak $a_m(n)$ je diskrétna aproximácia signálu $f(t)$ pri rozlíšení 2^m , potom platí:

$$a_m(n) = \sum_{k=-\infty}^{\infty} h_0(2n-k) \cdot a_{m-1}(k) \quad (3)$$

pričom zároveň platí:

$$\phi(t) = 2 \cdot \sum_n g_0(n) \cdot \phi(2t-n) \quad (4)$$

Vzťah (4) vyjadruje operáciu konvolúcie, po ktorej nasleduje decimácia faktorom 2.

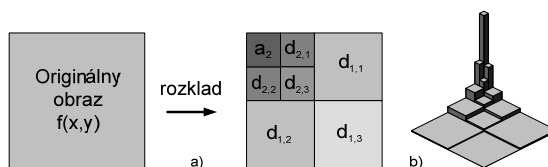
Podľa vzťahu (4) je možné mierkovú funkciu vyjadriť ako lineárnu kombináciu básových funkcií. Tento vzťah je jedným zo základných vzťahov ortogonálnej waveletovej transformácie, z ktorého sú odvodzované ďalšie vlastnosti filtra $H_0(z)$.

Waveletová transformácia je bezstratová, t.j. umožňuje bezstratovú rekonštrukciu signálu rekurzívnym postupom, ktorý možno vyjadriť v tvare:

$$a_{i-1}(n) = 2 \cdot \sum_k g_0(n-2k) \cdot a_i(k) + 2 \cdot \sum_k g_1(n-2k) \cdot d_i(k) \quad (5)$$

Ortogonalný rozklad signálov pomocou QMF bánk filtrov je možné zovšeobecniť aj pre viacrozmerné signály. Pri spracovaní obrazov je potrebné aplikovať dvojrozmerný prípad a analogicky aj v tomto prípade možno definovať dvojrozmernú mierkovú funkciu $\phi(x,y)$, z ktorej sa pomocou dilatácie a posunutia vytvoria ortogonálne bázy $\phi_{i,m,n}(x,y)$. Analogicky s jednorozmerným prípadom je možné odvodiť taktiež rýchly algoritmus výpočtu dvojrozmernej separovateľnej waveletovej transformácie.

Pri aplikácii L -úrovňového rozkladu na obraz, parameter a_L predstavuje hrubú aproximáciu obrazu pri rozlíšení 2^{-L} a $d_{i,j}$ predstavujú detailné obrazy. Waveletová reprezentácia je neredundantná, t.j. počet op originálneho obrazu je zhodný s počtom bodov waveletovej transformácie. Typické usporiadanie hrubej aproximácie a detailných obrazov je uvedený na obr. 2 a príklad waveletovej reprezentácie pre konkrétny obraz je uvedený na obr. 3.



Obr. 2 Usporiadanie komponentov reprezentácie pre separovateľný rozklad $f(x,y)$ a pre $L=2$;
 b) Rozloženie energie vo waveletovej transformácii pre $L=4$



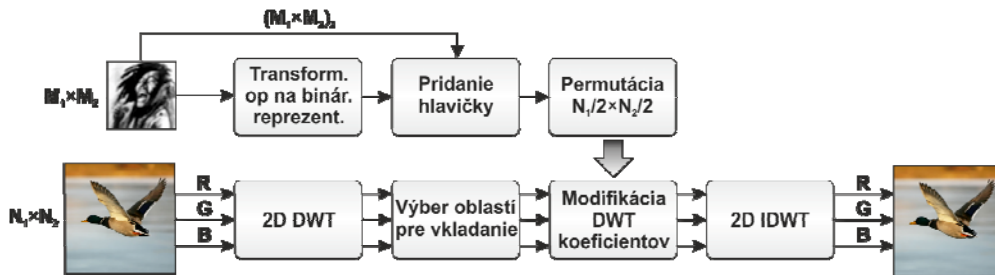
Obr. 3 Waveletová reprezentácia pre konkrétny obraz *Lena.bmp*

IV. NAVRHNUTÁ METÓDA BEZ POTREBY PRENOSU DODATOČNEJ INFORMÁCIE

Uvedená metóda využíva pre vkladanie, resp. extrakciu tajných správ dvojrozmernú diskretnú waveletovú transformáciu (2D DWT). Táto metóda, na rozdiel od metód navrhnutých na báze využívania 2D DCT transformácie, nevyžadujú v procese extrakcie tajnej správy originálny krycí obraz.

A. Proces vkladania tajnej správy

Navrhnutá metóda na báze 2D DWT pracuje so statickými viacúrovňovými farebnými krycími obrazmi a ako tajné správy využíva statické viacúrovňové obrazy. Blokovú schému navrhutej metódy ilustruje obr. 4.



Obr. 4 Blokovú schému navrhutej metódy

Vstupmi do procesu vkladania sú statický viacúrovňový farebný krycí obraz C s rozmerom $N_1 \times N_2$ a tajná správa m reprezentovaná statickým viacúrovňovým obrazom s rozmerom $M_1 \times M_2$. Zvolený krycí obraz sa v závislosti od vybraného farebného obrazového modelu, najprv rozloží na jednotlivé farebné zložky, ktoré sú následne transformované pomocou 2D DWT transformácie 1. úrovne. Takto získané transformačné DWT koeficienty C_{DWT} pozostávajú z oblasti aproximálnych transformačných koeficientov 1. úrovne a z oblasti horizontálnych, vertikálnych a diagonálnych detailov 1. úrovne. Navrhnutá metóda predpokladá výber minimálne dvoch, resp. troch oblastí DWT koeficientov, do ktorých bude možné realizovať vkladanie tajnej správy. Oblasť aproximálnych koeficientov je z daného výberu vylúčená, keďže nesie najväčšie množstvo informačného obsahu krycieho obrazu a jej modifikácia by mohla spôsobiť značnú degradáciu krycieho obrazu. Výber oblastí pre vloženie tajnej správy prebieha ľubovoľnou kombináciou vertikálnej, horizontálnej alebo diagonálnej oblastí. Vybraté oblasti DWT koeficientov C'_{DWT} spolu s upravenou tajnou správou m_p vstupujú do bloku modifikácie vybraných oblastí DWT koeficientov. Upravená tajná správa m_p sa získa z pôvodnej tajnej správy m úpravou na vektor $(M_1 \times M_2) \times 1$, ktorý je následne transformovaný na binárny tvar, čím vzniká matica upravenej tajnej správy veľkosti $(M_1 \times M_2) \times 8$. Táto matica je ešte doplnená o hlavičku veľkosti (1×16) , ktorá nesie informáciu o rozmere tajnej správy, pričom M_1 aj M_2 sú reprezentované 8 bitmi. Posledným krokom úpravy je permutácia takto vytvorenej matice na rozmer, ktorý je totožný s rozmerom vybraných oblastí DWT koeficientov. Pri prvej úrovni rozkladu DWT transformácie je tento rozmer $N_1/2$ a $N_2/2$.

Modifikované koeficienty vybraných oblastí DWT koeficientov sú získavané nasledujúcimi rovnicami:

$$ak \quad \text{mod}(\text{abs}(C'_{DWT}(x, y)), 1) \neq 0 \wedge m_p(x, y) = 0 \quad (6)$$

$$potom \quad C'_{m, DWT}(x, y) = C'_{DWT}(x, y) \quad (7)$$

$$ak \quad \text{mod}(\text{abs}(C'_{DWT}(x, y)), 1) \neq 0 \wedge m_p(x, y) = 1 \quad (8)$$

$$ak \quad C'_{DWT}(x, y) > 0 \quad (9)$$

$$potom \quad C'_{m, DWT}(x, y) = C'_{DWT}(x, y) - 0,5 \quad (10)$$

$$ak \quad C'_{DWT}(x, y) \leq 0 \quad (11)$$

$$potom \quad C'_{m, DWT}(x, y) = C'_{DWT}(x, y) + 0,5 \quad (12)$$

$$ak \quad \text{mod}(\text{abs}(C'_{DWT}(x, y)), 1) = 0 \wedge m_p(x, y) = 0 \quad (13)$$

$$ak \quad C'_{DWT}(x, y) > 0 \quad (14)$$

$$potom \quad C'_{m, DWT}(x, y) = C'_{DWT}(x, y) - 0,5 \quad (15)$$

$$ak \quad C'_{DWT}(x, y) \leq 0 \quad (16)$$

$$potom \quad C'_{m,DWT}(x, y) = C'_{DWT}(x, y) + 0,5 \quad (17)$$

$$ak \quad \text{mod}(\text{abs}(C'_{DWT}(x, y)), 1) = 0 \wedge m_p(x, y) = 1 \quad (18)$$

$$potom \quad C'_{m,DWT}(x, y) = C'_{DWT}(x, y) \quad (19)$$

kde $C'_{m,DWT}(x, y)$ predstavuje modifikované DWT koeficienty, $C'_{DWT}(x, y)$ predstavuje DWT koeficienty z vybraných oblastí a $m_p(x, y)$ predstavuje upravenú tajnú správu.

V nasledujúcom kroku sa pôvodné DWT koeficienty C_{DWT} nahradia modifikovanými vybranými koeficientmi $C'_{m,DWT}$, čím sa získajú modifikované DWT koeficienty $C_{m,DWT}$.

Aplikáciou inverznej dvojrozmernej diskkrétnej waveletovej transformácie (2D IDWT) na takto získané DWT koeficienty $C_{m,DWT}$ sa získa stego obraz C_m s vloženou tajnou správou.

B. Proces extrakcie tajnej správy

Vstupom procesu extrakcie tajnej správy je iba stego obraz C_m , na ktorý sa najprv aplikuje 2D DWT, čím sú získané transformačné koeficienty $\hat{C}_{m,DWT}$. Nasleduje výber oblastí $\hat{C}'_{m,DWT}$, t.j. oblasť vertikálnych, horizontálnych alebo diagonálnych transformačných koeficientov, ktoré boli v procese vkladania vybraté a ovplyvnené tajnou správou a vstupujú do bloku extrakcie vybraných oblastí DWT koeficientov. V tomto bloku sa z modifikovaných DWT koeficientov získava tajná správa $\hat{m}_p(x, y)$, pričom proces extrakcie tajnej správy možno vyjadriť nasledujúcimi vzťahmi:

$$ak \quad \text{mod}(\text{abs}(\hat{C}'_{DWT}(x, y)), 1) \neq 0 \quad (20)$$

$$potom \quad \hat{m}_p(x, y) = 0 \quad (21)$$

$$ak \quad \text{mod}(\text{abs}(\hat{C}'_{DWT}(x, y)), 1) = 0 \quad (22)$$

$$potom \quad \hat{m}_p(x, y) = 1 \quad (23)$$

V poslednom kroku sa získa rozmer tajnej správy, prenášanej prvými šesťástimi bitmi upravenej matice tajnej správy. Nasleduje prevod ostatných hodnôt získanej tajnej správy z bitovej reprezentácie na desiatkovú a spätné preusporiadanie získaných obrazových prvkov na pôvodný rozmer.

V. EXPERIMENTÁLNE VÝSLEDKY

Podstatou navrhnutej steganografickej metódy na vkladania tajných správ do farebných zložiek krycieho obrazu s využitím DWT sú modifikácie vybraných oblastí DWT koeficientov (koeficienty horizontálnych detailov $d_{1,1}$, koeficienty vertikálnych detailov $d_{1,2}$, resp. koeficienty diagonálnych detailov $d_{1,3}$). Táto metóda nerealizujú transformáciu po blokoch 8×8 , ale DWT transformácia je aplikovaná priamo na celý krycí obraz. Navrhnutá metóda umožňuje realizovať nezávislý steganografický systém, teda k úspešnej, 100% extrakcii tajnej správy zo stego obrazu postačuje iba stego obraz.

Samotné vkladanie tajnej správy sa realizuje na základe rovníc (6 - 19) opísaných v predošlej kapitole, ktoré opisujú spôsob modifikácie vybraných DWT transformačných koeficientov. Podstatou navrhnutej metódy je 2D DWT transformácia využívajúca Haarov wavelet. Charakteristickou vlastnosťou takto realizovanej DWT transformácie sú hodnoty transformačných koeficientov, vyjadrené v tvare násobkov čísla 0,5. Táto vlastnosť umožňuje realizovať vkladanie tajnej správy na báze vzťahov opísaných pomocou rovníc (1 - 14). Je potrebné tiež poznamenať, že pri vkladaní tajnej správy je nutné modifikovať minimálne dve vybraté oblasti DWT koeficientov krycieho obrazu, aby v procese extrakcie bolo možné úspešne získať vloženú tajnú správu zo stego obrazu.

Ďalšou vlastnosťou navrhnutej metódy je, že kapacita systému závisí iba na rozmere krycieho obrazu bez ohľadu na dynamické vlastnosti v krycom obraze. V prípade krycieho obrazu s priestorovou rozlišovacou schopnosťou 256×256 op s vysokou mierou detailov (D) je rozmer vybraných oblastí DWT koeficientov ($d_{1,1}$, $d_{1,2}$ alebo $d_{1,3}$), do ktorých bude realizované vkladanie, 128×128 . Množstvo bitov, ktoré je v tomto prípade možné ukryť, je teda 128×128 , čo predstavuje 16 384 bitov. Toto množstvo bitov zodpovedá tajnej správe reprezentovanej statickým viacúrovňovým obrazom s rozmerom $45 \times 45 \times 8$, ktorý predstavuje 16 200 bitov. Pri použití krycích obrazov s malým množstvom detailov (MD) $256 \times 256 \times 24b$ bude táto kapacita totožná, teda rozmer vlozenej tajnej správy bude tiež $45 \times 45 \times 8$.

Kvalita jednotlivých stego obrazov bola hodnotená objektívnymi mierami hodnotenia kvality PSNR, $PSNR_{CSF}$ a SSIM a ich výsledné hodnoty pre jednotlivé farebné obrazové modely boli vyhodnocované v tabuľkách. Testované boli farebné statické obrazy v štyroch farebných obrazových modeloch a to: RGB, HSV, YIQ a YCbCr. Zistené hodnoty sú uvedené v tab. 1.

Farebný obrazový model Použitá oblasť (Diag, Ver, Hor)		RGB Diag; Ver.	HSV Diag; Ver.	YIQ Diag; Hor.	YCbCr Diag; Ver.
KO _D <i>Lena.bmp</i>	PSNR [dB]	60,222	60,180	58,440	60,217
	$PSNR_{CSF}$ [dB]	108,697	106,750	87,482	88,896
	SSIM	0,9996	0,3123	0,9828	1
KO _{SD} <i>More.bmp</i>	Extrakcia dát [%]	99,987	99,938	100	100
	Počet chybných bitov	6/48600	20/48600	0/48600	0/48600
	PSNR [dB]	60,168	60,098	58,323	60,042
KO _{MD} <i>Lietadlo.bmp</i>	$PSNR_{CSF}$ [dB]	97,635	87,774	80,036	80,188
	SSIM	0,9992	0,4623	0,9719	0,9992
	Extrakcia dát [%]	99,993	99,973	99,996	100
KO _{MD} <i>Lietadlo.bmp</i>	Počet chybných bitov	1/48600	13/48600	2/48600	0/48600
	PSNR [dB]	59,901	59,887	58,067	59,873
	$PSNR_{CSF}$ [dB]	71,937	72,319	67,961	71,613
KO _{MD} <i>Lietadlo.bmp</i>	SSIM	0,9989	0,5376	0,9230	0,9988
	Extrakcia dát [%]	100	100	99,991	100
	Počet chybných bitov	0/48600	0/48600	4/48600	0/48600

Tab. 1 Experimentálne dosiahnuté výsledky pre farebné obrazové modely: RGB, HSV, YIQ a YCbCr

Z experimentálnych výsledkov z tab. 1 vyplýva, že najvhodnejší zo všetkých testovaných farebných modelov pre navrhnutú metódu na báze DWT transformácie je farebný obrazový model YCbCr. Najvyššie zistené hodnoty boli $PSNR = 60,2$ dB a hodnota $PSNR_{CSF} = 88,9$ dB.

VI. ZÁVER

Navrhnutá metóda na báze DWT transformácie vykazuje najvyššie hodnoty objektívnych mier hodnotenia kvality dosiahnuté využitím iba dvoch oblastí transformačných koeficientov. Vo všetkých farebných modeloch to boli koeficienty diagonálnych a vertikálnych detailov a iba v prípade modelu YIQ to boli koeficienty diagonálnych a horizontálnych detailov. Najväčšia kvalita stego obrazov bola dosiahnutá pre farebný model YCbCr a to pre krycie obrazy s vysokým množstvom detailov v luminančnej zložke Y. Hodnoty kvality stego obrazov boli približne 60 dB pre všetky testované farebné modely, ktoré pri množstve vložených bitov tajnej správy 16 200 predstavujú veľmi vysoké hodnoty.

Z pohľadu bezpečnosti steganografického systému je možné konštatovať, že metódy DWT, ktoré nevyžadujú v procese extrakcie tajných správ krycí obraz, poskytujú vyšší stupeň utajenia prebiehajúcej tajnej komunikácie, resp. znižujú možnosť odhalenia tejto komunikácie.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] BUGÁR, G. 2011. *Steganografia v statických farebných obrazoch* : dizertačná práca. Košice : KEMT FEI TUKE, 2011. 105 s.
- [2] SULEIMAN M. A. R. 2009. Colored Image-In-Image Hiding. In *CAD Systems in Microelectronics*. Lviv-Polyana: CADSM, 2009. Print ISBN: 978-966-2191-05-9 p. 445-450.
- [3] Antonini, M. – Barlaud, M. – Mathieu, P. – Daubechies, I. 1992. Image Coding Using Wavelet Transform. In *IEEE Transactions On Image Processing*. ISSN: 1057-7149, 1992, vol.1, no.2, p. 205-220.
- [4] RAJA, K. B. – SIDDARAJUL, S. – VENUGOPALL, K. R. – PATNAIK, L. M. 2007. Secure Steganography using Colour Palette Decomposition. In *Signal Processing, Communications and Networking, 2007. ICSCN '07. International Conference*. Print ISBN: 1-4244-0997-7, 2007, P. 74-80.
- [5] BUGÁR, G. 2008. Steganography in colour images. In *SCYR 2008 : 8th Scientific Conference of Young Researchers : proceeding from conference* [CD-ROM]. Košice : FEI TU, 2008. ISBN 978-80-553-0036-8, s. 126-128
- [6] HASSAN, M. – GHAZY, M. R. A. – ABDULAZIZ, A. M. – IBRAHIM, K. 2009. A Novel Approach for Hiding Messages in Images. In *International Conference on Signal Acquisition and Processing*. ISSN: 1057-7149, 2009, vol. 16, no. 2, p. 391-405.
- [7] YUAN-HUI, Y. – CHIN-CHEN, C. – IUON-CHANG, L. 2007. A new steganographic method for color and grayscale image hiding. In *Computer Vision and Image Understanding*. ISSN: 1077-3142, 2007, vol. 107, no. 3, p. 183-194.
- [8] KATZENBEISSER, S. and PETICOLAS, F.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, INC, 2000. ISBN1-58053-035-4.

Steganografia s vyhľadáváním perceptívne vhodných oblastí v krycích obrazoch

Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ,
Zita KLENOVIČOVÁ, Dušan LEVICKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

vladimir.banoci@tuke.sk, gabriel.bugar@tuke.sk, tomas.kanocz@tuke.sk,
zita.klenovicova@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Článok je zameraný na problematiku steganografie v statických farebných obrazoch. Prvá časť príspevku opisuje základné požiadavky kladené na steganografické systémy. Ďalšia kapitola pojednáva o operátoroch na detekciu hrán v obrazoch k výberu perceptívne vhodných oblastí na vkladanie tajných správ. Štvrtá kapitola opisuje základnú myšlienku navrhutej metódy a v závere sú uvedené experimentálne výsledky.

Kľúčové slová — steganografia, perceptívne vhodné oblasti, DCT, pseudonáhodné postupnosti, tajný kľúč

I. ÚVOD

Už od nepamäti sa ľudia pokúšali komunikovať pomocou rôznych prostriedkov, ako napr. reč, písmo takým spôsobom, ktorý by zabezpečoval ochranu obsahu pred neželanými osobami a ich vplyvu na prenášané správy. Jedným zo spôsobov na ochranu komunikácie a informácií je kryptografia, ktorej história siaha až do staroveku a zachovala sa až dodnes. V polovici 20. storočia bola využívaná na vojenské účely, ako ochrana komunikácie medzi vojenskými jednotkami. Základný princíp kryptografie spočíva v zašifrovaní tajnej správy pomocou kľúča. Takto zašifrovaná správa sa stáva nezrozumiteľnou bez použitia príslušného kľúča. Nevýhodou týchto šifrovaných správ je, že samotná komunikácia je viditeľná pre každého.

V súčasnosti sa pri ochrane tajných informácií kladie veľký dôraz práve na ich neviditeľnosť, teda na utajenie samotnej existencie takýchto informácií a práve to je základ vednej disciplíny zvanej steganografia (*steganography*). Základná myšlienka spočíva v utajení samotného faktu, že tajná informácia existuje, teda usiluje sa o vytvorenie tzv. "podprahového kanála" (*subliminal channel*), pomocou ktorej neviditeľná komunikácia s tajnými informáciami prebieha v pozadí viditeľnej, neutajenej a podozrenie nevzbudzujúcej komunikácie. Steganografia využíva bežne používané dáta (text, obrázky, audio alebo video) na utajenie tajných informácií, ktoré nevzbudzujú podozrenie o prenose informácií.

II. POŽIADAVKY KLADENÉ NA STEGANOGRAFICKÉ SYSTÉMY

Steganografia vykonáva skrytý, utajený prenos tajných informácií pomocou ich vkladia do krycích dát. Častejšie je využívaný pri komunikácii medzi dvoma skupinami (osobami). Medzi základné požiadavky patria:

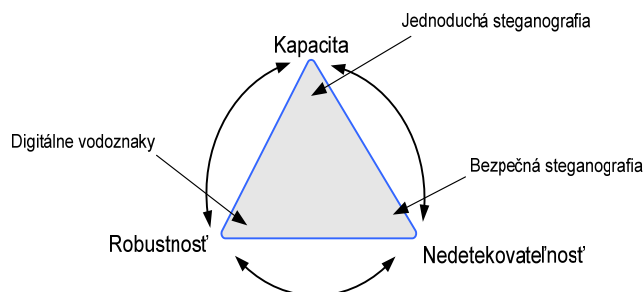
- **Robustnosť** (*robustness*)
 - vložená tajná informácia je nazývaná *robustnou*, ak je ľahko extrahovateľná, resp. detekovateľná z označených dát, ktoré boli predtým poškodené útokmi.
- **Nedetekovateľnosť** (*undetectability*)
 - je typickou vlastnosťou bezpečnej tajnej komunikácie. Vložená tajná informácia je nedetekovateľná, ak jej vloženie nevyvolá štatisticky významnú zmenu krycích dát, ktorou by sa označené dáta líšili od originálnych dát.
- **Perceptívna transparentnosť alebo nevnímateľnosť** (*invisibility*)
 - táto požiadavka je založená na vlastnostiach ľudského vizuálneho systému alebo ľudského sluchového systému. Vložená tajná informácia je perceptívne

transparentná, ak priemerný človek nie je schopný rozlíšiť originálne dáta od označených dát, teda od dát obsahujúcich skrytú informáciu.

- **Bezpečnosť** (*security*)
 - o bezpečnom algoritme sa hovorí v prípade, že vložená tajná informácia nemôže byť extrahovaná z označených dát bez znalosti algoritmu vkladania a extrakcie alebo čiastočnom poznaní procesu vkladania a extrakcie vlozenej informácie.
- **Kapacita** (*capacity*)
 - kapacita steganografických systémov predstavuje množstvo informácie, ktorú sme schopní vložiť, preniesť a úspešne extrahovať pomocou daného algoritmu.

A. Konflikt požiadaviek

Vyššie opísané požiadavky, resp. vlastnosti steganografických systémov sú vo vzájomnom konflikte, čo vylučuje ich súčasné splnenie. Znamená to, že nie je možné súčasne splniť všetky požiadavky tak, aby súčasne dosahovali dostatočne uspokojujúce hodnoty. V prípade, že je predmetom záujmu dosiahnuť vloženie veľkého množstva informácie (t.j. dôraz je na kapacite vložených dát), nie je možné súčasne požadovať to, aby vložená tajná správa vykazovala vysokú robustnosť alebo nedetekovateľnosť. Na druhej strane, ak sú vysoké nároky na robustnosť vložených dát dochádza k ich ľahšej detekovateľnosti. Na obr. 1 je schematicky zobrazená táto situácia a tiež požiadavky, ktoré sú uprednostňované v jednotlivých oblastiach ukrývania informácie.



Obr. 1 Konflikt požiadaviek

Na základe požiadaviek kladených na steganografické systémy uvedených na obr. 1 sa definujú tri oblasti utajovania informácií. Sú to:

o **oblasť jednoduchšej steganografie** (*simple steganography*)

- nie je tu očakávaná prítomnosť útokov pričom hlavný dôraz sa kladie na maximalizáciu kapacity prenosového steganografického kanála, pri súčasnom dodržaní kvality prenášanej informácie,

o **oblasť bezpečnej steganografie** (*secure steganography*)

- táto oblasť zahŕňa systémy s tajným a verejným kľúčom so základnou požiadavkou minimalizácie schopnosti detekcie utajenej komunikácie, teda maximalizácie nedetekovateľnosti a nevnímateľnosti,

o **oblasť digitálnych vodoznakov** (*digital watermarking*)

- cieľom digitálnych vodoznakov je vloženie prídavnej informácie za účelom označenia pôvodu, vlastníctva digitálnych dát a teda hlavnou požiadavkou tejto oblasti bude odolnosť označených dát voči možným útokom. Prítomnosť vodoznakov nemusí byť skrytá, ich viditeľnosť je závislá od účelu ich použitia. Na základe toho rozlišujeme 3 druhy vodoznakov:

1. viditeľné vodoznaky (*visible watermarks*),
2. neviditeľné vodoznaky (*invisible watermarks*),
3. krehké vodoznaky (*fragile watermarks*).

Ďalšími, tiež veľmi významnými požiadavkami steganografických systémov sú:

- úspešnosť, resp. efektívnosť extrakcie tajných správ,
- výpočtová nenáročnosť.

Úspešnosť, resp. efektívnosť extrakcie tajných správ - je určená pravdepodobnosťou chybné extrakcie. Jedná sa o extrakciu tajnej správy z neoznačených objektov (neadekvátna extrakcia), resp. neúspešnosť detekcie tajnej správy zo stego objektov. Pre efektívne steganografické metódy sa vyžaduje, aby táto pravdepodobnosť bola minimálna.

Výpočtová nenáročnosť – algoritmov vkladania a extrakcie, resp. detekcie tajnej správy v krycích objektoch je dôležitá pre uplatnenie v komerčnej sfére. Veľký dôraz na túto vlastnosť sa kladie v aplikáciách steganografie, kedy vkladanie, extrakcia, a detekcia tajných správ musí prebiehať v reálnom čase.

III. OPERÁTORY DETEKcie HRÁN V KRYCOM OBRAZE

Jedna z možností ako nájsť perceptívne vhodné oblasti v krycom obraze je vyhľadanie hrán objektov v obraze. Za hrany sa považujú oblasti v ktorých dochádza k prudkým zmenám hodnôt jasu v obrazovej funkcii. Na ich zistenie sa najčastejšie používajú gradientné operátory, ktoré určujú smer a veľkosť rastu funkcie dvoch premenných. Takto získané body s najväčšou hodnotou gradientu budú považované za hrany.

Gradient funkcie dvoch premenných je daný vzťahom

$$\nabla f(x, y) = \left(\frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y} \right) \quad (1)$$

kde $f(x, y)$ je obrazová funkcia krycieho obrazu.

Veľkosť gradientu funkcie dvoch premenných možno vyjadriť v tvare

$$|\nabla f(x, y)| = \sqrt{\left(\frac{\partial f(x, y)}{\partial x} \right)^2 + \left(\frac{\partial f(x, y)}{\partial y} \right)^2} \quad (2)$$

Smer hrany gradientu funkcie dvoch premenných je daný vzťahom

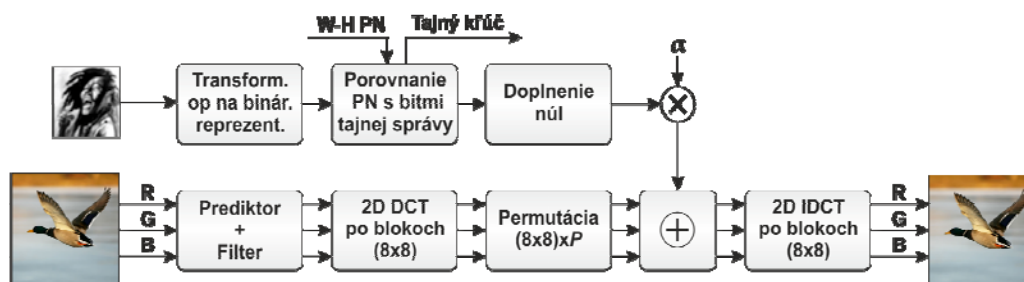
$$\Psi = \arctg \left(\frac{\frac{\partial f(x, y)}{\partial x}}{\frac{\partial f(x, y)}{\partial y}} \right) \quad (3)$$

Na základe využívania gradientných operátorov sú založené aj niektoré známe metódy detekcie hrán, ktoré sú podporované aj programovým prostredím *MatLab* a sú kategorizované nasledovne:

- o detektory hrán, ktorých podstata spočíva v aproximácii prvej derivácie obrazovej funkcie s využívaním viacerých masiek pri detekcii orientácie smeru hrany. Do tejto kategórie sú zaradované operátory *Robertsov*, *Sobelov* a *Prewittovej*,
- o druhá kategória je tvorená detektormi hrán, ktoré sú založené na zisťovaní prechodu nulou po druhej derivácii obrazovej funkcie. Zaradujú sa sem operátory *Zero-crossing*, *Cannyho* a *Logaritmický*.

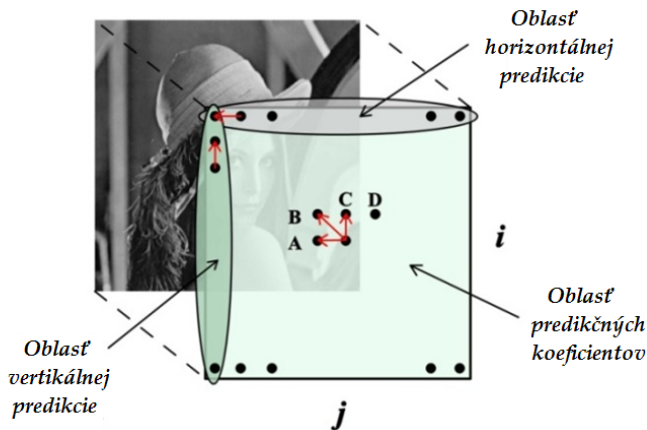
IV. NAVRHNUTÁ METÓDA ZALOŽENÁ NA BÁZE 2D DCT A VYHLADÁVANÍ PERCEPTÍVNE VHDNÝCH OBLASTÍ V KRYCÍCH OBRAZOCH S GENEROVANÍM TAJNÉHO KLÚČA

Navrhnutá metóda využíva vyhľadávanie perceptívne vhodných oblastí na vkladanie tajných dát v krycích obrazoch za účelom zvýšenia ich nevnímateľnosti. Vyhľadávanie týchto oblastí je realizované detektormi hrán opísaných v predchádzajúcich kapitolách. Ďalšou významnou črtou tejto metódy je generovanie tajného kľúča v procese vkladania tajnej správy na báze pseudonáhodných postupností, ktorý zvyšuje bezpečnosti utajenia komunikácie. Bloková schéma metódy používajúcej vyhľadávanie perceptívne vhodných oblastí v krycích obrazoch a generovanie tajného kľúča v procese vkladania tajnej správy v prostredí 2D DCT transformácie pre viacúrovňové statické farebné obrazy je uvedený na obr. 2.



Obr. 2 Bloková schéma navrhnutej metódy využívajúcej predikciu krycieho obrazu

Vstupmi procesu vkladania sú statický farebný krycí obraz C s rozmerom $N_1 \times N_2$, tajná správa m reprezentovaná statickým viacúrovňovým obrazom veľkosti $M_1 \times M_2$ a Walsh-Hadamardove pseudonáhodné postupnosti PN. Po zvolení vhodného krycieho obrazu a jeho rozložení na jednotlivé farebné zložky v závislosti od použitého farebného modelu nasleduje vyhľadanie perceptívne vhodných oblastí v obrazových zložkách na báze detektora hrán alebo vnútrosnímkovej predikcie. Vzhľadom na rozmiestnenie obrazových prvkov v krycom obraze, v procese predikcie je využitá kombinácia predikčných algoritmov s jednorozmernou horizontálnou a vertikálnou predikciou, ktorá je znázornená na obr. 3.



Obr. 3 Využívanie vnútrošnímkovovej predikcie v krycom obraze

Výstupom tejto predikcie budú tzv. chybové obrazové body, resp. chybový obraz s menšou disperziou ako mali obrazové body pred predikciou. Použitím filtra, ktorý potláča body s malou jasovou hodnotou sa dosiahne zvýraznenie tých oblastí, ktoré sa významnou mierou odlišujú od okolitých obrazových bodov a vytvárajú hrany. Správne nastavenie opísaného filtra umožňuje potláčať obrazové prvky, ktoré sú menšie ako dvojnásobná hodnota priemeru z celého chybového obrazu. Táto podmienka zabezpečí detekciu skutočných hrán od šumových vzoriek v obraze. Filtrovaný obraz sa následne rozdelí na bloky veľkosti 8×8 , pričom na vkladanie sa používajú bloky, ktoré obsahujú nenulové vzorky. Aplikáciou 2D DCT transformácie na tieto bloky sa transformujú ich obrazové prvky do spektra transformačných koeficientov. Preusporiadaním v slede „cik-cak“ sú tieto spektrálne koeficienty zoradené do vektorov, kde rovnako ako v predošlej metóde, prvý koeficient predstavuje jednosmernú DC zložku a ostatné koeficienty sú tvorené AC koeficientmi. Hodnota offsetu definuje počet prvých koeficientov preusporiadaného vektora, ktoré nebudú podliehať procesu modifikácie. V programovej realizácii je možné túto hodnotu offsetu meniť, čím sa priamo ovplyvňuje veľkosť vlozenej tajnej správy, resp. veľkosť degradácie stego obrazu.

Ďalším krokom je modifikácia tajnej správy m , ktorá je najprv upravená do vhodnej formy na vloženie do spektrálnych koeficientov nachádzajúcich sa v preusporiadaných vektoroch. Tajná správa, ktorá je reprezentovaná statickým viacúrovňovým obrazom s rozmerom $M_1 \times M_2$ sa permutuje na vektor o veľkosti $L \times 1$, kde L je rovný súčinu $M_1 \times M_2$. Následne je konvertovaná z dekadického do binárneho tvaru, čím vznikne matica tajných dát o rozmere $L \times 8$. Walsh-Hadamardove pseudonáhodné postupnosti sú upravené tak, že sa vynechá ich prvý riadok, keďže ten je tvorený iba jednotkami a jeho použitie nemá v tomto prípade žiadny praktický význam.

Po týchto úpravách začína samotný proces vkladania tajnej správy, ktorý pozostáva z porovnávania každého prvku upravenej matice tajných dát so všetkými siedmimi Walsh-Hadamardovými pseudonáhodnými postupnosťami po riadkoch. Cieľom porovnávania je zistiť najväčšiu podobnosť riadku tajných dát s pseudonáhodnými postupnosťami. Výstupom bude matica veľkosti $L \times 8$ obsahujúca Hadamardove postupnosti, ktoré sa najmenej odlišovali od porovnávaného riadku matice tajných dát. Ďalším výstupom je tajný kľúč K obsahujúci rozdiely medzi daným riadkom matice tajných dát a vybranou pseudonáhodnou postupnosťou, ktorá vykazovala od neho najmenšiu odchýlku. Týmto spôsobom dochádza k rozprestretiu tajnej správy, resp. dochádza k ukrytiu správy pod úroveň šumu. Následne dochádza k úprave rozprestretej tajnej správy z unipolárneho tvaru $(0, 1)$ na bipolárny tvar s hodnotami $(-1, 1)$, kvôli zvýšeniu robustnosti voči chybám pri prenose a kvôli lepšej extrakcii tajnej správy. Tieto dáta sú potom vynásobené škálovacím faktorom α a následne pričítané ku transformačným koeficientom.

Upravená matica dát krycieho obrazu obsahuje transformačné koeficienty s veľkosťou $R \times P$, kde rozmer R vychádza z veľkosti použitých blokov 8×8 upravený po preusporiadaní v slede „cik-cak“ na vektor. V danom prípade bude $R = 64$ a P je daný počtom vybraných blokov 8×8 z krycieho obrazu, ktoré obsahovali detekované hrany. V procese vkladania upravených dát tajnej správy sa krycí objekt transformačných koeficientov veľkosti $R \times P$ rozdelí na kanály o veľkosti $8 \times P$, kde jeden z kanálov bude ofsetový. Ofsetový kanál sa vyznačuje tým, že obsahuje transformačné koeficienty iba s najnižšími sekvenciami, ktoré nepodliehajú procesu modifikácie. Rozmer tohto kanála určujú parameter ofset, ktorý je v tejto metóde nastavený na hodnotu 8 a rozmer P . Ostatné kanály budú reprezentovať priestor na vloženie upravených dát tajnej správy. Rozprestreté tajné dáta sa rozdelia na bloky podľa veľkosti vytvorených kanálov a následne sú k nim pripočítané. Kvôli zvýšeniu robustnosti steganografického systému odolného voči šumu, modifikácia najprv nastáva v transformačných koeficientoch vo vyšších kanáloch.

Poslednými krokmi pri vytváraní stego obrazu je spätná permutácia modifikovaných transformačných koeficientov a aplikácia inverznej 2D DCT transformácie. Takto vzniknuté modifikované bloky 8×8 sú umiestnené späť do krycieho obrazu na miesta, odkiaľ boli po detekcii hrán vybrané.

Proces extrakcie tajnej správy je obdobný procesu jej vkladania. Vstupmi sú statický stego obraz C_m , tajný kľúč K a statický farebný viacúrovňový krycí obrázok C . V prvom kroku sa aplikuje vnútrosnímková predikcia na originálny krycí obraz a jeho filtráciou sú získané bloky obsahujúce hrany, pričom identické bloky sú vybrané aj zo stego obrazu. Na tieto bloky sa následne aplikuje 2D DCT transformácia, ktorá transformuje jasové hodnoty obrazových prvkov po blokoch 8×8 na transformačné koeficienty. Odčítaním spektrálnych transformačných koeficientov vybraných blokov stego a originálneho obrazu sa získa upravená matica dát tajnej správy. Následnou aplikáciou tajného kľúča, resp. opätovným porovnaním riadkov matice tajných dát a kľúča sa získajú presné miesta modifikácie v matici tajných dát. Po úprave týchto dát vznikne výsledná matica tajných správ. Následným prevodom týchto dát z binárneho do dekadického tvaru a po preusporiadaní tejto matice na pôvodný rozmer sa získa tajná správa.

V. EXPERIMENTÁLNE VÝSLEDKY

Podstata navrhutej steganografickej metódy, využívajúca farebné krycie obrazy vo farebných obrazových modeloch RGB, YCbCr, HSV a YIQ, spočíva v modifikácii vybraných DCT koeficientov na báze predikcie a detektorov hrán v krycom obraze, pomocou ktorých je možné proces vkladania prispôbiť podmienkam HVS. Vkladanie rozprestretých bitov tajnej správy pomocou pseudonáhodných postupností generovanej pomocou matice Hadamardových PN sa realizuje modifikáciou vybraných DCT transformačných koeficientov. Množstvo blokov, ktoré budú v procese vkladania modifikované závisí od množstva detailov v krycom obraze. V prípade využitia detailného (D) krycieho obrazu *Lena* s rozmerom $256 \times 256 \times 24b$ a pri využití predikcie v krycom obraze, možno na vkladanie tajnej správy použiť 392 blokov s rozmerom 8×8 (tento počet blokov vyplýva z experimentov). Vizuálna kvalita krycieho objektu je oveľa citlivejšia na modifikácie DCT koeficientov zodpovedajúcich nízkym priestorovým frekvenciám. Z tohto dôvodu, je v každom bloku na základe prednastaveného ofsetu vynechaných prvých 8 z možných 64 transformačných koeficientov, ktoré predstavujú najnižšie priestorové frekvencie. Takto ostáva pre vkladanie tajnej správy 56 transformačných koeficientov. Množstvo bitov, ktoré je v tomto prípade možné ukryť je teda 56×392 , teda 21 952 bitov. Rozprestretie tajnej správy je realizované porovnávaním so pseudonáhodnými postupnosťami z Hadamardovej matice PN, teda rozprestretá tajná správa má rovnakú veľkosť ako pred rozprestretím. Maximálne množstvo možných vložených bitov 21 952 zodpovedá približne tajnej správe reprezentovanej statickým viacúrovňovým obrazom s rozmerom $50 \times 50 \times 8$ (20 000 bitov).

Intenzita vlozenej tajnej správy a tým aj miera modifikácie vybraných DCT koeficientov je ovplyvnená hodnotou parametra α . Väčšie hodnoty tohto parametra spôsobia väčšiu degradáciu vo vizuálnej kvalite stego obrazu. Je potrebné poznamenať, že v procese vkladania tajnej správy sú modifikované iba DCT koeficienty, ktoré boli detekované ako hrany. Modifikácia DCT koeficientu je vykonaná podľa podmienok opísaných v predošlej kapitole. Celkový počet modifikovaných DCT koeficientov a tým aj výsledná kvalita stego obrazu okrem hodnoty parametra α teda závisí aj od charakteru krycieho obrazu.

Kvalita jednotlivých stego obrazov bola hodnotená objektívnymi mierami PSNR, $PSNR_{CSF}$ a SSIM, ktorých výsledky boli vyhodnocované v tabuľkách. Uvedená tabuľka tab. 1 obsahuje najlepšie dosiahnuté výsledky týchto parametrov, ktoré boli získané pre krycie obrazy vo farebnom obrazovom modeli YCbCr spolu s percentuálnymi vyjadreniami úspešnosti extrakcií tajných správ extrahovaných zo stego obrazov v rôznych farebných modeloch.

Faktor intenzity α		0,3	0,4	0,5	0,6	0,7
Detailné KO _D <i>Lena.bmp</i>	PSNR [dB]	54,122	53,306	52,556	51,842	51,195
	PSNR _{CSF} [dB]	71,982	71,298	70,556	69,826	68,994
	SSIM	0,9986	0,9983	0,9980	0,9976	0,9972
	Extrakcia dát [%]	99,861	99,943	100	100	100
	Počet chybných bitov	88/63248	36/63248	0/63248	0/63248	0/63248
Stredne detailné KO _{SD} <i>More.bmp</i>	PSNR [dB]	53,875	53,055	52,303	51,600	50,930
	PSNR _{CSF} [dB]	66,601	65,958	65,209	64,405	63,473
	SSIM	0,9983	0,9979	0,9975	0,9971	0,9966
	Extrakcia dát [%]	99,840	99,936	99,994	100	100
	Počet chybných bitov	100/63040	40/63040	4/63040	0/63040	0/63040
Málo detailné KO _{MD} <i>Lietadlo.bmp</i>	PSNR [dB]	54,672	53,829	53,088	52,384	51,722
	PSNR _{CSF} [dB]	62,307	61,376	60,585	59,812	59,083
	SSIM	0,9984	0,9980	0,9977	0,9973	0,9968
	Extrakcia dát [%]	99,898	99,938	99,994	100	100
	Počet chybných bitov	56/51152	28/51152	4/51152	0/51152	0/51152

Tab. 1 Experimentálne dosiahnuté výsledky navrhutej metódy vo farebnom modeli YC_bC_r

VI. ZÁVER

Pre farebný model YC_bC_r sú získané experimentálne výsledky uvedené v tab. 1. Na rozdiel od ďalších testovaných farebných obrazových modelov je možné konštatovať, že v prípade YC_bC_r modelu najmenšia degradácia stego obrazu nastala v prípade, keď ako krycie obrazy boli využívané obrazy s veľkým množstvom detailov. Najväčšia hodnota PSNR bola 52,5dB a pre PSNR_{CSF} to bolo až 70,55dB. Rovnako ako aj index štruktúrálnej podobnosti, ktorý bol približne v rozsahu 0,997 – 0,998 dosahoval vysoké hodnoty pre všetky typy krycích obrazov.

Kvalita stego obrazov je vysoká aj pri maximálnej zvolenej hodnote $\alpha=1,3$, o čom svedčia aj vysoké hodnoty PSNR, PSNR_{CSF} a SSIM. Navrhnutá metóda pri všetkých typoch testovaných krycích obrazov sa teda vyznačuje vysokou úrovňou perceptívnej transparentnosti vkladanej tajnej správy. Podľa subjektívneho hodnotenia autora, zvolené hodnoty α nespôsobili postrehnuteľné zhoršenie kvality obrazov a výskyt vnímateľných artefaktov. V prípade nízkej hodnoty faktora intenzity α však nie je zabezpečená požadovaná úspešnosť extrakcie vložených tajných správ. Za optimálne riešenie konfliktu extrakcie a perceptívnej transparentnosti vlozenej tajnej správy sa odporúča zvolenie hodnoty minimálne $\alpha=1,1$. Dosiahnuté výsledky ukazujú, že uvedená metóda vykazuje veľmi vysokú perceptívnu transparentnosť vo všetkých farebných obrazových modeloch.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] KATZENBEISSER, S. and PETICOLAS, F.: Information Hiding Techniques for Steganography and Digital Watermarking. *Artech House, INC*, 2000. ISBN1-58053-035-4.
- [2] PROAKIS, J. G.: Digital Communications. *McGraw Hill*, USA 1989.
- [3] Agrawal, N. – Gupta, A. 2009. DCT Domain Message Embedding in Spread-Spectrum Steganography System. In *Data Compression Conference*. ISSN: 1068-0314, 2009, p. 433-433.
- [4] HOPPER, N. J. – LANGFORD, J. – AHN, L. 2009. Provably Secure Steganography. In *Computers, IEEE Transactions*. ISSN: 0018-9340, 2009, vol. 58, no. 5, p. 662-676.
- [5] BRUNDICK, F. S. – MARVEL, L. M. 2001. Implementation of Spread Spectrum Image Steganography. *Army Research Laboratory*, Aberdeen Proving Ground, 2001, MD 210055067.
- [6] KER, A., D. 2005. Steganalysis of LSB matching in grayscale images. In *Signal Processing Letters, IEEE*. ISSN: 1070-9908, 2005, vol. 12, no. 6, p. 441-444.
- [7] SULEIMAN M. A. R. 2009. Colored Image-In-Image Hiding. In *CAD Systems in Microelectronics*. Lviv-Polyana: CADSM, 2009. Print ISBN: 978-966-2191-05-9 p. 445-450.
- [8] BUGÁR, G. – BÁNOCI, V. – LEVICKÝ, D. 2010. Information hiding using pseudo-random number sequences with error correction. In *Radioelektronika 2010: proceedings of 20th international conference : April 19-21*.

Steganografia v statických farebných obrazoch

*Vladimír BÁNOCI, Gabriel BUGÁR, Tomáš KANÓCZ,
Zita KLENOVIČOVÁ, Dušan LEVICKÝ*

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

vladimir.banoci@tuke.sk, gabriel.bugar@tuke.sk, tomas.kanocz@tuke.sk,
zita.klenovicova@tuke.sk, dusan.levicky@tuke.sk

Abstrakt — Článok je zameraný na problematiku utajovania informácií, resp. steganografie v statických farebných obrazoch. Jednotlivé časti príspevku opisujú základné princípy steganografie a perceptívne vhodné oblasti v statických farebných obrazoch. V ďalších kapitolách je rozpracovaná základná myšlienka navrhutej metódy. V závere sú uvedené experimentálne výsledky navrhutej metódy.

Kľúčové slová — steganografia, ukrývanie informácií, DCT, podprahový kanál, perceptívne vhodné oblasti, pseudonáhodné postupnosti

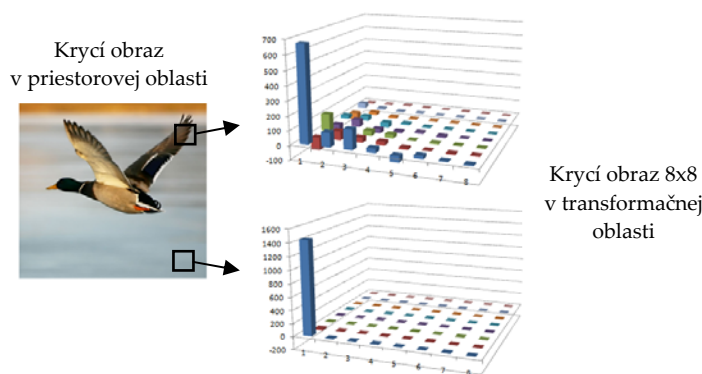
I. ÚVOD

Už od nepamäti sa ľudia pokúšali komunikovať pomocou rôznych prostriedkov, ako napr. reč, písmo takým spôsobom, ktorý by zabezpečoval ochranu obsahu pred neželanými osobami a ich vplyvu na prenášané správy. Jedným zo spôsobov na ochranu komunikácie a informácií je kryptografia, ktorej história siaha až do staroveku a zachovala sa až dodnes. V polovici 20. storočia bola využívaná na vojenské účely, ako ochrana komunikácie medzi vojenskými jednotkami. Základný princíp kryptografie spočíva v zašifrovaní tajnej správy pomocou kľúča. Takto zašifrovaná správa sa stáva nezrozumiteľnou bez použitia príslušného kľúča. Nevýhodou týchto šifrovaných správ je, že samotná komunikácia je viditeľná pre každého.

V súčasnosti sa pri ochrane tajných informácií kladie veľký dôraz práve na ich neviditeľnosť, teda na utajenie samotnej existencie takýchto informácií a práve to je základ vednej disciplíny zvanej steganografia (*steganography*). Základná myšlienka spočíva v utajení samotného faktu, že tajná informácia existuje, teda usiluje sa o vytvorenie tzv. "podprahového kanála" (*subliminal channel*), pomocou ktorej neviditeľná komunikácia s tajnými informáciami prebieha v pozadí viditeľnej, neutajenej a podozrenie nevzbudzujúcej komunikácie. Steganografia využíva bežne používané dáta (text, obrázky, audio alebo video) na utajenie tajných informácií, ktoré nevzbudzujú podozrenie o prenose informácií.

II. PERCEPTÍVNE VHODNÉ OBLASTI V KRYCÍCH OBRAZOCH

Ako perceptívne vhodné oblasti pre vkladanie tajných správ do krycích obrazov je možné označiť obrazové dáta s väčšou dynamikou obrazových prvkov. Väčšinou sú to hrany objektov v obraze, alebo detaily v obraze z pohľadu rozmiestnenia energie vo frekvenčnej oblasti. Monotónna plocha v obraze je tvorená obrazovými prvkami, ktoré sa so svojou hodnotou jasú navzájom výraznejšie nelíšia. Ich energia v transformačnej oblasti po aplikácii dvojrozmernej diskkrétnej kosínusovej transformácie bude sústredená do veľmi malého počtu spektrálnych transformačných koeficientov s najnižšími priestorovými frekvenciami v okolí jednosmernej zložky DC. Spektrálne koeficienty so strednými a vysokými priestorovými frekvenciami nie sú potom pre vkladanie tajných správ vhodné, vzhľadom na ich malú energiu, resp. na nedostatok redundancie. Následné modifikácie vykonané v týchto transformačných koeficientoch sú ľahšie detekovateľné. V oblasti hrán a detailov obrazu je rozloženie energie v transformovanej oblasti rovnomernejšie. Obsahuje dostatok energie a redundantnej zložky pri stredných a vyšších sekvenciách pre vloženie tajnej správy, pričom je zachovaná prijateľná robustnosť a nevnímateľnosť. Uvedené skutočnosti ilustruje obr. 1.

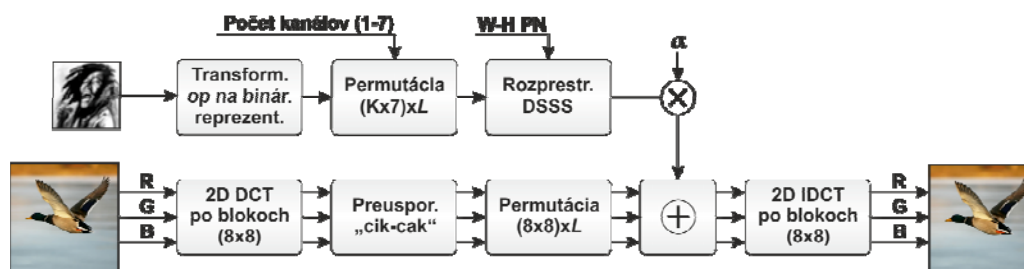


Obr. 1 Vplyv štruktúry obrázku na rozmiestnenie energie v transformovanej oblasti

III. NAVRHNUTÁ METÓDA MAXIMALIZÁCIE MNOŽSTVA VLOŽENÝCH TAJNÝCH DÁT

Navrhnutá metóda využíva pri vkladaní a extrakcii tajných správ, ktoré sú reprezentované statickými viacúrovňovými obrazmi do statických farebných obrazov, dvojrozmernú diskretnú kosínusovú transformáciu (2D DCT) realizovanú po blokoch 8×8 . V rámci týchto blokov sú modifikované určité transformačné koeficienty v závislosti od offsetu a veľkosti, resp. množstva dát vlozenej tajnej správy.

Bloková schéma metódy využívajúcej krycie obrazy pri extrakcii v prostredí 2D DCT transformácie pre viacúrovňové statické farebné obrazy je uvedená na obr. 2.

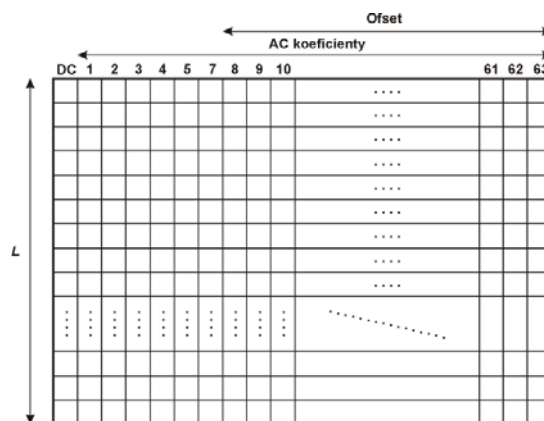


Obr. 2 Bloková schéma navrhnujetej metódy

Vstupmi do procesu vkladania tajných správ sú krycí statický farebný viacúrovňový obraz C s rozmerom $N_1 \times N_2$, tajná správa m s rozmerom $M \times M$ reprezentovaná statickým viacúrovňovým šedým obrazom a Hadamardová matica s Walsh-Hadamardovými pseudonáhodnými postupnosťami (W-H PN). Krycí obraz je najprv rozložený na jednotlivé farebné zložky v príslušnom farebnom obrazovom modeli a následne je rozdelený na bloky s veľkosťou 8×8 obrazových prvkov. Nasleduje transformácia každého bloku pomocou 2D DCT, čím sa získajú bloky DCT transformačných koeficientov C_{DCT} , ktorých počet je daný vzťahom:

$$L = \left(\left(\frac{N_1}{8} \right) \cdot \left(\frac{N_2}{8} \right) \right). \quad (1)$$

Matica transformačných koeficientov C_{DCT} je následne preusporiadaná pomocou následnosti „cik-cak“ do vektorov tak, že na prvej pozícii je vždy umiestnený DC koeficient a ostatné zložky sú tvorené AC koeficientmi s narastajúcimi priestorovými frekvenciami. Takto vytvorené vektory následne budú tvoriť maticu preusporiadaných transformačných koeficientov C'_{DCT} o rozmere $L \times 64$ (obr. 3). Počet transformačných koeficientov v tejto matici, ktoré nebudú ovplyvňované tajnou správou určuje offset (dôvod tejto skutočnosti bol uvedený v predošlej kapitole III.). Napr. v prípade, že sa offset zvolí rovný 8, tak prvých 8 koeficientov v matici transformačných koeficientov sa neovplyvní tajnými dátami a zabráni sa tak modifikácii transformačných koeficientov s najväčšou energiou. Pri úpravách tajnej správy reprezentovanej statickými viacúrovňovými obrazmi sa najprv pozmenia hodnoty obrazových prvkov z dekadického na binárny tvar a pre zvýšenie odolnosti sa ďalej modifikujú na hodnoty $(-1,1)$. Následne na základe zvoleného počtu kanálov K (možný počet 1-7) sa preusporiadajú bity tajnej správy na rozmer $(K \times 7) \times L$. Takto sú pripravené na rozprestrie pomocou pseudonáhodných postupností, kde každý informačný bit tajnej správy bude rozprestieraný PN postupnosťou.

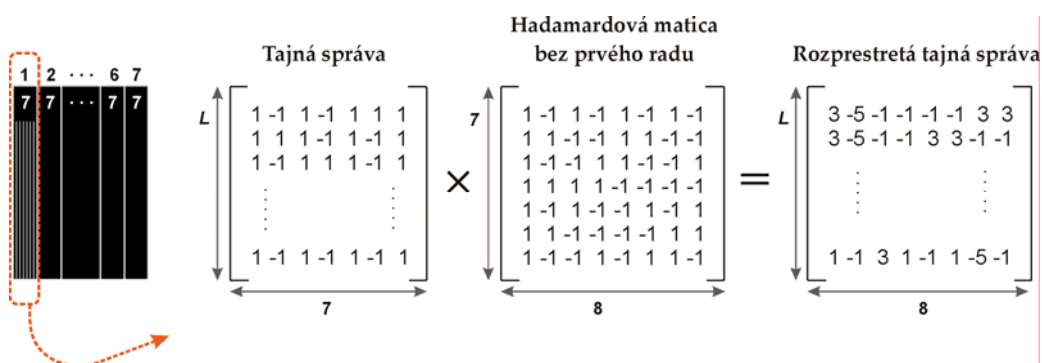


Obr. 3 Matica transformačných koeficientov po preusporiadaní metódou „cik-cak“

Pri použití viacerých kanálov interferuje väčší počet spektrálnych transformačných koeficientov a pri použití viacerých PN postupností sa tieto koeficienty ovplyvnia väčšou energiou. Metóda využíva Walsh-Hadamardove pseudonáhodné postupnosti dĺžky 8, z ktorých je prvý riadok vynechaný, keďže je tvorený samými jednotkami. Preto sa využíva iba zvyšných sedem PN postupností. Rozprestretá tajná správa m_{pn} sa ešte násobí škálovacím faktorom α a následne sa pripočíta k transformačným koeficientom krycieho obrazu C'_{DCT} . Posledným krokom je spätné preusporiadanie modifikovaných transformačných koeficientov pomocou následnosti „cik-cak“ a aplikácia inverznej 2D DCT transformácie, čím vzniká stego obraz C_m . Celý proces vkladania tajnej správy možno symbolicky vyjadriť v nasledujúcom tvare:

$$C_m(x, y) = C'_{DCT}(x, y) + \alpha \cdot m(x, y) \cdot PN(x, y). \quad (2)$$

Samotné rozprestieranie tajných dát je navrhnuté tak, že na základe zvoleného počtu PN postupností a počtu zvolených kanálov K sa výsledné rozprestreté dáta spočítajú a tým sa vytvoria nové kanály. Napríklad, ak sú zvolené dve PN postupnosti a dva kanály, tajná správa bude rozdelená na štyri rovnaké vektory. Prvé dva vektory tajných dát sú rozprestreté dvomi PN postupnosťami a sú spočítané s koeficientmi v prvom kanále. Druhé dva vektory sú rozprestreté rovnakými PN postupnosťami ako v prvom prípade a sú spočítané s koeficientmi v druhom kanále. Na obr. 4 je znázornený tento spôsob rozprestierania, kde bol vytvorený 1 kanál a použitých 7 PN postupností.



Obr. 4 Rozprestretie tajnej správy

Vstupmi do procesu extrakcie tajnej správy sú viacúrovňový farebný statický stego obraz C_m , počet použitých hadamardových pseudonáhodných postupností, počet zvolených kanálov K a statický viacúrovňový farebný krycí obraz, bez ktorého by extrakcia tajnej správy pri danej metóde nebola možná. Postup extrakcie tajnej správy je veľmi podobný procesu vkladania tajnej správy. Začína sa aplikáciou 2D DCT transformácie po blokoch 8×8 krycieho a stego obrazu do spektra transformačných koeficientov, ktoré sú následne preusporiadané podľa následnosti „cik-cak“. Takto získané bloky sú permutované na bloky o veľkosti $(8 \times 8) \times L$ a následne sa jednotlivé transformačné koeficienty na rovnakých pozíciách navzájom odčítajú, čím sú získané rozprestreté dáta tajnej správy m_{pn} . Takto získané výsledné dáta predstavujú rozprestretú tajnú správu ovplyvnenú chybami spôsobenými pri prenose, resp.

v tomto prípade chybami, ktoré vznikli dôsledkom zaokrúhľovania jasových úrovní po inverznej 2D DCT transformácie v procese vkladania tajnej správy. Z týchto dát sa získava vložená tajná správa koreláciou rozprestretých dát s Hadamardovými pseudonáhodnými postupnosťami.

$$m_Z(j) = \frac{1}{N} \sum_{i=1}^N m_{pn}[i] p_{n_H}[i], \quad (3)$$

kde $m_Z(j)$ - predstavujú dáta tajnej správy reprezentované bitmi po zvinutí,

$m_{pn}(i)$ - predstavuje dáta tajnej správy pred zvinutím,

$p_{n_H}(i)$ - predstavujú jednotlivé bity Hadamardovej pseudonáhodnej postupnosti.

Následne vhodne nastaveným komparátorom je získaný vektor tajných dát v binárnom tvare. Použitý komparátor s rozhodovacou úrovňou je opísaný nasledujúcimi rovnicami:

$$ak \quad m_Z(j) > 0 \quad potom \quad m(j) = 1, \quad (4)$$

$$ak \quad m_Z(j) \leq 0 \quad potom \quad m(j) = -1. \quad (5)$$

Takto získanú tajnú správu po prevode na dekadický tvar a preusporiadaní na pôvodný rozmer je možné zobrazit' vo forme viacúrovňového šedého obrazu.

IV. EXPERIMENTÁLNE VÝSLEDKY

Uvedená navrhnutá steganografická metóda vo farebných obrazoch využíva techniku rozprestretého spektra. Vkladanie tajnej správy je realizované modifikáciou vybraných transformačných koeficientov rozprestretou tajnou správou. Pri získavaní experimentálnych výsledkov boli použité krycie obrazy s priestorovou rozlišovacou schopnosťou 256×256 op v 4 rozličných farebných obrazových modeloch RGB, HSV, YCbCr a YIQ. Ako tajné správy boli použité statické šedé obrazy s priestorovou rozlišovacou schopnosťou od 40×40 až po 79×79 obrazových prvkov, ktoré predstavujú 12800 až 49928 bitov informácie. Obmedzenie maximálneho rozmeru tajnej správy 79×79 op vychádza z návrhu metódy. Krycí obraz s rozmerom 256×256 op po transformácii 2D DCT po blokoch 8×8 obsahuje 32×32 , teda 1024 blokov transformačných koeficientov. V každom bloku je na základe prednastaveného offsetu vynechaných prvých 8 z možných 64 transformačných koeficientov, ktoré predstavujú najnižšie priestorové frekvencie, čím na vkladanie tajnej správy zostáva 56 transformačných koeficientov. Z dĺžky použitých pseudonáhodných postupností, ktorá je v prípade matice Hadamardových PN rovná 8, vyplýva, že maximálny počet možných vytvorených kanálov je 7, pretože do 56 transformačných koeficientov je možné vložiť 7 PN postupností dĺžky 8. Použitím 7 pseudonáhodných postupností na jeden kanál, množstvo bitov, ktoré je možné vložiť je 7, t.j. každý blok je schopný ukryť 49 bitov tajnej správy. Pri počte blokov 1024 to bude 49×1024 bitov, čo limituje množstvo tajnej správy na 50 176 bitov. Z uvedeného vyplýva, že maximálny rozmer utajovanej správy môže byť $79 \times 79 \times 8$, čo predstavuje 49 928 bitov. Uvedené obmedzené množstvo informácie platí pre každú zložku farebného obrazu, pretože do každej zložky farebného obrazového modelu je vložená rovnaká informácia.

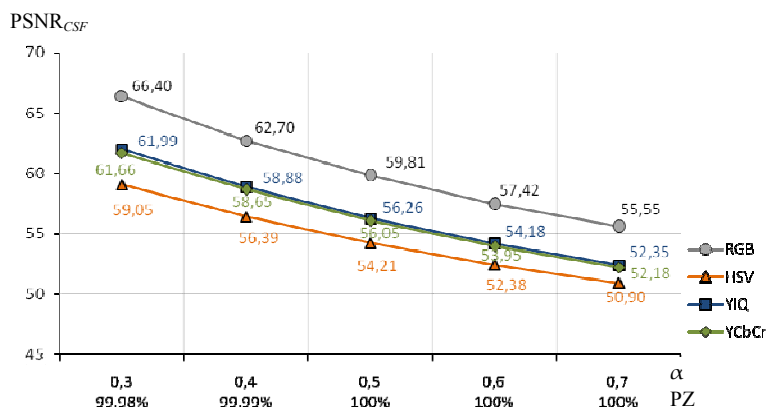
Miera modifikácie spektrálnych transformačných koeficientov je daná hodnotou parametra α , ktorý má konštantnú hodnotu pre všetky modifikované transformačné koeficienty. Predstavuje faktor intenzity tajnej správy v krycom obraze, resp. veľkosť energie vlozenej tajnej správy, ktorá ovplyvňuje perceptívnu transparentnosť. Väčšie hodnoty tohto parametra spôsobia väčšiu degradáciu pôvodnej kvality krycieho obrazu a tým zníženie perceptívnej transparentnosti vlozenej tajnej správy.

Na hodnotenie kvality stego objektov a miery vnímateľnosti tajných správ boli použité objektívne miery PSNR, $PSNR_{CSF}$ a SSIM.

V tab. 1 sú uvedené najlepšie dosiahnuté hodnoty objektívnych mier hodnotenia kvality s hodnotami percentuálnych zhôd extrahovaných tajných správ s vloženými. Pri experimentoch bolo zistené, že farebný obrazový model RGB poskytuje najlepšie výsledky pre danú navrhnutú metódu pre obrazy s rozdielnym množstvom detailov (D , SD a MD). Navrhnutá steganografická metóda patrí do skupiny závislých steganografických systémov, teda k úspešnej, 100% extrakcii tajnej správy zo stego obrazu je potrebný krycí obraz. Kvalita stego obrazov, pri ktorej je možné úspešne extrahovať tajnú správu z jednotlivých zložiek farebného obrazu, sa však môže navzájom líšiť. Možné odchýlky výsledkov vyplývajú predovšetkým z charakteru farebnej zložky a použitej hodnoty parametra α . Pre jednotlivé farebné zložky krycieho obrazu možno zvoliť rôzne hodnoty parametra α .

Faktor intenzity α		0,3	0,4	0,5	0,6	0,7
Detailné KO _D <i>Lena.bmp</i>	PSNR [dB]	50,117	47,877	46,077	44,548	43,263
	PSNR _{CSF} [dB]	66,400	62,701	59,811	57,429	55,558
	SSIM	0,9968	0,9936	0,9903	0,9863	0,9818
	Extrakcia dát [%] Počet chybných bitov	99,833 83/49928	99,997 1/49928	100 0/49928	100 0/49928	100 0/49928
Stredne detailné KO _{SD} <i>More.bmp</i>	PSNR [dB]	50,147	47,878	46,078	44,550	43,266
	PSNR _{CSF} [dB]	60,966	59,202	56,424	54,173	52,444
	SSIM	0,9939	0,9897	0,9846	0,9782	0,9709
	Extrakcia dát [%] Počet chybných bitov	99,831 84/49928	99,997 2/49928	100 0/49928	100 0/49928	100 0/49928
Málo detailné KO _{MD} <i>Lietadlo.bmp</i>	PSNR [dB]	50,117	47,877	46,077	44,548	43,263
	PSNR _{CSF} [dB]	55,847	53,412	51,500	49,847	48,526
	SSIM	0,7280	0,7246	0,7203	0,7149	0,7089
	Extrakcia dát [%] Počet chybných bitov	99,833 83/49928	99,997 1/49928	100 0/49928	100 0/49928	100 0/49928

Tab. 1 Experimentálne výsledky vo farebnom obrazovom modeli RGB



Obr. 5 Experimentálne najvyššie dosiahnuté hodnoty kvality stego obrazov pre jednotlivé farebné obrazové modely

Ako to naznačuje tab. 1, vo farebnom obrazovom modeli RGB percentuálna zhoda extrahovanej tajnej správy s vloženou tajnou správou, nezávisí od množstva detailov použitého krycieho obrazu. Maximálnu, 100% zhodu tajnej správy, bolo možné dosiahnuť pri nastavenej hodnote faktora intenzity $\alpha = 0,5$. Vzhľadom na to, že farebné zložky R, G a B obsahujú veľmi podobné hodnoty jasových zložiek, modifikácia ich transformačných koeficientov vyvolá veľmi podobnú degradáciu vo všetkých farebných zložkách krycieho obrazu. Táto degradácia stego obrazu je ale závislá od množstva miery detailov použitého krycieho obrazu. Z tab. 1 je ďalej zrejmé, že objektívna miera hodnotenia kvality stego obrazu špičkový pomer signálu ku šumu (PSNR), ktorý dosahoval hodnoty 46 dB, neodhalil badateľné zhoršenie kvality stego obrazu v závislosti od množstva detailov v krycom obraze. Objektívne miery PSNR_{CSF} a SSIM, ktoré sa lepšie prispôbujú k ľudskému vizuálnemu systému HVS ukazujú, že pri využití farebného obrazového modelu RGB, najmenšie skreslenie stego obrazu nastáva v prípadoch, keď tajná správa je vkladaná do krycieho obrazu s veľkým množstvom detailov. Pri maximálnej zhode extrahovanej a vlozenej tajnej správy boli zistené najvyššie hodnoty 58,9 dB pre PSNR_{CSF} a 0,99 pre SSIM.

V. ZÁVER

Pri analýze navrhnutéj steganografickej metódy boli všetky nastaviteľné parametre systému nastavené na hodnoty, pri ktorých je možné dosiahnuť maximálny kompromis medzi kapacitou systému a perceptívnou transparentnosťou vložených tajných správ do stego obrazov. Počet pseudonáhodných postupností z Hadamardovej matice postupností bol nastavený na hodnotu 7

a počet zvolených kanálov sa stanovil tiež na 7. Takto navolený steganografický systém umožní vložiť do jednej farebnej zložky krycieho obrazu 50 176 bitov tajnej správy, čo predstavuje viacúrovňový obraz o rozmere $79 \times 79 \times 8$.

Porovnaním získaných hodnôt objektívnych mier hodnotenia kvality stego obrazov so subjektívnym hodnotením týchto obrazov je možné konštatovať, že najlepšie výsledky pri aplikácii navrhutej metódy je možné dosiahnuť pre krycie obrazy vo farebných modeloch YIQ a RGB. V prípade YIQ sú to obrazy s malým množstvom detailov a v prípade RGB sú to obrazy s vysokým množstvom detailov.

Z dosiahnutých výsledkov vyplýva, že vizuálna kvalita stego obrazov sa pre všetky použité typy farebných modelov s narastajúcou hodnotou α zhoršuje a tým sa znižuje aj perceptívna transparentnosť, resp. detekovateľnosť vlozenej tajnej správy (obr. 5). Na základe subjektívneho hodnotenia sa výraznejšie zhoršenie vo vizuálnej kvalite neprejavil ani pri maximálnej použitej hodnote $\alpha = 0,8$. Nastavenie $\alpha = 0,5$ zabezpečuje vyhovujúcu rovnováhu vo všetkých farebných modeloch medzi týmito konfliktnými požiadavkami kladenými na stego obrazy a zároveň zabezpečuje 100% extrakciu vlozenej tajnej správy. Badateľnejšia degradácia stego obrazu nastala iba po dôkladnejšom porovnávaní detailných krycích obrazov so stego obrazmi vo farebnom obrazovom modeli $YCbCr$ a vo farebnom obrazovom modeli YIQ pri použití málo detailných krycích obrazov.

Na základe uvedených výsledkov navrhnutá metóda vkladania tajných správ do krycích obrazov vykazuje vysokú úroveň kapacity pri všetkých zvolených hodnotách parametra α . Je to v dôsledku toho, že každá farebná zložka je schopná ukryť 50 176 bitov informácie tajnej správy, čo v konečnom dôsledku predstavuje 3×50176 bitov teda 150 528 bitov a v prípade použitia iba jednej tajnej správy reprezentovanej statickým viacúrovňovým obrazom to bude obraz s rozmerom až $137 \times 137 \times 8$.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] KATZENBEISSER, S. and PETICOLAS, F.: Information Hiding Techniques for Steganography and Digital Watermarking. *Artech House, INC*, 2000. ISBN1-58053-035-4.
- [2] HASSAN, M. – GHAZY, M. R. A. – ABDULAZIZ, A. M. – IBRAHIM, K. 2009. A Novel Approach for Hiding Messages in Images. In *International Conference on Signal Acquisition and Processing*. ISSN: 1057-7149, 2009, vol. 16, no. 2, p. 391-405.
- [3] AGRAWAL, N. – GUPTA, A. 2009. DCT Domain Message Embedding in Spread-Spectrum Steganography System. In *Data Compression Conference*. ISSN: 1068-0314, 2009, p. 433-433.
- [4] RAJA, K. B. – SIDDARAJUL, S. – VENUGOPALL, K. R. – PATNAIK, L. M. 2007. Secure Steganography using Colour Palette Decomposition. In *Signal Processing, Communications and Networking, 2007. ICSCN '07. International Conference*. Print ISBN: 1-4244-0997-7, 2007, p. 74-80.
- [5] BUGÁR, G. 2008. Steganography in colour images. In *SCYR 2008 : 8th Scientific Conference of Young Researchers : proceeding from conference [CD-ROM]*. Košice : FEI TU, 2008. ISBN 978-80-553-0036-8, s. 126-128
- [6] SULEIMAN M. A. R. 2009. Colored Image-In-Image Hiding. In *CAD Systems in Microelectronics*. Lviv-Polyana: CADSM, 2009. Print ISBN: 978-966-2191-05-9 p. 445-450.
- [7] YUAN-HUI, Y. – CHIN-CHEN, C. – IUON-CHANG, L. 2007. A new steganographic method for color and grayscale image hiding. In *Computer Vision and Image Understanding*. ISSN: 1077-3142, 2007, vol. 107, no. 3, p. 183-194.
- [8] BUGÁR, G. – BÁNOCI, V. – LEVICKÝ, D. 2010. Information hiding using pseudo-random number sequences with error correction. In *Radioelektronika 2010: proceedings of 20th international conference : April 19-21, 2010, Brno, Czech Republic. - Brno: Brno University of Technology*. 2010. ISBN: 978-1-4244-6320-6, p. 1-4.
- [9] BÁNOCI, V. – BUGÁR, G. – LEVICKÝ, D. 2008. Moderná steganografia v dátových komunikačných systémoch. In *Manažérstvo bezpečnosti : zborník vedeckých prác*. Košice : Vysoká škola bezpečnostného manažérstva v Košiciach, 2008. ISBN: 978-80-89282-29-6, s. 40-51.
- [10] FORIŠ, P. 2006. *Perceptívne metódy digitálnej vodotlač v statických obrazoch* : dizertačná práca. Košice : KEMT FEI TUKE, 2006. 138 s.
- [11] BUGÁR, G. 2011. *Steganografia v statických farebných obrazoch* : dizertačná práca. Košice : KEMT FEI TUKE, 2011. 105 s.

Tvorba modelu slovenského jazyka pre doménovo orientovanú úlohu v systéme automatického rozpoznávania reči

Ján STAŠ, Jozef JUHÁR

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach,
Slovenská republika

jan.stas@tuke.sk, jozef.juhar@tuke.sk

Abstrakt – Systémy automatického rozpoznávania reči sa často navrhujú pre konkrétnu doménovo orientovanú úlohu, napr. na prepis rečových prejavov z oblastí medicíny, jurisdikcie a pod. V takýchto prípadoch je nevyhnutné uprietiť zvýšenú pozornosť aj na oblasť predspracovania textových dát, ktoré sa používajú na tréning jazykových modelov. Navyše výsledný jazykový model musí byť čo najviac prispôsobený oblasti jeho použitia. Z toho dôvodu je taktiež nevyhnutné uprietiť pozornosť aj na oblasť kombinácie a adaptácie jazykových modelov. Tento článok opisuje proces predspracovania textových dát, ktorý zahŕňa rad dodatočných a dôležitých úprav, ako aj samotný proces modelovania slovenského jazyka a adaptácie výsledného modelu na doménovo orientovanú úlohu v rozpoznávaní reči. Experimenty boli zamerané na vyhodnotenie chybovosti systému automatického rozpoznávania plynulej reči s veľkým slovníkom v reálnej úlohe rozpoznávania slovenského jazyka.

Kľúčové slová – adaptácia jazykového modelu, automatické rozpoznávanie plynulej reči s veľkým slovníkom, doménovo-zameraný jazykový model, predspracovanie textových dát

I. ÚVOD

Štatistické modelovanie jazyka je široko uplatniteľné v systémoch na automatické rozpoznávanie reči (ARR), štatistický strojový preklad, či v dialógových systémoch. V prípade systémov ARR, jazykový model plní veľmi dôležitú úlohu obzvlášť v prípade automatického rozpoznávania plynulej reči s veľkým slovníkom (v angl. „*large vocabulary continuous speech recognition*“, skr. LVCSR). Tento určuje zákonitosť formovania postupnosti slov, t.j. jej pravdepodobnosť, ako aj pravdepodobnosť slova samotného, ktorá korešponduje s vyslovenou akustickou informáciou, vyslovenou používateľom systému.

Jazykové modely sú zvyčajne tréňované na veľkej množine textových dát, ktorej veľkosť závisí od samotného jazyka, a ktoré sa zvyčajne získavajú z rôznych elektronických dokumentov alebo z internetových stránok. Keďže jazyk môže obsahovať veľké množstvo unikátnych slovných tvarov, ktorých počet je daný jazykom aplikácie, jazykové modely sú tak vytvárané na obmedzenej množine slovných tvarov. Tento jav je možné vysvetliť na príklade anglického jazyka, pri ktorom na dostatočné pokrytie slovnej zásoby postačuje približne 60 000 unikátnych slovných tvarov [1]. Zatiaľ čo v prípade vysoko flektívnych jazykov, akým je slovenčina, je na rovnaké pokrytie jazyka potrebných takýchto tvarov mnohonásobne viac.

Ďalším problémom slovenského jazyka je jeho neusporiadanosť, t.j. relatívne voľný slovosled vo vetách, ktorý v konečnom dôsledku vedie k nedostatočnosti dát používaných v procese tréningu jazykového modelu. Z toho dôvodu je nevyhnutné zozbierať také množstvo textových dát, ktoré dokážu dostatočne štatisticky reprezentovať daný jazyk. Problém nedostatočnosti dát sa snažia v nemalej miere odstrániť aj *metódy vyhladzovania jazykového modelu* (v angl. „*smoothing methods*“) [2], ktoré prerozdeľujú časť pravdepodobností veľmi frekventovaných javov medzi javy málo frekventované.

V prípade tvorby jazykového modelu pre aplikácie prebiehajúce v reálnom čase je tiež nevyhnutné ohraničiť veľkosť výsledného modelu, pretože s narastajúcou slovnou zásobou vzrastá aj množstvo možných n -gramov v modeli. Niektoré z nich sa v konečnom dôsledku môžu vyskytovať v jazyku málo frekventovane a nemajú veľký vplyv na kvalitu samotného jazykového

modelu. Z toho dôvodu je na odstránenie takýchto javov často vhodné použiť rôzne *techniky prerezávania jazykového modelu* (v angl. „*pruning techniques*“) [3].

Súčasný stav problematiky v oblasti modelovania jazyka je zameraný na zvyšovanie výkonnosti LVCSR systému použitím *adaptácie jazykového modelu* (v angl. „*language model adaptation*“) [4] [5], obzvlášť v prípade doménovo zameraného rozpoznávania reči. Zvyčajne sa to vykonáva adaptáciou všeobecného modelu na používateľskú doménu pomocou malého, doménovo zameraného modelu, kedy adaptačné dáta sú tiež použité na nastavenie parametrov všeobecného jazykového modelu.

Tento článok je zameraný na modelovanie slovenského jazyka v úlohe doménovo zameraného rozpoznávania reči v súdnictve. Proces predspracovania textových dát zahŕňa rad dodatočných a dôležitých úprav, napr. problém anonymizácie osobných údajov osôb vystupujúcich v zápisoch zo súdnych pojednávaní. Adaptácia jazykového modelu na oblasť súdnictva bola vykonaná aplikáciou lineárnej interpolácie doménovo zameraného modelu so všeobecným jazykovým modelom, pričom výsledný model bol mierne modifikovaný modelom vytvoreným z textových dát anotovaných akustických nahrávok, používaných pri tvorbe akustického modelu. Experimentálne výsledky opísané v záverečnej časti ukazujú výrazné zníženie chybovosti LVCSR systému v prípade navrhutej metodiky spracovania textových dát a adaptácie jazykového modelu.

II. PREDSPRACOVANIE TEXTOVÝCH DÁT

A. Doménovo-zameraný textový korpus

Jednou z dôležitých úloh pri tvorbe jazykového modelu je spracovať textové dáta do možno čo najlepšej celoslovnej podoby. V prípade spracovania textových dát z oblasti súdnictva sme narazili na niekoľko problémov. V týchto textových dokumentoch sa vyskytovalo veľké množstvo špecifických abreviácií, akronymov, symbolov, ako aj časových a číselných údajov. V určitých prípadoch boli slová tiež zvýrazňované pridaním medziznakových medzier. Množstvo mien, priezvisk, a osobných údajov bolo anonymizovaných. Z toho dôvodu bolo nutné tieto špecifické prípady ošetriť tak, aby výsledný textový korpus bol použiteľný v procese tréningu jazykového modelu. V nasledujúcich častiach budú tieto dodatočné úpravy textu v krátkosti opísané.

Zvýraznenie slov. Bolo detekovaných a automaticky ošetrených približne 150 rôznych slov zvýraznených medziznakovými medzerami v nasledujúcom tvare

S c h v a Ľ u j e z m i e r	~	Schvaľuje zmier
R o z s u d o k	~	Rozsudok
r o z h o d o l	~	rozhodol

Abreviácie. 400 najviac frekventovaných abreviácií bolo manuálne ošetrených na základe ich okolitého kontextu za pomoci približne 1 400 regulárnych výrazov nasledovne

Podľa ust. § 3	~	Podľa ustanovenia paragrafu tri
ods. 2 Obch. z.	~	odsek dva Obchodného zákonníka
v § 8 ObZ.	~	v paragrafe osem Obchodného zákonníka

Anonymizácia názvov entít. Bol navrhnutý automatický generátor názvov entít z čiastkových zoznamov mien, priezvisk, názvov ulíc, miest a obcí v Slovenskej republike. Tieto vzory boli automaticky detekované na základe kľúčových slov a nahradené ich adekvátnou formou na základe ich okolitého kontextu v nasledujúcom tvare

JUDr. D. H.	~	JUDR. Danou Hlavatou
bytom v N., ul. K.	~	bytom v Nitre, ulica Krivá
so sídlom v B. B.	~	so sídlom v Banskej Bystrici

Číselné údaje boli prepísané do ich celoslovnej podoby na základe ich okolitého kotextu. Taktiež bolo nutné rozhodnúť o tom, ktorá z detekovaných čísloviek je základná, a ktorá radová, a tiež ktoré z čísiel bolo vhodnejšie prepísať do celoslovnej podoby po samostatnej číslici.

s 20%-ným úrokom	~	s dvadsať percentným úrokom
118,- Sk do 3 dní	~	sto osemnásť SK do troch dní
Okr. súd Košice II	~	Okresný súd Košice rímska dva

Zjednotenie akronymov. Akronymy, abreviácie, skratky titulov alebo izolované písmená sú zvyčajne hláskované po jednotlivých písmenách. Z dôvodu ich lepšieho rozlíšenia od jednotlivých slov sme takéto prípady konvertovali na veľké písmená.

Eur	~	EUR	rtg	~	RTG
CSc.	~	CSC.	PhDr.	~	PHDR.
osp	~	O.S.P.	s.r.o.	~	S.R.O.

Je vhodné poznamenať, že v určitých prípadoch sa niektoré často vyskytované skratky titulov vyslovujú v ich celoslovnej podobe. Preto boli niektoré z nich vo výslednom textovom korpuse prepísané do celoslovnej podoby, ako aj ponechané v ich skrátenej forme.

Tabuľka 1
 ŠTATISTIKA KORPUSOV TEXTOVÝCH DÁT.

Druh textového korpusu	Počet viet	Počet slov	Veľkosť slovníka
Webový korpus	86 872 192	1 195 033 631	328 693
Doménovo orientovaný korpus	5 913 142	184 534 662	148 999
Korpus anotácií	20 009	234 016	15 672

B. Textový korpus anotovaných nahrávok

Špeciálnym prípadom textových korpusov sú texty anotácií z akustických dát používaných pri tvorbe akustického modelu. Tieto anotácie obsahujú veľké množstvo šumových značiek obsiahnutých priamo v texte. Experimentálne sme zistili, že tieto šumové javy majú pozitívny vplyv na kvalitu výsledného jazykového modelu, pretože rečník takéto šumy, ako napr. zaváhanie, kašeľ a pod., často v reči používa. Z toho dôvodu sme sa rozhodli zahrnúť šumové značky aj do tréningového procesu. Keďže existuje veľké množstvo šumových značiek v anotovaných textoch, zaradili sme ich preto do piatich veľkých skupín

[@]	krátka pauza
[@:]	dlhá pauza
[fil]	hlboký nádych, zakašľanie, prežúvanie, kýchnutie, a pod.
[int]	šum v pozadí
[sil]	zaváhanie, t.j. hmm, ehmm, a pod.

Tieto boli následne zahrnuté do slovníka používaného pri tréningu jazykového modelu a v procese rozpoznávania LVCSR systémom boli považované za transparentné slová.

C. Webový korpus textových dát

Okrem doménovo zameraného korpusu je nevyhnutné zozbierať aj veľké množstvo textových dát zo všeobecnej oblasti s cieľom dostatočne štatisticky reprezentovať daný jazyk. Preto sme v *Lab. rečových technológií v telekomunikáciách* vyvinuli automatický systém na zber textových dát, zvaný *webAgent* [6]. Tento systém získava textové dáta z rôznych webových stránok a iných elektronických zdrojov textu, ktoré sú písané v slovenskom jazyku. Následne sú tieto automaticky spracované pomocou dodatočných nástrojov na segmentáciu textu do viet, na expanziu skratiek, prepis čísiel a pod. Systém tiež zahŕňa rad obmedzení v podobe filtrácie gramaticky nesprávnych slov, kontroly duplicitných textových dokumentov a i. [7]. Je vhodné poznamenať, že webový korpus bol následne dodatočne spracovaný podľa metodiky opísanej v kap. II-A.

III. MODELOVANIE JAZYKA

Ako bolo spomenuté v predchádzajúcej časti, spracované textové dáta boli rozdelené do troch základných korpusov: *a.* webový korpus použitý k natrénovaniu všeobecného jazykového modelu; *b.* korpus z oblasti súdnictva na natrénovanie doménovo zameraného modelu a *c.* korpus anotácií, v ktorom sú obsiahnuté aj šumové javy. Použitý slovník bol vygenerovaný z každého korpusu zvlášť a prešiel automatickou, ako aj manuálnou kontrolou pravopisu. Štatistiku textových korpusov, t.j. počet viet, počet slov ako aj veľkosť slovníkov je možné vidieť v Tab. 1.

K tvorbe čiastkových jazykových modelov z jednotlivých textových korpusov boli použité štandardné programové nástroje sady *SRI Language Modeling Toolkit* [8]. Trigramové jazykové modely boli obmedzené slovníkmi rôznej veľkosti (viď Tab. 1) a vyhladené pomocou *modifikovaného Kneserovho-Neyovho algoritmu* (v angl. „*modified Kneser-Ney algorithm*“) [2].

IV. ADAPTÁCIA JAZYKOVÉHO MODELU

Jednou zo štandardných a účinných metód kombinácie a adaptácie doménovo zameraného jazykového modelu so všeobecným modelom je *metóda lineárnej interpolácie* (v angl. „*linear interpolation*“, skr. LI) [1], ktorá je definovaná vzťahom

$$P_{LI}(w|h) = (1 - \lambda)P_A(w|h) + \lambda P_B(w|h), \quad (1)$$

kde $P_A(w|h)$ označuje pravdepodobnosť slova w v závislosti od jeho histórie h v malom adaptačnom korpusu a $P_B(w|h)$ udáva pravdepodobnosť slova w v rozsiahlom všeobecnom jazykovom modeli. V tomto prípade je výsledná pravdepodobnosť daná váženým súčtom týchto dvoch modelov. Hodnota interpolačnej váhy λ ($\lambda \geq 0$) môže byť jednoducho vypočítaná pomocou „*expectation-maximization*“ (EM) algoritmu [4] maximalizáciou vierohodnosti na odloženej množine textových dát (v angl. „*held-out data*“), príp. pomocou algoritmu založeného na *minimalizácii perplexity modelu* (v angl. „*perplexity minimization*“) [8] vyčíslenej na množine odložených dát, alebo môže byť stanovená empiricky.

Tabuľka 2
 VPLYV PREREZÁVANIA NA POČET N-GRAMOV A VEĽKOSŤ VÝSLEDNÉHO JAZYKOVÉHO MODELU PO ADAPTÁCII.

Jazykový model	Počet bigramov	Počet trigramov	Veľkosť modelu [MB]
Model pred prerezaním	77 078 469	90 557 054	5 144
Model po prerezaní	12 181 367	7 910 370	601

V našom prípade všetky tri modely boli lineárne interpolované súčasne, pričom interpolačné váhy boli určené algoritmom minimalizácie perplexity určenej na množine odložených dát (približne 5% textových dát z oblasti súdnictva). Po adaptácii vzrástla veľkosť slovníka na hodnotu 329 949 jedinečných slov.

Keďže výsledný adaptovaný model bol príliš obsiahly pre prípad systému LVCSR pracujúceho v reálnom čase, na prerezanie modelu bola použitá *metóda založená na relatívnej entropii* (v angl. „*relative-entropy based pruning*“) [3]. Porovnanie počtu bigramov a trigramov pred a po prerezaní modelu je možné vidieť v Tab. 2. Hodnota prahu prerezávania (10^{-10}) bola stanovená empiricky zohľadňujúc podmienku fungovania systému v reálnom čase.

V. NASTAVENIA LVCSR SYSTÉMU

Ako už bolo spomenuté v kap. III a kap. IV, boli vytvorené tri rôzne jazykové modely, obmedzené slovníkom a vyhladené modifikovaným Kneserovým-Neyovým algoritmom, natréňované pomocou troch rôznych textových korpusov. Následne boli kombinované použitím lineárnej interpolácie, pričom interpolačné váhy boli prispôbené na oblasť súdnictva. Výsledný jazykový model bol prerezaný pomocou metódy založenej na relatívnej entropii s cieľom jeho použitia v doménovo orientovanej aplikácii systému LVCSR, pracujúceho v reálnom čase.

Použitý bol trifónový kontextovo závislý akustický model vytvorený na báze *skrytých Markovových modelov* (v angl. „*hidden Markov models*“, skr. HMM), ktorého každý stav bol modelovaný 32 Gaussovými zmesami. Model bol vytvorený z príznakových vektorov, tvorených 39 *mel-frekvenčnými keprálnymi koeficientami* (v angl. „*mel-frequency cepstral*“, skr. MFC). Natréňovaný bol na anotovaných rečových nahrávkach v rozsahu 170 hodín súdnych pojednávaní a parlamentnej reči. Všetky možné skutočne rozpoznané trifóny boli mapované do menšej množiny obsahujúcej približne 2 000 rôznych trifón. Trénovacia množina zahŕňala tiež model ticha, model krátkej pauzy a ďalšie dodatočné šumové javy [9].

Na samotné rozpoznávanie reči bol použitý vysoko-výkonný LVCSR dekodér [10], využívajúci dvojpásmovú rozpoznávaciu stratégiu. V tomto prípade je v smere zľava doprava použitý bigramový jazykový model, pričom pri spätnom prehľadávaní sú výsledné hypotézy rozpoznávania z priameho smeru dotvorené a vyhodnotené trigramovým jazykovým modelom.

Testovacie dáta reprezentovali nahrávky súdnych pojednávaní, ktoré obsahovali 41 862 slov v 3 426 vetách a frázach. Do testovania sme zahrnuli tiež frázy, pretože rečník často v reálnych podmienkach umiestňuje dlhé pauzy nielen na hraniciach viet, ale aj na hraniciach určitých fráz.

Na hodnotenie jazykových modelov sa zvyčajne používajú dve štandardné miery hodnotenia, a to: *a.* extrinžické hodnotenie na základe *chybovosti systému* (v angl. „*word error rate*“, skr. WER) a *b.* intrinžické hodnotenie pomocou *perplexity* (PPL), vyhodnotenej na testovacích dátach. WER je štandardnou mierou hodnotenia výkonnosti LVCSR systémov a udáva sa ako pomer hypotéz referenčného textu k textu po rozpoznaní, ktorý v sebe zahŕňa chyby vložení, zmazaní a chyby zámenou. V prípade, že nemáme k dispozícii LVCSR systém, na hodnotenie sa používa miera, zvaná perplexita, resp. zložitost' modelu. Perplexita je definovaná ako reciproká hodnota váženého (geometrického) priemeru pravdepodobnosti určenú modelom pre každé slovo v testovacích dátach. Táto miera však nemusí nevyhnutne sama o sebe odrážať hodnotu presnosti rozpoznania, ale zvyčajne s ňou vysoko koreluje.

VI. EXPERIMENTÁLNE VÝSLEDKY

Experimenty s jazykovým modelom boli orientované na vyhodnotenie WER a PPL na testovacích dátach, ktoré obsahovali reálne nahrávky súdnych pojednávaní. Experimentálne výsledky adaptovaného jazykového modelu sme porovnali s modelom, ktorý nezahŕňal dodatočné úpravy pri predspracovaní textových dát a nebol ani tematicky adaptovaný na oblasť súdnictva. Je vhodné poznamenať, že množina trénovacích dát bola v oboch prípadoch rovnaká. Aby bolo možné oba modely korektne ohodnotiť, testovacie dáta pre adaptovaný model boli taktiež upravené podľa metodiky predspracovania textových dát uverejnenej v kap. II-A.

Tabuľka 3
VÝSLEDOK EXPERIMENTÁLNEHO TESTOVANIA JAZYKOVÝCH MODELOV.

Jazykový model	Počet bigramov	Počet trigramov	PPL	WER [%]
Model bez adaptácie	14 130 666	8 202 437	143,8	9,03
Model s adaptáciou	12 181 367	7 910 370	96,7	7,63

Napriek tomu, že jazykový model bez adaptácie bol o trochu väčší než model po adaptácii v počte bigramov a trigramov, je možné vidieť v Tab. 3, že v prípade adaptácie modelu na doménu, ako aj lepším predspracovaním textových dát sme dosiahli výrazné zníženie chybovosti WER LVCSR systému, ako aj zníženie perplexity PPL výsledného modelu. Je vhodné poznamenať, že po porovnaní referenčného textu s textom po rozpoznaní, chyba LVCSR systému sa prejavila väčšinou v prípadoch, kedy medzi slovami nastáva jav spodobovania, alebo sú rozdielne, výslovnostne príbuzné slová v rečovom prejave vyslovované rovnako. Napriek týmto chybám sme dosiahli zníženie chybovosti WER približne o 1,5% v porovnaní s modelom bez adaptácie (viď Tab. 3). Avšak tento poznatok bude predmetom ďalšieho výskumu pri znižovaní chybovosti WER systému na automatické rozpoznávanie plynulej reči s veľkým slovníkom, a taktiež na zvýšenie kvality modelu slovenského jazyka v úlohe doménovo orientovanej aplikácie rozpoznávania reči prebiehajúcej v reálnom čase.

VII. ZÁVER

V tomto článku bola v krátkosti predstavená nová metodika predspracovania textových dát s cieľom zvýšiť kvalitu modelu slovenského jazyka v úlohe doménovo orientovaného rozpoznávania reči. Zlepšenie jazykového modelu sme dosiahli tiež adaptáciou modelu pomocou lineárnej interpolácie s nastavením interpolačných váh na odloženej množine textových dát z oblasti súdnictva. Napriek chybám v rozpoznaní, výsledný jazykový model je použiteľný v aplikácii systému LVCSR, pracujúcej v reálnom čase.

Nasledujúci výskum bude zameraný na ďalšie zvyšovanie kvality modelu slovenského jazyka ešte lepším predspracovaním textových dát, zväčšením množiny adaptačných dát, príp. zameraním výskumu na jazykové modely obsahujúce viacсловné výrazy, s cieľom odstrániť vyššie spomenuté chyby pri automatickom rozpoznávaní reči so zreteľom na slovenský jazyk.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (100%).

LITERATÚRA

- [1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd edition), Prentice Hall, Pearson Education, pp. 998, ISBN 978-0-13-504196-3, 2009.
- [2] S. F. Chen and J. Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling*, Tech. Report TR-10-98, Computer Science Group, Cambridge, Harvard University, Massachusetts, pp. 1–63, August 1998.
- [3] A. Stolcke, *Entropy-based Pruning of Backoff Language Models*, In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, pp. 270–274, February 1998.
- [4] J. Gao, H. Suzuki and W. Yuan, *An Empirical Study on Language Model Adaptation*, In: ACM Transaction on Asian Language Information Processing, TALIP'06, New York, USA, Vol. 5, No. 3, pp. 209–227, 2006.
- [5] J. Staš, D. Hládek and J. Juhár, *Language Model Adaptation for Slovak LVCSR*, In: Proceedings of International Conference on Applied Electrical Engineering and Informatics, AEI'10, Venice, Italy, pp. 101–106, ISBN 978-80-553-0519-6, September 2010.
- [6] D. Hládek and J. Staš, *Text Mining and Processing for Corpora Creation in Slovak Language*, Journal of Computer Science and Control Systems, Vol. 3, No. 1, pp. 65–68, ISSN 1844-6043, 2010.
- [7] J. Staš, D. Hládek, M. Pleva and J. Juhár, *Slovak Language Model from Internet Text Data*, A. Esposito et al. (Eds.): Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues, LNCS 6456, Springer-Verlag, Berlin, Heidelberg, pp. 352–358, ISBN 978-3-642-18183-2, 2010.
- [8] A. Stolcke, *SRILM – An Extensible Language Modeling Toolkit*, In: Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP'02, pp. 901–904, September 2002.
- [9] R. Sabo and S. Darjaa, *Automatický prepis rokování Národní rady Slovenské republiky – rečová databáza, akustické modely, rozpoznávání reči*, In: Proceedings of the 15th International Acoustic Conference Noise and Vibration in Practice, Kočovce, Slovak Republic, pp. 107–110, ISBN 978-80-227-3305-2, May 2010.
- [10] T. Lee, T. Kawahara and K. Shikano, *Julius – An Open Source Real-time Large Vocabulary Recognition Engine*, In: Proceedings of European Conference on Speech Communications and Technology, EUROSPEECH'01, pp. 1691–1694, 2001.

Detekcia akustických udalostí na báze rečových parametrizácií a GMM

Eva VOZÁRIKOVÁ, Anton ČIŽMÁR

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

eva.vozarikova@tuke.sk, anton.cizmar@tuke.sk

Abstrakt — V súčasnosti narastá záujem o implementáciu systémov detekujúcich akustické udalosti. Tento záujem je sústredený najmä na oblasť bezpečnosti. Detekciou špecifických zvukov je možné vytvoriť systémy, ktoré budú poskytovať informácie o výskyte týchto zvukov. Pre bezpečnostné aplikácie je záujem sústredený na zvuky, ktoré môžu predstavovať nebezpečenstvo. Tento článok je zameraný na detekciu abnormálnych zvukov konkrétne zvuku výstrelu zo zbrane a zvuku rozbíjania skla. V článku je popísaný základný prístup a metodika detekcie akustických udalostí. Obsahuje popis vybraných metód výpočtu charakteristických príznakov ako Mel-frekvenčných kepstrálnych koeficientov (Mel-Frequency Cepstral Coefficients - MFCC), koeficientov percepčnej lineárnej predikcie (Perceptual Linear Prediction coefficients - PLP) a lineárne predikčné kepstrálne koeficienty (Linear Prediction Cepstral Coefficients - LPCC), výkonnosť ktorých je overená pomocou Gaussových modelov (Gaussian Mixture Models - GMMs).

KLúčové slová — akustická udalosť, GMM, LPCC, MFCC, PLP

I. ÚVOD

Verejné a súkromné priestory sú bežne vybavené kamerovými systémami, v ktorých akustická informácia nie je využitá. Niektoré situácie je jednoduchšie detekovať pomocou zvukovej než obrazovej informácie. Príkladom takýchto zvukov je zvuk výstrelu zo zbrane, výkrik, volanie o pomoc, atď. Preto bezpečnostné systémy, ktoré pracujú s obrazovou a zvukovou informáciou sú v súčasnej dobe veľmi žiadanými systémami najmä pre ich vyššiu účinnosť v porovnaní so systémami pracujúcimi iba s jedným druhom informácie. Aj z tohto dôvodu detekcia akustických udalostí predstavuje výskumnú oblasť s vysokým potenciálom. Systémy detekcie akustických udalostí [1], [2], [3], majú potenciál chrániť zdravie, súkromný a verejný majetok prostredníctvom generovania varovných hlásení pri výskyte potenciálne nebezpečných zvukov. Okrem použitia v oblasti bezpečnosti majú tieto systémy narastajúci význam aj v iných odvetviach napr. v priemyselných aplikáciách na odhaľovanie skrytých väd, kde sa odchýlka od správnej činnosti zariadenia prejaví zmenou charakteristického zvuku, ďalej v oblasti starostlivosti o zdravie pacientov, kde je snahou uľahčiť prácu zdravotnému personálu a pacientom zabezpečiť vyšší komfort služieb. Tieto systémy je možné použiť aj pri diagnostike niektorých ochorení. Taktiež nemalý prínos majú aj pre osamelo žijúcich starších ľudí, kde systém môže privolať pomoc napr. pri detekcii takého zvuku akým je pád. Veľkou oblasťou, kde je možné využiť systémy detekcie akustických udalostí, je ich aplikácia v inteligentných domoch, kde je záujem sústredený na ovládanie niektorých zariadení alebo systému ako celku. Ďalšie využitie nachádzajú v priestoroch konferenčných miestností, kde sa využívajú napr. na monitorovanie priebehu stretnutí, atď. [4].

Pojmom akustická udalosť je nazvaný krátky segment zvuku, ktorý má zriedkavý výskyt, pričom tento výskyt nevieme predpovedať. Samozrejme, daný zvuk je relevantný pre konkrétnu aplikáciu.

Z pohľadu spoľahlivej činnosti detekčného systému je veľmi dôležité redukovať varovné hlásenie vplyvom nesprávnej klasifikácie tzv. false alarm. Systém by mal detekovať varovné hlásenie iba v prípade, že sú rozpoznané konkrétne akustické udalosti. Nemalý problém spôsobuje aj rozpoznávanie podobne znejúcich zvukov, z ktorých jeden predstavuje akustickú udalosť a druhý zvuk spôsobí chybnú klasifikáciu. Príkladom takýchto podobne znejúcich zvukov je napr. prudké brzdenie automobilu (akustická udalosť) a naopak rýchly rozbeh, ktorý vo väčšine prípadov nepredstavuje nebezpečenstvo.

Vo všeobecnosti, každý detekčný systém je z principiálneho hľadiska možné rozdeliť na dve časti, t.j. na časť extrakcie charakteristických príznakov a na časť, v ktorej sa vykonáva klasifikácia. Efektívny algoritmus výpočtu relevantných príznakov hrá v kľúčovú úlohu pri detekcii a klasifikácii akýchkoľvek zvukov, pretože rozpoznávací proces priamo závisí na kvalite extrahovaných parametrov zo vstupného akustického signálu.

K problematike rozpoznávania špecifických zvukov, akými je aj reč, bolo vykonaných aj publikovaných veľa vedeckých prác. Na parametrickú reprezentáciu rečového signálu bolo vytvorených a otestovaných viacero metód. Niektoré z nich poskytujú veľmi efektívnu a kompaktnú reprezentáciu rečového signálu. K týmto základným a veľmi populárnym metódam výpočtu charakteristických príznakov patrí predovšetkým reprezentácia akustického signálu pomocou Mel-frekvenčných keprálnych koeficientov (Mel-Frequency Cepstral Coefficients - MFCC), ďalej pomocou koeficientov percepčnej lineárnej predikcie (Perceptual Linear Prediction coefficients - PLP) a tiež pomocou lineárne predikčných keprálnych koeficientov (Linear Prediction Cepstral Coefficients - LPCC). Rečový signál má najviac informačného obsahu sústredeného pri nižších frekvenciách a štandardne používaná vzorkovacia frekvencia je 8kHz alebo 16kHz. Akustické udalosti sú charakteristické inými vlastnosťami, napr. informačný obsah je sústredený prevažne v oblasti stredných a vyšších frekvencií, z toho vychádza aj potreba použitia vyšších vzorkovacích frekvencií 24kHz, resp. 48kHz. Ako vyplýva z predchádzajúcich riadkov, rozdielny spektrálny a časový priebeh pre reč a akustické udalosti môže viesť k použitiu rozdielnych prístupov k výpočtu charakteristických príznakov.

Na výpočet príznakov akustických udalostí je možné použiť metódy známe z rozpoznávania reči [5], [6], [7], z ktorých sa veľmi často využívajú MFCC [8], PLP [8], LPCC [9], atď. Tieto prístupy primárne navrhnuté pre rečový signál dosahujú dobrú výkonnosť aj pre takú skupinu zvukov akými sú akustické udalosti. Ďalší veľmi populárny prístup reprezentácie akustických udalostí je založený na nízkoúrovňových deskriptoroch definovaných v rámci štandardu MPEG-7 [10]. 17 nízkoúrovňových deskriptorov môže byť použitých na tento účel, napr. deskriptor spektrálneho ťažiska (Audio Spectral Centroid - ASC), deskriptor plochosti spektra (Audio Spectral Flatness - ASF), deskriptor popisujúci spektrálnu obálku (Audio Spectral Envelope - ASE), deskriptor základnej frekvencie (Audio Fundamental Frequency - AFF), deskriptor výkonu signálu (Audio Power - AP), atď. Pre systémy detekcie akustických udalostí môže byť veľmi výhodné, aby množina charakteristických parametrov bola tvorená kombináciou rečových príznakov a MPEG-7 deskriptorov.

Pri klasifikácii akustických udalostí sa využívajú dva hlavné prístupy. Prvý z nich je založený na generatívnych modeloch akými sú skryté Markovove modely (Hidden Markov Models - HMM) alebo Gaussove modely zmesí (Gaussian Mixture Models - GMM). Druhý prístup je založený na diskriminatívnej metóde napr. metóde podporných vektorov (Support Vector Machine - SVM), ktoré predstavujú veľmi populárny klasifikačný prístup [11], [12].

Tento článok je zameraný na detekciu dvoch typov akustických udalostí: výstrelov zo zbrane a zvuk rozbiť skla. V úlohe klasifikátora sú použité GMM modely s rôznym počtom zmesí (Probability Density Function - PDFs). Rozličné metódy extrakcie príznakov ako MFCC, PLP, LPCC sú stručne uvedené v tomto článku.

Článok je rozdelený podľa nasledovnej štruktúry: sekcia II. popisuje extrakciu príznakov, sekcia III. podáva základné informácie o použitom klasifikátore, v IV. sekcii sú zhrnuté informácie o použitej databáze zvukov a v sekcii V. sú uvedené bližšie informácie o vykonaných experimentoch, za ktorými nasleduje VI. sekcia popisujúca dosiahnuté výsledky. Zhrnutie poznatkov je možné nájsť v závere tohto článku.

II. EXTRAKCIA PRÍZNAKOV

Extrakcia (výpočet) charakteristických príznakov je realizovaná v bloku extrakcie príznakov. Vstupom do tohto bloku je akustický signál, výstupom sú parametre popisujúce daný akustický signál. Spôsob výpočtu charakteristických príznakov závisí od zvolenej parametrizačnej metódy. V tejto práci je vyhodnotený vplyv MFCC, PLP a LPCC parametrizačných metód [8], [9]. Základné informácie o týchto metódach sú popísané v nasledujúcej časti.

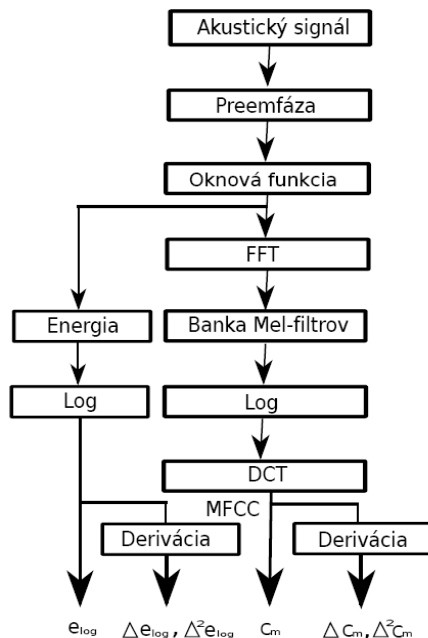
A. Mel-frekvenčná keprálna analýza

Analýza akustického signálu pomocou MFCC keprálnych koeficientov sa snaží zohľadniť poznatky z oblasti percepcie zvuku. MFCC keprálne koeficienty rešpektujú nelineárnu závislosť vnímania frekvencie. Tento aspekt je zohľadnený použitím trojuholníkových

pásmových priepustov s nelineárnym frekvenčným rozložením, tzv. Melovskou frekvenčnou škálou. Vzťah medzi frekvenciou a mel-frekvenciou je daný vzťahom:

$$f_{mel} = 2595 \cdot \log_{10}(1 + 700f), \quad (1)$$

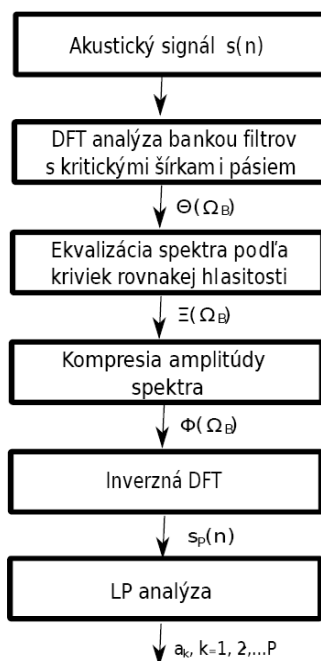
kde f je frekvencia v Hz a f_{mel} je jej odpovedajúca frekvencia na nelineárnej frekvenčnej škále v meloch [mel]. Proces výpočtu Mel-frekvenčných kepstrálnych koeficientov je zobrazený na Obr. 1.



Obr. 1 Proces výpočtu MFCC koeficientov

B. Percepčná lineárna predikcia

Percepčná lineárna predikcia využíva najmä poznatky o frekvenčnej závislosti vzhľadom k vnímaniu výšky tónu, hlasitosti, a maskovacím javom, teda sa svojím charakterom snaží priblížiť k vnímaniu zvuku sluchovými orgánmi človeka. Využíva Barkovu nelineárnu frekvenčnú škálu. Tieto vlastnosti radia PLP k populárnym a často používaným parametrizáciám. Proces výpočtu PLP koeficientov je znázornený na Obr. 2.



Obr. 2 Proces výpočtu PLP koeficientov

C. Lineárna predikčná kepstrálna analýza

Lineárne predikčné kepstrálne koeficienty sú odvodené zo spektrálnej obálky získanej pomocou lineárnej predikčnej analýzy (LPC). Výhoda LPCC koeficientov oproti tradičným LPC koeficientom spočíva v ich dekorelovanosti. Výpočet LPCC je odvodený z výpočtu LPC koeficientov pomocou logaritmovania prenosovej funkcie vokálneho traktu $H(z)=G/A(z)$:

$$\log(G / A(z)) = c(0) + c(1)z^{-1} + \dots = \sum_{k=0}^{\infty} c(k)z^{-k}, \quad (2)$$

kde $c(k)$ sú kepstrálne koeficienty lineárnej predikcie - LPCC.

III. KLASIFIKÁTOR - GMM

Jadrom rozpoznávacieho systému je klasifikátor, ktorý na základe vedomosti získanej z procesu tréningu dokáže účinne rozklasifikovať (rozdeliť) vstupné parametre do vopred definovaných tried. V tejto práci v úlohe klasifikátora vystupujú Gaussove modely zmesí (GMM) [12], [13]. GMM patrí ku klasifikátorom využívajúcim štatistickú informáciu. Hlavným parametrom GMM modelu je počet zmesí gaussovských rozdelení pričom jedna gaussovská funkcia predstavuje jeden zhuk v priestore. Zmiešaním M gaussovských pravdepodobnostných funkcií vzniká funkcia hustoty pravdepodobnosti tzv. PDF - Probability Density Function. GMM je určený vektorom stredných hodnôt, kovariančnou maticou a váhami jednotlivých zložiek. Parametre GMM sú v procese tréningu odhadované pomocou EM algoritmu. Cieľom modelovania je nájsť také parametre modelu, aby bola maximalizovaná pravdepodobnosť $P(X|C_i)$:

$$P_{GMM}(x | C_i) = \sum_{m=1}^M c_{im} N(x, \mu_{im}, \Sigma_{im}), \quad (3)$$

kde x predstavuje vektor prízankov a C_i je akustický model,

$$N(x, \mu_{im}, \Sigma_{im}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \times \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right], \quad (4)$$

kde c_{im} predstavuje váhu, μ_{im} je stredná hodnota a Σ_{im} je kovariančná matica m -tej zmesi pre triedu i .

V tejto práci GMM modely s rôznym počtom zmesí (do 256 PDFs) boli natréňované a ich výkonnosť bola porovnaná s použitím rôznych parametrizačných techník.

IV. DATABÁZA ZVUKOV

Akustické udalosti použité v týchto experimentoch boli nahrávané v relatívne tichom prostredí. Následne bola vytvorená tréningová a testovacia sada zvukov. Pozadie použité v týchto experimentoch bolo nahrávané v blízkosti rušnej ulice. Pri nahrávaní bola použitá vzorkovacia frekvencia 48kHz s rozlíšením 16 bitov na vzorku. Všetky dáta boli manuálne označené s použitím Transcriber-u.

GMM modely boli natréňované na 40 výstreloch zo zbrane, 40 zvukov rozbitia skla a 11 minút trvajúcom pozadí. Testovacia sada bola použitá na ohodnotenie výkonnosti detekčného systému, ktorého úlohou bolo detekovať dva druhy akustických udalostí t.j. rozbitie skla a výstrel zo zbrane. Testovanie bolo uskutočnené s akustickými dátami v trvaní 40 sekúnd. Tieto nahrávky zahŕňali 7 + 7 akustických udalostí spolu s primiešaným pozadím.

V. POPIS EXPERIMENTOV

Experimenty boli zamerané na detekciu zvukov predstavujúcich akustické udalosti, t.j. zvuk výstrelu zo zbrane a zvuk rozbitia skla. Súbor HTK nástrojov bol použitý na extrakciu prízankov z akustického signálu, tréning a testovanie GMM modelov.

Pri extrakcii prízankov bola použitá Hammingova oknová funkcia s trvaním 20 ms a prekryvaním 10 ms. Niekoľko typov parametrizácií bolo vykonaných, no výsledky len tých, ktoré dosiahli najlepšie výsledky sú prezentované. Jedná sa o parametrizácie MFCC, PLP, LPCC s použitím časových derivácií koeficientov a koeficientom vyjadrujúcim energiu skúmaného zvukového segmentu. Tiež lineárna predikčná analýza (LPC) bola vykonaná a otestovaná. Výsledné hodnoty nedosahovali dobré výsledky v porovnaní s predchádzajúcimi parametrizáciami. Z tohto dôvodu nie sú v článku prezentované.

Príznakový vektor popisujúci zvukový segment mal buď z 36 alebo z 39 koeficientov. Vektor s 36 koeficientmi pozostával zo statických koeficientov a ich prvej (D - delta)

a druhej(A - acceleration) časovej derivácie. Koeficienty vyjadrujúce energiu skúmaného rámca signálu boli vynechané. Pri použití vektora príznakov s rozmerom 39 boli tak, ako v predošlom prípade použité statické a dynamické koeficienty, plus koeficienty popisujúce energetické vlastnosti skúmaného rámca ((0)-tý koeficient resp. (E)-koeficient logaritmu energie).

V procese tréningu akustických modelov boli vytvorené modely do počtu zmesí 256 pre akustické udalosti (výstrel zo zbrane a rozbitie skla) a pre zvuky pozadia.

Na ohodnotenie výkonnosti detekčného systému bol použitý parameter „presnosť“ (accuracy) [%], ktorý je definovaný nasledovne [14]:

$$ACC[\%] = \frac{N - D - S - I}{N} \times 100, \quad (5)$$

kde D (deletion) predstavuje chybu spôsobenú zmažaním, S (substitution) chybu spôsobenú zámenou, I (insertion) chybu spôsobenú vložením a N predstavuje celkový počet zvukov podľa priloženého popisného súboru.

VI. VÝSLEDKY EXPERIMENTOV

Výsledky experimentov zameraných na detekciu akustických udalostí sú zobrazené v Tab. 1 – Tab. 3.

Prvý experiment bol zameraný na detekciu rozbitia skla. Najlepšie výsledky boli dosiahnuté pre parametrizácie PLP_D_A, MFCC_D_A, LPCC_E_D_A a MFCC_E_D_A ako je možné vidieť v Tab. 1. Hrubým písmom sú označené hodnoty presnosti (ACC), ktoré sú nad úrovňou 85%. Z výsledkov tiež vyplýva, že vysoký počet zmesí GMM modelu neprináša zvýšenie výkonnosti detekčného systému.

Tabuľka 1
 VÝSLEDKY DETEKcie ROZBITIA SKLA

ACC [%]	Počet zmesí GMM								
Rozbitie skla	1	2	4	8	16	32	64	128	256
PLP_D_A	26,67	6,67	33,33	26,67	100	86,67	73,33	40	20
PLP_0_D_A	73,33	46,67	40	20	20	6,67	20	6,67	6,67
PLP_E_D_A	26,67	26,67	46,67	46,67	26,67	20	13,33	20	6,67
MFCC_D_A	53,33	60	53,33	100	100	80	80	66,67	60
MFCC_0_D_A	60	60	53,33	6,67	26,67	46,67	33,33	20	6,67
MFCC_E_D_A	60	40	93,33	100	86,67	86,67	20	6,67	6,67
LPCC_E_D_A	33,33	100	86,67	73,33	80	46,67	46,67	33,33	40

V Tab. 2 sú zobrazené dosiahnuté výsledky detekcie výstrelů zo zbrane. V porovnaní s predchádzajúcim experimentom boli dosiahnuté horšie hodnoty parametra ACC. Presnosť rozpoznania testovacej nahrávky bola na úrovni max. 86,67%. Podobne ako v predchádzajúcom prípade, nárast počtu zmesí GMM modelu nepriniesol očakávané zlepšenie.

Tabuľka 2
 VÝSLEDKY DETEKcie VÝSTRELOV ZO ZBRANE

ACC [%]	Počet zmesí GMM								
Výstrely	1	2	4	8	16	32	64	128	256
PLP_D_A	33,33	20	33,33	46,67	46,67	40	33,33	33,33	33,33
PLP_0_D_A	66,67	73,33	86,67	46,67	60	46,67	46,67	46,67	20
PLP_E_D_A	33,33	20	73,33	60	33,33	33,33	33,33	20	20
MFCC_D_A	60	60	73,33	53,33	40	60	60	26,67	6,67
MFCC_0_D_A	73,33	86,67	86,67	26,67	46,67	73,33	26,67	20	6,67
MFCC_E_D_A	46,67	60	80	60	40	40	6,67	6,67	6,67
LPCC_E_D_A	6,67	6,67	86,67	73,33	73,33	46,67	60	33,33	33,33

Priemerné výsledky, ktoré podávajú obraz o celkovej výkonnosti detekčného systému sú zobrazené v Tab. 3. Detekčný systém bol navrhnutý na detekciu dvoch typů akustických udalostí (rozbitia skla a výstrelu zo zbrane). Iba v jednom prípade bola dosiahnutá úroveň presnosti 86,67% pre MFCC_E_D_A s GMM modelom so 4 zmesami.

Tabuľka 3
 VÝSLEDKY DETEKCÉHO SYSTÉMU

ACC [%]	Počet zmesí GMM								
	1	2	4	8	16	32	64	128	256
PLP_D_A	30	13,33	33,33	36,67	73,33	63,33	53,33	36,67	26,67
PLP_O_D_A	70	60	63,33	33,33	40	26,67	33,33	26,67	13,33
PLP_E_D_A	30	23,33	60	53,33	30	26,67	23,33	13,33	13,33
MFCC_D_A	56,67	60	63,33	76,67	70	70	70	46,67	33,33
MFCC_O_D_A	66,67	73,33	70	16,67	36,67	60	30	20	6,67
MFCC_E_D_A	53,33	50	86,67	80	63,33	63,33	13,33	6,67	6,67
LPCC_E_D_A	20	66,67	66,67	63,33	70	46,67	53,33	33,33	36,67

VII. ZÁVER

Výsledky prezentované v predchádzajúcej časti poukazujú na skutočnosť, že metódy extrakcie príznakov, ktoré boli navrhnuté primárne pre rečový signál dokážu účinne reprezentovať akustické udalosti. Testované metódy parametrizácie vhodnejšie reprezentovali zvuk rozbitia skla než zvuk výstrelu zo zbrane. Z výsledkov experimentov vyplýva že vysoký počet zmesí GMM nepriniesol zlepšenie výkonnosti systému. Táto skutočnosť môže byť ovplyvnená nedostatkom tréningových dát.

V blízkej budúcnosti náš záujem sústredíme na aplikáciu MPEG -7 deskriptorov, ktoré by mohli vylepšiť výsledky rozpoznania. Taktiež, v prípravnej fáze je rozšírenie existujúcej databázy o nové nahrávky akustických udalostí a o nahrávky rôznych druhov pozadí.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] S. Ntalampiras, I. Potamitis, N. Fakotakis, "On acoustic surveillance of hazardous situations," ICASSP, Taiwan, 2009, pp. 165-168, ISBN: 978-1-4244-2353-8.
- [2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems," London, 2007, pp. 21-26, ISBN: 978-1-4244-1696-7.
- [3] J. L. Rouas, J. Louradour, S. Ambellouis, "Audio events detection in public transport vehicle," ITSC, Canada, 2006, pp. 733-738, ISBN: 1-4244-0093-7.
- [4] A. Temko, R. Malkin, Ch. Zieger, D. Macho, C. Nadeu, M. Omologo, "Acoustic Event Detection and Classification in Smart-Room Environment: Evaluation of CHIL Project Systems," in The IV Biennial Workshop on Speech Technology, 2006.
- [5] J. Portêlo, M. Bugalho, I. Trancoso J. Neto, A. Abad, A. Serralheiro, "Non-speech audio event detection," ICASSP, Taiwan, 2009, pp. 1973-1976, ISBN: 978-1-4244-2353-8.
- [6] X. Zhuang, X. Zhou, T.S. Huang, M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," ICASSP, Nevada, 2008, pp. 17-20, ISBN: 978-1-4244-1483-3.
- [7] X. Huang, A. Acero, H.-W. Hon, "Spoken language processing," 2001, ISBN 0-13-022616-5.
- [8] J. Psutka, L. Müller, J. V. Psutka, "Comparison of MFCC and PLP parametrizations in the speaker independent continuous speech recognition task," Eurospeech, Scandinavia, 2001, pp.1813-1816.
- [9] G. Antoniol, V. F. Rollo, G. Venturi, "Linear predictive coding and cepstrum coefficients for mining time variant information from software repositories," International workshop MRS, Missouri, 2005.
- [10] M. Casey, "General sound classification and similarity in MPEG-7," Cambridge University Press, 2001, pp. 153-164.
- [11] Ch. Zieger, M. Omologo, "Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm," INTERSPEECH, Australia, 2008.
- [12] M. Nazari, A. Sayadiyan, S. Valiollahzadeh, "Probabilistic SVM/GMM classifier for speaker-independent vowel recognition in continues speech," ICASSP, Taiwan, 2009, ISBN: 978-1-4244-2353-8.
- [13] J. K. Shah, A. N. Iyer, B. Y. Smolenski, R. E. Yantorno, "Robust voiced/unvoiced classification using novell features and Gaussian mixture model," ICASSP, Canada, 2004, ISBN: 0-7803-8484-9.
- [14] S. Young, et. al., The HTK Book. Cambridge University 2009.

Klasifikácia akustických udalostí na báze SVM

Jozef VAVREK, Eva VOZÁRIKOVÁ, Peter VISZLAY, Anton ČIŽMÁR

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

jozef.vavrek@tuke.sk, eva.vozarikova@tuke.sk, peter.viszlay@tuke.sk

Abstrakt — Táto práca sa zaoberá základnou problematikou klasifikácie a detekcie dvoch typov akustických udalostí a to výstrelů z rôznych typů zbraní a rozbitia skla. K základným komponentom takéhoto systému patrí extrakcia príznakov, alebo tiež parametrizácia spracovávaného signálu a následná klasifikácia do jednotlivých tried. Parametrizácia bola realizovaná pomocou troch typů metód, menovite Mel-frekvenčných keprálnych koeficientů (MFCC), koeficientů lineárnej predikcie (PLP) a analýzy hlavných koeficientů (PCA). Klasifikácia jednotlivých udalostí bola uskutočnená pomocou SVM klasifikátora, ktorého koncepcia vychádza z princípů štatistickej teórie učenia a nájdenia optimálnej deliacej hyperplochy. Hlavné úsilie je smerované k optimalizácii parametrov SVM klasifikátora za účelom dosiahnutia čo najvyššej presnosti klasifikácie.

Kľúčové slová — akustická udalosť, parametrizácia, SVM

I. ÚVOD

Účelom systémů rozpoznávania zvuků (udalostí) je vo všeobecnosti určiť, či daný zvuk patrí do príslušnej triedy, alebo oddeliť jednu triedu zvuků od ostatných. Takéto systémy umožňujú detekovať a klasifikovať rôzne typy akustických udalostí, reči a tiež zmeny rečníkov. Pozostávajú z dvoch základných modulů, konkrétne *modul detekcie udalostí*, ktorý izoluje jednotlivé segmenty signálů od pozadia, v závislosti od spektrálnych a časových zmien v audio toku. Následne *modul extrakcie príznakov a klasifikácie* priradí akustický signál do príslušnej kategórie, na základe popisných parametrov signálu a optimalizácie parametrov modelu klasifikátora. K tomu aby klasifikátor vedel s čo najvyššou presnosťou určiť do ktorej triedy príslušný zvukový segment patrí, je potrebné vykonať čo najvernejšiu parametrizáciu signálu, teda extrahovať súbor parametrov zachytávajúcich spektrálne a časové charakteristiky takéhoto signálu. Z toho dôvodu boli na jednotlivé akustické signály aplikované tri často používané parametrizačné techniky a to MFCC, PLP a PCA.

Metóda podporných vektorů (Support Vector Machines - SVM) je jednou z najpoužívanějších metód určených na klasifikáciu rôznych typů akustických udalostí. Táto práca sa preto bližšie zameriava na popis charakteristik a základných princípů tejto metódy, ako aj jej aplikáciu na klasifikáciu akustických udalostí pre bezpečnostné systémy.

Akustická udalosť predstavuje krátky zvukový segment, ktorý ma zvyčajne náhodný charakter, jej priebeh nie je možné predpovedať a jej charakteristika závisí od danej aplikácie použitia. Veľkým úsilím v bezpečnostných systémoch je preto detekcia akustických udalostí, ktoré predstavujú hrozby, teda situácie ohrozujúce ľudský život, ako napr. výstrely zo zbrane, volanie o pomoc, výkriky, rozbitie skla, explózie, atď.

Štruktúra práce je rozdelená do nasledujúcich sekcií: sekcia 2 opisuje súvisiace práce, sekcia 3 poskytuje prehľad použitých parametrizačných metód a SVM klasifikátora a v sekcií 4 sú zahrnuté realizované experimenty a ich výsledky. V sekcií 5 sú zhrnuté prínosy práce a ciele do budúcnosti.

II. SÚVISIACE PRÁCE

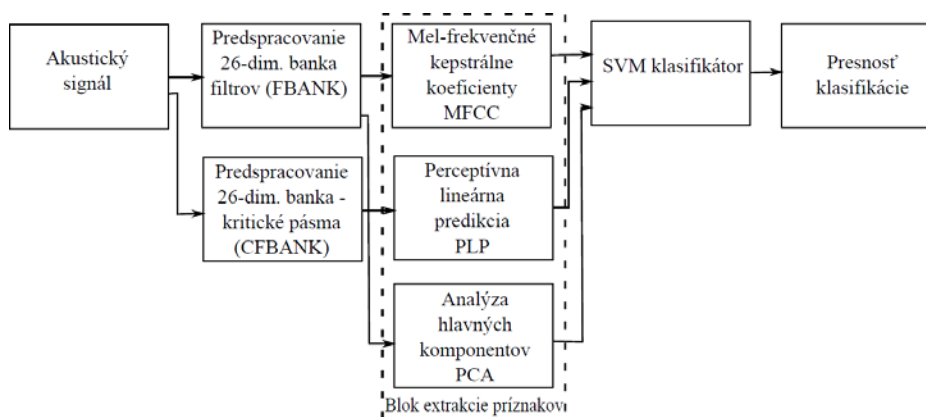
Niektoré práce [1, 2] sa zameriavajú na návrh všeobecného systému klasifikácie zvuků, pričom sú použité rôzne typy parametrov, ktoré sú extrahované z akustických signálů, napr. MPEG-7 nízko-úrovňové deskriptory (plochosť spektra, spektrálna entropia, spektrálne ťažisko, obálka spektra, a pod.), počet prechodů nulou (zero crossing rate - ZCR), základná

frekvencia (fundamental frequency – FF), atď. Tieto parametre procesu analýzy akustických signálov zabezpečujú veľkú výkonnosť systémov rozpoznávania hovorenej reči a tiež rôznych typov akustických udalostí [3].

Niekoľko ďalších prác je zameraných na detekciu a klasifikáciu niekoľkých typov akustických udalostí ako sú: zvuky vtákov [4], zvuk gitary [5], potlesk [6], výstrely zo zbrane [7], a iné [8], pomocou klasifikátora SVM. V iných prácach je hlavným cieľom nájdenie optimálnych parametrov klasifikátorov ako sú GMM, HMM [9], KNN [10], umelé neurónové siete (ANN) [11] a ich kombinácie. Teda pri navrhovaní takýchto klasifikačných systémov je hlavné úsilie kladené na vytvorenie takého optimálneho detektora a klasifikátora, pomocou ktorého by bolo možné klasifikovať akustické udalosti s čo najvyššou presnosťou a dôveryhodnosťou.

III. METÓDY EXTRAKCIE PRÍZNAKOV A POPIS KLASIFIKÁTORA

Ako bolo spomínané v predošlej kapitole, každý systém rozpoznávania pozostáva z dvoch základných blokov, bloku extrakcie príznakov a bloku klasifikátora. Úlohou bloku extrakcie príznakov je spracovať vstupný akustický signál do formy vhodnej na ďalšie spracovanie v klasifikačnom systéme. Všeobecná blokovaná schéma klasifikačného systému s tromi základnými typmi transformačných technik a SVM klasifikátora je zobrazená na nasledujúcom obrázku (Obr. 1).



Obr. 1 Blokovaná schéma navrhnutého systému detekcie a klasifikácie akustických udalostí

A. Analýza hlavných komponentov

Analýza hlavných komponentov (Principal Component Analysis - PCA) [12] predstavuje optimálnu lineárnu ortogonálnu transformáciu, ktorá mapuje n -dimenzionálny vstupný vektor dát do m -dimenzionálneho, pričom platí, že $m < n$. Úlohou PCA je nájdenie takého vektora, ktorý najlepšie popisuje charakteristiky vstupných dát v zmysle najväčšieho rozptylu. K popisu veľkosti rozptylu slúžia tzv. *hlavné komponenty*. Tieto základné komponenty môžeme určiť pomocou výpočtu kovariančnej matice spracovávaných dát, pričom hľadáme ich vlastné vektory s najväčšou vlastnou hodnotou. K vyriešeniu problému vzťahujúcemu sa na zvolený počet hlavných komponentov je použité porovnávacie kritérium s optimálnou prahovou hodnotou. Zvolené optimálne dimenzie pre jednotlivé akustické udalosti sú uvedené v sekcii IV.

B. Perceptívna lineárna predikcia

Perceptívna lineárna predikcia (Perceptual Linear Prediction - PLP) [13] zabezpečuje extrakciu užitočnej informácie z rečového signálu na báze ľudského percepčného systému. Na rozdiel od mel-keprálnej analýzy, PLP používa odlišný typ banky filtrov a výsledné koeficienty sú počítané pomocou lineárnej predikcie. Banka filtrov je určená z barkovej mierky, pričom prevod z frekvenčnej do barkovej mierky je definovaný nasledujúcim vzťahom:

$$\Omega = \text{Bark}(f) = 6 \ln \left(\frac{f}{600} + \sqrt{\left(\frac{f}{600} \right)^2 + 1} \right) \quad (1)$$

PLP koeficienty sú potom vypočítané použitím autokorelačných koeficientov, ktoré sú definované pomocou inverznej Fourierovej transformácie (IDFT) a z lineárnych percepčných koeficientov, získaných lineárnou predikciou.

C. Mel-frekvenčné kepstrálne koeficienty

Mel-frekvenčná kepstrálna analýza (Mel-Frequency Cepstral - MFC) [13] sa stala populárnou metódou reprezentácie signálu v mnohých úlohách klasifikácie. Analýza akustického signálu pomocou MFCC kepstrálnych koeficientov sa snaží zohľadniť aj poznatky z oblasti percepcie zvuku. MFCC kepstrálne koeficienty rešpektujú nelineárnu závislosť vnímania frekvencie. Tento aspekt je zohľadnený použitím trojuholníkových pásmových priepustov s nelineárnym frekvenčným rozložením tzv. Melovskou frekvenčnou škálou. Vzťah medzi frekvenciou a mel-frekvenciou je definovaný nasledovne:

$$f_{mel} = 2595 \log_{10}(1 + 700f), \quad (2)$$

kde f je frekvencia v lineárnej škále [Hz] a f_{mel} je jej odpovedajúca frekvencia na nelineárnej frekvenčnej škále v Meloch [Mel]. Výsledné koeficienty sú vypočítané pomocou Fourierovej transformácie a diskkrétnej kosínusovej transformácie (Discrete Cosine Transformation - DCT).

D. SVM klasifikátor

Metóda podporných vektorov (Support Vector Machines - SVM) [15] je jednou z najpoužívanějších techník v mnohých odvetviach strojového učenia. Algoritmus SVM je založený na klasifikácii vzoriek vo feature priestore (vektorovom priestore vzoriek) pomocou *oddelovacej hyperplochy*, ktorá plní úlohu separácie vzoriek podľa tried, do ktorých jednotlivé vzorky patria. V prípade ak vzorky patriace do odlišných tried nie je možné oddeliť v priestore lineárne, použijeme mapovanie do viacrozmerného vektorového priestoru vzoriek a následne oddelenie pomocou lineárnej hyperplochy. Túto transformáciu realizujeme použitím *kernelových funkcií*: polynomiálna, sigmodálna, viacvrstvový perceptrón, RBF funkcia majúca ako základ radiálnu symetrickú Gaussovskú funkciu. Pri hľadaní optimálnej oddeľovacej hyperplochy sa potom zohľadňuje šírka hraničnej oblasti, ktorá určuje celkovú chybovosť klasifikácie a počet vzoriek, ktoré definujú deliace hranice, teda *podporných vektorov*. Podporné vektory sú základnými prvkami pri určovaní hraníc rozpätia hyperplochy a počítajú sa pomocou Lagrangeových multiplikátorov pri riešení štandardného optimalizačného problému.

Majme súbor tréningových dát, ktoré sú reprezentované vo forme N-dimenzionálnych vektorov:

$$(x_1, y_1), \dots, (x_n, y_n) \in X \times \{\pm 1\}, \quad (3)$$

kde X je neprázdna množina n -rozmerných vektorov x_i (vzoriek, koeficientov), ktoré reprezentujú danú triedu a y_i predstavujú označenie jednotlivých tried. V prípade binárnej klasifikácie je potrebné oddeliť iba dve triedy vzoriek označených $+1$ pre vzorky z prvej triedy a -1 pre vzorky patriace do druhej triedy. Teda pre všetky vzorky $x_i \in X$ definujeme príslušnú triedu s označením $y \in \{\pm 1\}$. Potom spomedzi všetkých možných deliacich hyperplôch musíme vybrať tú, ktorá minimalizuje chybu pri tréningu a súčasne definuje optimálne deliace hranice, teda *optimálnu deliacu hyperplochu*, ktorú možno zapísať v tvare:

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^n w_i x_i + b, \quad (4)$$

kde $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, a b nazývame *odchýlkou* (bias). Výstupom tréningovej fázy je získanie optimálnych váh $\mathbf{w}$, pomocou ktorých potom definujeme výstup o podľa *indikačnej funkcie* danej vzťahom:

$$if = o = \text{sign}(d(\mathbf{x}, \mathbf{w}, b)). \quad (5)$$

Teda rozhodovacie kritérium môžeme definovať nasledovným spôsobom:

- ak $d(\mathbf{x}_p, \mathbf{w}, b) > 0$, $\mathbf{x}_p$ patrí do triedy 1 ($o = y_1 = +1$),
- a ak $d(\mathbf{x}_p, \mathbf{w}, b) < 0$, potom $\mathbf{x}_p$ patrí do príslušnej triedy 2 ($o = y_2 = -1$).

Parametre optimálnej hyperplochy $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_m]^T$, teda *váhy*, definujeme pomocou Lagrangeovej funkcie:

$$L_d(\mathbf{w}, b, \alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle, \quad (6)$$

ktorej riešením je súbor Lagrangeových koeficientov α_i . Optimálnu hyperplochu potom určíme maximalizáciou (6), za predpokladu splnenia nasledujúcich podmienok:

$$\sum_{i=1}^l \alpha_i y_i = 0; \quad \alpha_i \geq 0. \quad (7)$$

V prípade potreby klasifikácie vzoriek z viacerých tried je nutné použiť klasifikátor viacerých tried (multi-class classifier). SVM používa dva základné typy takýchto klasifikátorov a to *jeden-voči-jednému* (one-against-one) a *jeden-voči-ostatným* (one-against-all). Pri metóde jeden-voči-jednému realizujeme klasifikáciu vždy pre dve triedy. Pre M tried potrebujeme $(M-1)M/2$ binárnych klasifikátorov. Ak klasifikujeme testovacie vzorky, vyhodnocujeme všetky binárne klasifikátory a klasifikujeme ich podľa toho, na ktorej triede dostanú najvyšší počet hlasov (vzoriek priradených do príslušnej triedy). Na druhej strane je metóda jeden-voči-všetkým založená na vytvorení súboru binárnych klasifikátorov, pričom každý je trénovaný spôsobom klasifikácie vzoriek patriacich do jedinej triedy a zvyšných vzoriek patriacich do ostatných tried.

IV. EXPERIMENTY A VÝSLEDKY

A. Predspracovanie použitím banky filtrov

Systém klasifikácie je navrhnutý pre klasifikáciu dvoch typov akustických udalostí (výstrely a rozbitie skla). Na tréovanie a testovanie boli použité nahrávky so vzorkovacou frekvenciou $f_{vz}=48kHz$ a počtom bitov na vzorku $N_{bit}=16bit/vzorku$, v jednokanálovom mono formáte. Počet nahrávok použitých na tréovanie a testovanie je uvedený v Tab. 1.

Tabuľka 1
 SÚBOR POUŽITÝCH AKUSTICKÝCH UDALOSTÍ

Akustická udalosť	Trénovacie dáta	Testovacie dáta
Sklo	28	5
Výstrely	31	10
Pozadie	128	31

Jednotlivé príznaky (koeficienty) boli počítané pomocou segmentácie na rámce o veľkosti 20ms s prekryvaním 10ms, na ktoré bola aplikovaná Hammingova oknová funkcia. V prvom kroku bol vstupný signál filtrovaný pomocou banky 26 filtrov - 26 FBANK a následne boli vypočítané koeficienty pre jednotlivé parametrizačné metódy. Súbor FBANK koeficientov bol vypočítaný v bloku predspracovania (Obr. 1) pomocou nástroja HTK – Hidden Markov Model Toolkit. Optimálna dimenzia (počet prvkov vektora) každej vzorky bola nezávisle vypočítaná pomocou PCA metódy, teda pre triedu *sklo* bola optimálna dimenzia 3, pre *výstrely* bola 4 a pre *pozadie* 7. V prípade mel-frekvenčnej kepstrálnej analýzy bolo použitých 12 kepstrálnych koeficientov a nultý koeficient popisujúci logaritmus energie. Výstupom perceptívnej lineárnej predikcie bol súbor 12 perceptívnych koeficientov, podobne ako v prípade MFCC, avšak okrem nultého koeficientu.

B. Metodika tréovania a testovania

Po rozdelení spracovávaných dát na trénovacie a testovacie podľa Tab. 1, bolo vykonané predspracovanie jednotlivých signálov pomocou nástroja HTK, ktorého výstupom boli FBANK koeficienty. Následne bola vykonaná parametrizácia pomocou MFC, PLP a PCA techník. Každá spracovaná nahrávka bola reprezentovaná maticou koeficientov, pričom výsledný trénovací a testovací set bol vytvorený zlúčením týchto matíc. Výsledná transformačná matica MFC, PLP a PCA metód bola potom prispôbená do formy vhodnej pre ďalšie spracovanie SVM klasifikátora.

Trénovacie modely pre každý typ udalostí a pozadia boli vytvorené pomocou softvéru LIBSVM (Library for Support Vector Machines). LIBSVM je nástroj, využívajúci prostredie MATLAB, určený pre klasifikáciu a regresiu dát pomocou SVM. Tento nástroj podporuje viacero funkcií ako sú: C-SVM, v-SVM, regresiu, klasifikáciu viacerých tried a pod. Na vytvorenie trénovacieho modelu bola použitá C-SVM klasifikácia, z dôvodu dosahovania najvyššej presnosti klasifikácie, pričom bolo potrebné definovať optimálny penalizačný

parameter C (penalty parameter). Ten bol vypočítaný pomocou 5-vrstvovej *kros-validačnej techniky* (cross-validation). Jednotlivé experimenty boli realizované pomocou RBF kernelovej funkcie, kde bolo potrebné definovať parameter γ , ktorý bol taktiež získaný použitím kros-validácie. Vo väčšine prípadov boli z dôvodu vytvorenia optimálneho tréningového modelu zvolené parametre $C = 8$ a $\gamma = 2$. Všetko o týchto funkciách a parametroch softvéru LIBSVM môžete nájsť v [16].

V prípade metódy PCA boli jednotlivé akustické udalosti reprezentované pomocou vektorov príznakov s optimálnymi dimenziami (Dim3, Dim4, Dim7). Klasifikácia súboru tréningových vzoriek bola realizovaná pomocou metódy jeden-voči-jednému s použitím troch binárnych SVM klasifikátorov. Dosiahnuté výsledky jednotlivých experimentov pri použití metódy MFC sú zobrazené v Tab. 2, pri použití metódy PLP v Tab. 3 a pre PCA metódu v Tab. 4. Výsledná presnosť klasifikácie jednotlivých tried udalostí pre PCA metódu je uvedená v Tab. 5.

Na vyhodnotenie navrhnutého klasifikačného systému bol použitý parameter presnosti (úspešnosti) klasifikácie [%], definovaný nasledovne:

$$Presnost' = \frac{pocet_spravne_klasifikovanych_vzoriek}{pocet_vsetkych_vzoriek} * 100\%$$

Tabuľka 2
PRESNOSŤ SVM KLASIFIKÁCIE PRI POUŽITÍ METÓDY MFC

Testovacie vzorky	Tréningový model	Presnosť [%]	Počet správne klasifikovaných testovacích vzoriek
Pozadie_sklo_výstrely	model_rbf_C8_g2	96.16	12655/13202
Výstrely	model_rbf_C8_g2	41.44	281/678
Rozbitie skla	model_rbf_C8_g2	23.25	93/400

Tabuľka 3
PRESNOSŤ SVM KLASIFIKÁCIE PRI POUŽITÍ METÓDY PLP

Testovacie vzorky	Tréningový model	Presnosť [%]	Počet správne klasifikovaných testovacích vzoriek
Pozadie_sklo_výstrely	model_rbf_C8_g2	96.97	12803/13202
Výstrely	model_rbf_C8_g2	7.27	50/678
Rozbitie skla	model_rbf_C8_g2	8.41	34/400

Tabuľka 4
PRESNOSŤ SVM KLASIFIKÁCIE PRI POUŽITÍ METÓDY PCA

Testovacie vzorky	Tréningový model	Presnosť [%]	Počet správne klasifikovaných testovacích vzoriek
Pozadie_sklo_výstrely_3dim	model_rbf_C16_g4	93.24	12310/13202
Pozadie_sklo_výstrely_4dim	model_rbf_C16_g4	94.56	12485/13202
Pozadie_sklo_výstrely_7dim	model_rbf_C64_g4	96.82	12780/13202

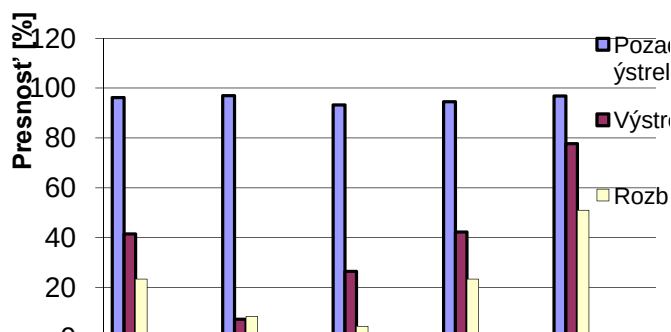
Tabuľka 5
VÝSLEDNÁ PRESNOSŤ SVM KLASIFIKÁCIE JEDNOTLIVÝCH TRIED PRI POUŽITÍ METÓDY PCA

Presnosť klasifikácie testovacích vzoriek [%]	Dim 3	Dim 4	Dim 7
Rozbitie skla	4.45	23.26	50.99
Výstrely	26.34	42.21	77.63
Pozadie	100	99.91	99.43

Z výsledkov uvedených v jednotlivých tabuľkách vyplýva, že najoptimálnejšou metódou extrakcie príznakov pre klasifikáciu vzoriek pomocou SVM je PCA technika. Z Tab. 5 možno vidieť, že znižovanie dimenzie vektorov príznakov spôsobuje výrazné zníženie presnosti klasifikácie pre jednotlivé akustické udalosti ako aj celkovú presnosť predikcie jednotlivých testovacích vzoriek. Porovnanie výsledkov klasifikácie pre použité metódy je uvedené na Obr. 2.

V. ZHRNUTIE

Cieľom práce a jednotlivých experimentov bolo navrhnúť efektívny klasifikačný systém, ktorý by bol schopný klasifikovať akustické udalosti s pozadím, s čo najvyššou presnosťou a spoľahlivosťou. Klasifikácia vybraných typov akustických signálov bola realizovaná pomocou klasifikátora SVM za použitia troch parametrizačných metód MFC, PLP a PCA. Pomocou týchto metód extrakcie príznakov a klasifikátora sme určili typ najvhodnejších koeficientov na klasifikáciu tréningových a testovacích vzoriek, s cieľom dosiahnutia najvyššej presnosti klasifikácie jednotlivých akustických udalostí. Výsledky experimentov ukazujú, že metóda SVM je vhodná na klasifikáciu dvoch typov akustických udalostí s primiešaným pozadím, s použitím techniky PCA. Maximálna presnosť klasifikácie pre jednotlivé triedy zvukov bola dosiahnutá použitím tejto techniky pre 7 dim. vektor príznakov. Veľké množstvo tréningových a testovacích dát s pozadím spôsobilo nevyvážené výsledky pri vyhodnocovaní presnosti klasifikácie testovacích vzoriek realizovaných experimentov. Ciele budúcej práce sa sústreďujú na použitie príznakov štandardu MPEG-7 a kombinácie iných klasifikačných metód s klasifikátorom SVM.



Obr. 2 Porovnanie výsledkov SVM klasifikácie použitých parametrizačných metód

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] M. Casey, *General Sound Classification and Similarity in MPEG-7*, vedecký článok, Organised Sound, Volume 6, pp. 153-154.
- [2] G. Muhammad, K. Alghathbar, *Environment recognition from audio using MPEG-7 features*, vedecký článok, Proc. EN-Com 2009, ISBN: 978-1-4244-4995-8.
- [3] X. Zhuang, X. Zhou, T.S. Hhuang, M. Hasegawa-Johnson, *Feature analysis and selection for acoustic event detection*, ICASSP 2008, pp. 17-20, ISBN:1-4244-1484-9.
- [4] S. Fagerlund, *Bird Species Recognition Using Support Vector Machines*, EURASIP Journal on Advances in Signal Processing, 2007, pp. 3-5.
- [5] K. Dosenbach, W. Fohl, A. Meisel, *Identification of Individual Guitar Sounds by Support Vector Machines*, Proc. of the Conference on Digital Audio Effects, Finland, 2008.
- [6] L. Li, G. Fengpei, Z. Qingwei, Y. Yonghong, *A SVM-Based Audio Event Detection System*, icece, pp.292-295, 2010 International Conference on Electrical and Control Engineering, 2010.
- [7] J. Martin, *Support Vector Machine Classification of Gunshots*, 2007, dostupné na internete: <<http://www.cs.duke.edu/~jmartin/files/machineLearning-paper.pdf>>.
- [8] A. Temko, C. Nadeu, J. Biel, *Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07*, Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, 2008, Volume 4625/2008.
- [9] X. Zhou, et.al., *HMM-based Acoustic Event Detection with Adaboost Feature Selection*, Multimodal Technologies for Perception of Humans, 2008.
- [10] W. Jia-Ching, et.al., *Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor*, International Joint Conference on Neural Networks 2006.
- [11] C. Lopes, F. Perdigao, *Event Detection by HMM, SVM and ANN: A Comparative Study*, Computational Processing of the Portuguese Language 2008.
- [12] L. I. Smith, *A Tutorial on Principal Components Analysis*, 2002, dostupné na internete <http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf>.
- [13] X. Huang, A. Acero, H. Hsiao-Wuen, *Spoken Language Processing: A Guide to Theory, algorithm, and system development*, Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2001, ISBN 0-13-022616-5.
- [14] K. PALIWAL, *Decorrelated and liftered filter-bank energies for robust speech recognition*, Eurospeech'99, pp. 85-88, 1999.
- [15] V. KECMAN, *Learning and Soft Computing, Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. The MIT Press, Massachusetts Institute of Technology, 2001. ISBN 0-262-11255-8, pp.121-189.
- [16] Ch. Chin-Chung, Ch. Chin-Jen, *Library for Support Vector machines*, National Taiwan University, February 27, 2009, dostupné na internete: <<http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>>.

Porovnanie dekorelačných metód príznakových vektorov v akustickom modelovaní založenom na HMM

Peter VISZLAY, Matúš PLEVA, Jozef JUHÁR

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky a
informatiky, Technická univerzita v Košiciach, Slovenská republika

peter.viszlay@tuke.sk, matus.pleva@tuke.sk, jozef.juhar@tuke.sk

Abstrakt – Tento článok porovnáva dva akustické modely založené na štruktúre skrytých Markovových modelov (HMM). Prvý model je konvenčný akustický model založený na príznakoch typu MFCC (Mel-frequency cepstral coefficients). Druhý model je založený na lineárnej transformácii PCA (Analýza hlavných komponentov, v angl. Principal Component Analysis). Cieľom príspevku je porovnať výsledky rozpoznávania pre tieto dva podobné akustické modely. Hlavný rozdiel medzi nimi je v aplikácii dekorelačnej techniky v procese extrakcie príznakov. PCA je známa metóda spracovania dát používaná v systémoch rozpoznávania reči, ktorej cieľom je dekorelovať príznakové vektory a redukovať ich rozmer. Pre experimentálne účely bola použitá rečová databáza TIMIT s mapovaním anglických foném.

Kľúčové slová – dekorelácia, fonéma, kovariancia, lineárna transformácia, rozmer, vlastný vektor

I. ÚVOD

Systém automatického rozpoznávania reči (ARR) pracujúci na štatistickom prístupe obvykle pozostáva z bloku predspracovania (akustického front-end-u) a štatistického klasifikátora. Vo fáze predspracovania dochádza k extrakcii vhodných príznakových vektorov, ktoré by mali obsahovať len relevantnú informáciu umožňujúcu jednoznačnú klasifikáciu do jednotlivých fonetických tried s ohľadom na rozmer vektorov. Článok sústreďí pozornosť na predspracovanie rečového signálu s využitím lineárnej transformácie.

Obsah článku je členený nasledovným spôsobom. Ďalšia kapitola opisuje lineárne transformácie používané k redukcii rozmeru vektorov príznakov. Tretia kapitola poskytuje detailný opis metódy - analýzy hlavných komponentov. Štvrtá kapitola prezentuje podmienky experimentov pre oba modely. Piata kapitola uvádza výsledky experimentov a ich porovnanie. Posledná časť poskytuje krátke zhrnutie našej práce a záver.

II. LINEÁRNE TRANSFORMÁCIE PRÍZNAKOVÝCH VEKTOROV

Lineárne transformácie (LT) sú v systémoch ARR používané na konverziu originálnej množiny príznakových vektorov do alternatívneho, kompaktnejšieho príznakového priestoru s ohľadom na zachovanie čo najväčšieho množstva informácie. LT taktiež pomáhajú zvýšiť robustnosť a úspešnosť rozpoznávacieho systému [1], [2].

Počas realizácie samotnej lineárnej transformácie sa často vykonáva aj krok redukcie rozmeru príznakov. Dosiahne sa to zachovaním len relevantných rozmerov po transformácii podľa určitého kritéria. Krok redukcie rozmeru pomáha riešiť problém nazývaný *nevýhoda viacrozmernosti*. Tento jav spôsobuje exponenciálnu degradáciu výkonu rozpoznávacieho systému v závislosti od nárastu rozmeru príznakového priestoru. Príznakové vektory získané ako výsledok procesu parametrizácie rečového signálu sú často korelované, preto je dôležité ich v nasledujúcom kroku dekorelovať.

Lineárna transformácia je aplikovaná na sekvenciu akustických vektorov ako matematická procedúra. Transformovaná množina vektorov je potom vhodná na trénovanie akustického modelu. Všeobecný matematický model LT môže byť formulovaný v nasledovnom tvare [3]:

$$y = \mathbf{W}^T x, \quad (1)$$

kde y predstavuje výstupnú transformovanú množinu, $\mathbf{W}$ je transformačná matica a x je vstupný súbor príznakov. Cieľom LT je nájsť transformačnú maticu s ohľadom na niektoré optimalizačné kritérium (strata informácie, diskriminácia tried a pod.).

V praxi môžu byť LT realizované *pod dohľadom* a *bez dohľadu*. Hlavný rozdiel medzi týmito možnosťami je nevyhnutná potreba prídavnej informácie o dátach (triedach) v prípade metód pracujúcich pod dohľadom. Najznámejšou LT pracujúcou pod dohľadom je *Lineárna diskriminačná analýza* a jedna z najpoužívanejších metód pracujúcich bez dohľadu je *Analýza hlavných komponentov*. Posledne menovaná metóda je tiež použitá v našich experimentoch.

III. ANALÝZA HLAVNÝCH KOMPONENTOV

A. Charakteristika

Analýza hlavných komponentov je známa technika spracovania a redukcie rozmeru dát, ktorá sa s úspechom aplikuje aj v spracovaní reči. Je to optimálna lineárna ortogonálna transformácia, ktorá mapuje vstupné n -rozmerné dáta na m -rozmerné, $m < n$, s ohľadom na ich rozptyl.

Hlavnou myšlienkou PCA je redukovať rozmer množiny dát pozostávajúcej z veľkého množstva vzájomne súvisiacich premenných. Tento zámer je dosiahnutý transformáciou na novú množinu premenných, nazývaných *hlavné komponenty* (HK), ktoré sú nekorelované a vhodným spôsobom zoradené, tzn., že niekoľko prvých HK zachováva najväčšiu časť rozptylu prítomného v originálnych premenných [4]. Obyčajne, prvých niekoľko HK reprezentuje približne 80% rozptylu.

HK môžu byť determinované výpočtom kovariančnej matice centrovaných vstupných dát s jej nasledovným spektrálnym rozkladom, výsledkom ktorého je nájdenie jej vlastných čísel a vlastných vektorov [2]. HK sú vlastné vektory kovariančnej matice reprezentované súradnicovými osami nového, transformovaného súradnicového systému.

B. Matematický opis

Predpokladajme, že $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M$ sú stĺpcové vektory s rozmerom $N \times 1$. Potom, podľa [5] a [6] môže byť PCA vykonaná v nasledovných krokoch.

- 1) Najprv je potrebné vypočítať vektor strednej hodnoty $\bar{\mathbf{x}}$ ako priemer elementov:

$$\bar{\mathbf{x}} = \frac{1}{M} \sum_{i=1}^M \mathbf{x}_i. \quad (2)$$

- 2) Pre správnu funkciu PCA, vypočítaný vektor strednej hodnoty musí byť odčítaný od každého vstupného vektora. Nové, *centrované* vektory majú potom tvar:

$$\Phi_i = \mathbf{x}_i - \bar{\mathbf{x}}. \quad (3)$$

Tento krok produkuje dátový súbor, ktorého globálna stredná hodnota je rovná nule.

- 3) V ďalšom kroku sa z centrovaných vektorov vytvorí centrovaná matica A s rozmerom $N \times M$:

$$A = [\Phi_1 \Phi_2 \dots \Phi_M]. \quad (4)$$

- 4) Z matice A sa vypočíta kovariančná matica C nasledovným spôsobom:

$$C = \frac{1}{M} \sum_{n=1}^M \Phi_n \Phi_n^T = \frac{1}{M} \sum_{n=1}^M (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{M} A A^T. \quad (5)$$

Kovariančná matica je symetrická štvorcová matica s rozmerom $N \times N$ s reálnymi prvkami.

- 5) Spektrálnym rozkladom kovariančnej matice sa získajú jej vlastné čísla $\lambda_i, i \in \langle 1; N \rangle$. Je potrebné ich zostupne zoradiť:

$$\lambda_1 > \lambda_2 > \dots > \lambda_N. \quad (6)$$

- 6) Potom sú určené vlastné vektory $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N$ prislúchajúce vlastným číslam. Pretože matica C je symetrická, vlastné vektory formujú tzv. *lineárnu bázu*. To znamená, že každý vektor $\mathbf{x}$, resp. $(\mathbf{x} - \bar{\mathbf{x}})$ môže byť vyjadrený ako lineárna kombinácia vlastných vektorov:

$$(\mathbf{x} - \bar{\mathbf{x}}) = b_1 \mathbf{u}_1 + b_2 \mathbf{u}_2 + \dots + b_N \mathbf{u}_N = \sum_{i=1}^N b_i \mathbf{u}_i. \quad (7)$$

- 7) Krok redukcie rozmeru je vykonaný zachovaním len tých vlastných vektorov, ktoré zodpovedajú K najväčším vlastným hodnotám:

$$\hat{\mathbf{x}} - \bar{\mathbf{x}} = \sum_{i=1}^K b_i \mathbf{u}_i, \quad \text{kde } K \ll N. \quad (8)$$

Výstupná reprezentácia množiny $\hat{\mathbf{x}} - \bar{\mathbf{x}}$ v báze $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K$ je teda $Y = [b_1, b_2, \dots, b_K]^T$.

8) Finálna lineárna transformácia $\mathbf{R}^N \rightarrow \mathbf{R}^K$ je vypočítaná ako multiplikácia matíc:

$$Y = U^T(\mathbf{x} - \bar{\mathbf{x}}), \quad (9)$$

kde Y reprezentuje transformované dáta a U je transformačná matica PCA.

Pri určovaní finálneho rozmeru modelu sa môže vyskytnúť problém týkajúci sa determinácie konštanty K . Pre tento účel je vhodné využiť nasledovné komparatívne kritérium pre voľbu K :

$$\frac{\sum_{i=1}^K \lambda_i}{\sum_{i=1}^N \lambda_i} > T. \quad (10)$$

V praxi sa prah (v angl. threshold) T volí v rozsahu 0,9-0,95 [5]. Táto hodnota vyjadruje stupeň väzby medzi vstupnými a výstupnými premennými. Ak sa prah zvolí blízky jednej, potom výstupná reprezentácia bude mať vyššie K a naopak.

IV. OPIS A PODMIENKY EXPERIMENTOV

Všetky experimenty boli vykonané na rečovej databáze TIMIT, ktorá obsahuje celkovo 6300 nahrávok rozdelených do trénovacej a testovacej množiny. Extrakcia príznakov, tréning akustických modelov a samotné testovanie bolo realizované s využitím nástroja HTK (Hidden Markov Model Toolkit). Operácie spojené s transformáciou PCA boli vykonané v prostredí Matlab.

Pred tréningom modelov bolo využité mapovanie foném. 60 anglických foném bolo redukovaných na počet 40. Mapovacia schéma je uvedená v Tab. 1. Skryté Markovove modely boli

Tabuľka 1
 MAPOVACIA SCHÉMA FONÉM

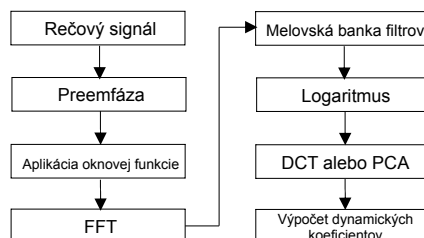
TIMIT	ao	ax	axh	axr	hv	ix	el	em	en
MAP	aa	ah	ah	er	hh	ih	l	m	n
TIMIT	nx	eng	zh	*cl	h#	pau	epi	ux	q
MAP	n	ng	sh	-	sil	sil	epi	uw	sp

trénované pre 40 foném (po mapovaní). Každý stav HMM bol modelovaný 1-1024 Gaussovskými zmesami. Za účelom testovania akustických modelov bol vytvorený aj jazykový model (bigramový backoff model vyhladený Wittenovou-Bellovou metódou). Bol natrénovaný z celej rečovej databázy s využitím nástroja SRI Language Modeling Toolkit [7].

Základné fázy predspracovania sú rovnaké pre obe metódy. Hlavný rozdiel je v použití rozdielnych dekorelačných metód. Prvá je klasická diskretná kosínusová transformácia (DCT), ktorá sa obyčajne používa v MFC parametrizácii. My sme sa sústredili na druhú metódu dekorelácie a transformácie - PCA.

A. Akustický model na báze MFC

Vstupný rečový signál je predspracovaný klasickým postupom. Najprv sa aplikuje preemfáza a Hammingova oknová funkcia, ktorej dĺžka trvania bola nastavená na 25ms (jeden rámec reči). Perióda rámcovania bola 10ms. Na oknované segmenty bola aplikovaná rýchla Fourierova transformácia (FFT). V ďalšom kroku bola aplikovaná melovská analýza a nakoniec bola aplikovaná logaritmická funkcia na lineárne výstupy melovskej banky filtrov. Výsledkom tohto postupu boli príznaky typu FBANK. Táto procedúra je ilustrovaná na Obr. 1. Po aplikácii logaritmu bolo



Obr. 1 Predspracovanie rečového signálu

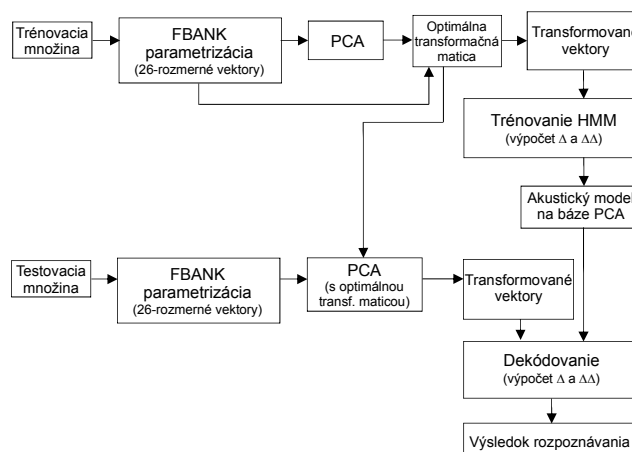
nutné dekorelovať vektory. V prípade MFC modelu bola pre tento účel zvolená DCT. Výsledkom tejto operácie boli 6-rozmerné vektory (statické kestrálne koeficienty). Toto číslo bolo zvolené v konfiguračnom súbore pri MFC extrakcii. Je vhodné poznamenať, že najprv bola vykonaná PCA, kde bol determinovaný optimálny rozmer na 18 (s delta a akceleračnými koeficientami). Za

účelom porovnania oboch modelov bolo potrebné vytvoriť MFC model s rovnakým rozmerom. 18-rozmerný MFC model vznikol výpočtom delta a akceleračných koeficientov z uvedených šiestich statických, výsledkom čoho boli 18-rozmerné MFC vektory.

B. Akustický model na báze PCA

Pre tento model sú všetky kroky predspracovania identické s predchádzajúcim modelom, rozdiel je len v aplikácii PCA namiesto DCT. Počet kanálov v melovskej banke filtrov bol nastavený na 26. Tieto 26-rozmerné vektory boli použité ako vstup pre PCA. Ako bolo spomenuté, po PCA transformácii boli výsledkom 6-rozmerné transformované vektory. Delta a akceleračné koeficienty boli dopočítané z týchto základných prostredníctvom nástroja HTK počas tréningovej a testovacej fázy (viď Obr. 2). Tieto boli potom pridané k základným, výsledkom čoho boli 18-rozmerné vektory.

Pre získanie transformačnej matice bola hlavná PCA analýza aplikovaná na tréningovú množinu nasledujúcim spôsobom. Každá rečová nahrávka je reprezentovaná separátnou dátovou maticou s logaritmickými melovskými príznakmi. Vstupná matica pre PCA bola vytvorená zrežaním týchto jednotlivých separátnych matic do jednej veľkej matice. Na túto maticu bola potom aplikovaná PCA podľa matematického postupu opísaného v kapitole III-B. Výsledkom bola optimálna transformačná matica s vhodným rozmerom, ktorá bola použitá na transformáciu tréningovej aj testovacej množiny vektorov. Pri testovaní bola teda použitá matica z tréningového procesu. Transformované vektory boli potom použité na tréningovanie HMM.



Obr. 2 Tréningový a testovací proces s využitím PCA

V. POROVNANIE VÝSLEDKOV EXPERIMENTOV

Pri vyhodnocovaní experimentov boli využité dva parametre. Prvým je tzv. *chybovosť* (v angl. word error rate, skr. WER) a druhým je *presnosť* (v angl. accuracy, skr. Acc.). Výsledky sú uvedené v Tab. 2 a Tab. 3. Z tabuliek je zrejmé, že výsledky pre PCA model sú lepšie približne o 1%. Najlepšie výsledky boli dosiahnuté pre prípad modelovania HMM stavu 256 a 512 Gaussovskými zmesami.

Tabuľka 2
 PARAMETER WER PRE MFC A PCA MODEL

Zmes	1	2	4	8	16	32	64	128	256	512	1024
MFC	40,07%	38,66%	35,87%	35,15%	34,35%	33,12%	32,39%	32,03%	31,63%	31,82%	33,63%
PCA	39,33%	37,95%	36,36%	34,19%	33,32%	32,61%	32,0%	31,09%	30,92%	31,09%	32,86%

Tabuľka 3
 PARAMETER ACC. PRE MFC A PCA MODEL

Zmes	1	2	4	8	16	32	64	128	256	512	1024
MFC	62,97%	64,4%	66,97%	67,86%	68,48%	69,66%	70,32%	70,87%	71,21%	71,2%	69,9%
PCA	63,9%	65,36%	66,9%	68,98%	69,85%	70,6%	71,32%	72,12%	72,47%	72,61%	71,19%

VI. ZÁVER

V tejto práci je prezentované porovnanie dvoch rozdielnych akustických modelov s využitím rozdielnych dekorelačných techník. Výsledky ukazujú, že model založený na PCA vykazuje lepšie presnosti a nižšiu chybovosť v porovnaní s klasickým MFC modelom.

V budúcnosti plánujeme použiť inú lineárnu transformáciu okrem PCA (napr. lineárnu diskriminačnú analýzu - LDA).

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] P. Cunningham, "Dimension reduction," University College Dublin, Tech. Rep. UCD-CSI-2007-7, Aug. 2007.
- [2] X. Wang and D. O'Shaughnessy, "Improving the efficiency of automatic speech recognition by feature transformation and dimensionality reduction," in *Proc. of the 8th European Conference on Speech Communication and Technology, EUROSPEECH – INTERSPEECH*, Geneva, Switzerland, Sep. 2003, pp. 1025–1028.
- [3] P. Somervuo, B. Chen, and Q. Zhu, "Feature transformations and combinations for improving ASR performance," in *Proc. of European Conf. on Speech Communication and Technology*, 2003, pp. 477–480.
- [4] I. T. Jolliffe, *Principal Component Analysis*. New York, USA: Springer-Verlag, 1986.
- [5] G. Bebis, *Principal Components Analysis*, Department of Computer Science, University of Nevada, Reno, 2003. [Online]. Available: <http://www.cse.unr.edu/~bebis/MathMethods/PCA/lecture.pdf>
- [6] L. I. Smith, *A Tutorial on Principal Components Analysis*, Feb. 2002. [Online]. Available: <http://users.ecs.soton.ac.uk/hbr03r/pa037042.pdf>
- [7] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of the 7th Intl. Conf. on Spoken Language Processing*, Denver, Colorado, Sep. 2002, pp. 901–904.

Bezdrôtové senzorové siete v automobiloch

Kamil ŠINDLERY, Stanislav MARCHEVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

kamilsindlery@rls-chiptuning.sk, stanislav.marchevsky@tuke.sk

Abstrakt — V tomto článku opisujeme technológiu bezdrôtových senzorových sietí použitú v automobiloch. Aplikáciu bezdrôtových senzorových sietí v automobiloch považujeme za bezdrôtový senzorový systém schopný zbierať, spracovávať a poskytovať viaceré typy technických informácií počas cesty automobilom. Bezdrôtové senzorové siete poskytujú redundantnú, bezchybnú bezdrôtovú komunikáciu medzi senzormi, ovládačmi a kontrolérmi.

Kľúčové slová — Bezdrôtové Senzorové Siete (BSS), Car to Car (C2C), Senzor, Wireless Sensor Networks (WSNs)

I. ÚVOD

A. Sensory

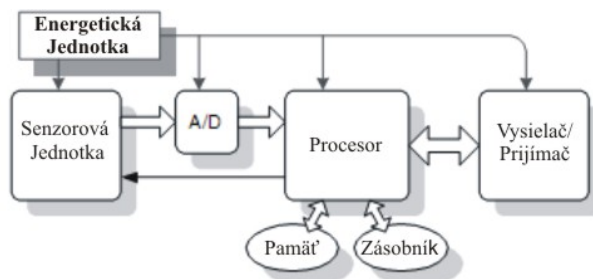
Senzor je fyzické zariadenie alebo biologický orgán, ktorý zisťuje signál, fyzikálne podmienky alebo chemickú zlúčeninu. Snímač prevádza fyzikálnu veličinu na inú veličinu. Všeobecná definícia snímača je nasledovná; snímač prevádza informáciu z fyzikálnej oblasti meranej veličiny do inej fyzikálnej oblasti, napr. na unifikovaný signál, najčastejšie na elektrický signál [1].

Rozdelenie snímačov používaných v automobiloch:

- Laserové
- Radarové (dosah > 150 m)
- Ultrazvukové (oblasť v blízkosti vozidla < 3 m)
- Infračervené (oblasť nočného videnia < 150 m)
- Videokamery (stredný rozsah < 80 m)

B. Senzorové uzly

Senzorový uzol je časť v senzorovej sieti, ktorá je schopná vykonávať určité spracovanie dát, zbierať senzorové informácie a komunikovať s ostatnými uzlami zapojenými v sieti. Typická architektúra senzorového uzla je ukázaná na Obr. 1.



Obr. 1 Štruktúra senzorového uzla

II. SENZOROVÉ SIETE PRE AUTOMOBILOVÚ PREVÁDZKU

Senzorové siete sa stali počas posledných rokov predmetom intenzívneho výskumu. Aplikácie senzorových sietí sú využívané v širokom rozsahu v priemyslových a štandardných prostrediach. Nepreskúmanou oblasťou zostávajú doprava a automobilové oblasti zamerané na

bezpečnostné systémy so senzorovými sieťami. Riadenie dopravy v mestách je stále komplikovanejšie a je teda potrebné, aby sme pomocou kontroly nad dopravou dokázali redukovať rozsah dopravnej nehodovosti a zlepšiť plynulosť dopravy. Použitie bezdrôtovej senzorovej siete môže byť vhodné riešenie pre blízku budúcnosť. Bezdrôtové inteligentné senzory by mohli byť nainštalované na každej križovatke. Tieto senzory by komunikovali medzi sebou a sledovali dopravnú situáciu. Správca komunikácie by mohol potom použiť tieto informácie pre skorú registráciu nehôd a vodiči by mohli využívať tieto informácie na skrátenie času cestovania [2].

Služby priamo súvisiace s bezpečnosťou premávky:

- varovanie na križovatkách
- varovanie pri dopravných značkách
- varovanie pri nehodách
- varovanie pri práci na vozovke
- varovanie pri zlých poveternostných podmienkach

A. Komunikácia Car-to-Car

Pre vodičov je informácia o vzájomnej vzdialenosti ich automobilov rozhodujúcim faktorom pre schopnosť bezpečne riadiť automobil.

Autá v súčasnej dobe už používajú systémy, ktoré sú schopné snímať životné prostredie, ale tieto systémy nevedia posielat informácie medzi vozidlami ako aj medzi vozidlami a krajinou prostredníctvom bezdrôtovej komunikácie. Informácie získané vozidlom nie sú dostupné pre ostatných. Pri komunikáciách V2V (Vehicle to Vehicle), C2C (Car to Car) a V2R (Vehicle to Roadside) informácie zosnímané senzorom a prístrojmi vozidla sú dostupné a môžu byť rýchlo odovzdané k ostatným vozidlám v bezprostrednej blízkosti, pomocou ad-hoc sieťovej komunikácie. Komunikácie C2C a V2V sú v podstate rovnaké, rozdiel je v tom, že v USA sa to nazýva V2V a v Európe C2C. Rozšírené informácie zo senzorov umožňujú vodičom rýchlejšie reagovať pri núdzových situáciách, napr. kratší reakčný čas pri brzdení. Komunikácia vozidiel s okrajom cesty použitím rovnakej základnej technológie poskytuje dodatočné informácie pre zvýšenie pohodlia a komfortu, ako napr. špeciálne informácie dostupné pozdĺž diaľnice.

B. Sledovanie automobilov

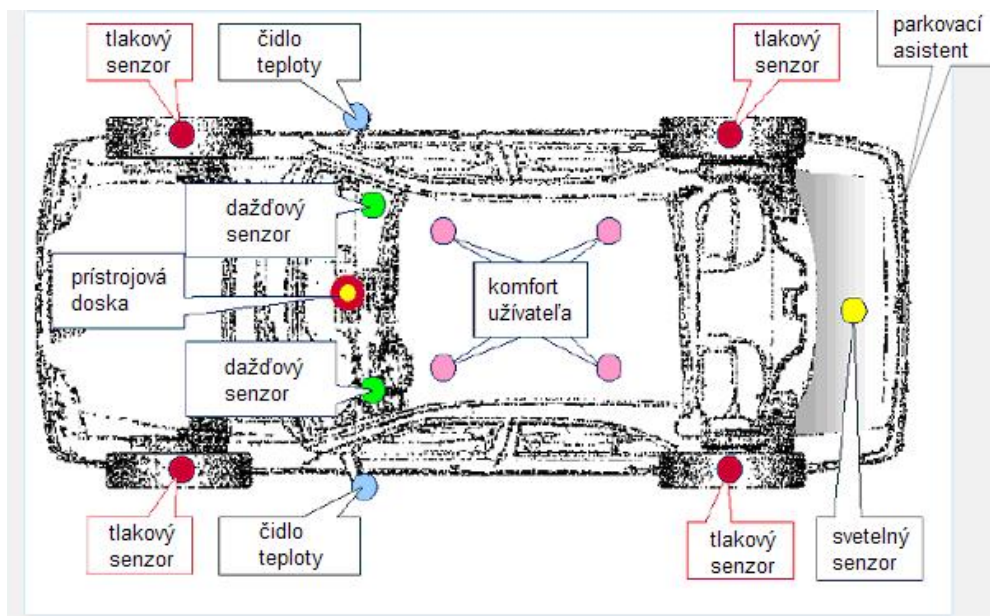
Bezdrôtová senzorová sieť je tvorená kompaktnými a autonómnymi uzlami, ktoré sú vybavené funkciami pre zhromažďovanie a spracovanie dát alebo bezdrôtovú komunikáciu. Uzly sú použité na monitorovanie a určenie geografickej polohy. Každý uzol zbiera dáta o svojej polohe. Zozbierané údaje sú potom použité na rôzne účely. Pretože uzly pracujú vo fyzikálnom prostredí, znalosť o ich vlastnej fyzikálnej polohe je rozhodujúca. Polohová informácia môže byť získaná prostredníctvom niekoľko metód. Systém GPS je jednou z metód, ktoré poskytujú absolútnu polohovú informáciu. Z ekonomických dôvodov, majú GPS prijímače len niektoré senzorové uzly. Pravidelne vysielajú trigonometrický signál o vlastnej polohe a tak ďalšie uzly bez GPS prijímačov môžu určiť svoju aproximatívnu polohu v teréne. Táto aplikácia umožňuje senzorom v sieti lokalizovať špecificky pohybujúce sa automobily a hlásiť ich geografickú polohu [3], [4].

Základňou stanica vysielá ID sledovaného automobilu. Senzory vo všetkých automobiloch prijímajú toto vysielanie a preposielajú správu, ak nie sú predmetom hľadania. Vozidlo, ktoré je hľadané začne po prijatí tejto správy vysielat svoju zemepisnú šírku a dĺžku [5].

III. SENZORY V AUTE

Senzor je snímač schopný zmerať fyzikálnu, chemickú alebo biologickú hodnotu veličiny a reprezentovať ju na výstupe, najlepšie ako elektrický signál. Senzor predstavuje aj akčný člen, ktorý práve naopak, na základe elektrického signálu generuje výstup. Výstupom môže byť zopnutie nejakého väčšieho zariadenia, prípadne jeho riadenie či regulácia. Inteligentný snímač je rozšírenou verziou svojho jednoduchého predchodcu. Rovnako obsahuje obvody pre úpravu meranej veličiny, A/D prevodník, mikroprocesor pre ďalšie spracovanie a vyhodnocovanie a jednotku pre komunikáciu s ostatnými senzormi v sieti [5].

Moderné automobily obsahujú množstvo rôznych senzorov nachádzajúcich sa v interiéri ako aj v exteriéri auta. Senzory v exteriéri auta ako napríklad parkovacie senzory zvyšujú komfort vodiča pri riadení auta. Senzory tlaku v kolesách automobilu zvyšujú bezpečnosť počas jeho prevádzky.

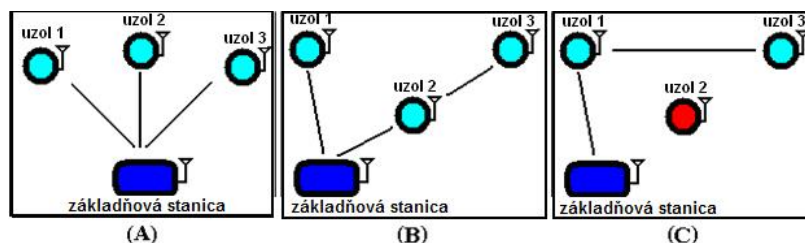


Obr. 2 Senzory v aute

IV. SMEROVANIE

Bezdrôtové siete vyžadujú pohotovú smerový protokol na udržanie kroku s rýchlymi zmenami v topológii siete, kvôli zmene stavu v uzloch alebo spojeniach. Preto sú použité metódy ako geografické alebo energiu kontrolujúce smerovanie. Kombinácia týchto techník môže byť použitá na vytvorenie vysoko kvalitných spojení so základňovou stanicou, ktoré minimalizujú energiu. Prediktory kvality spojenia môžu byť ľubovoľné parametre výkonu, ako napríklad výkon alebo oneskorenie z konca na koniec alebo čokoľvek iné. Prioritou v protokole je spoľahlivosť spojenia. Spoľahlivosť je definovaná ako paket správne dekódovaný v prijímači. Napríklad, vysielateľ sa vždy pokúsi poslať pakety priamo najprv na základňovú stanicu. Ak pakety nemôžu byť úspešne vysielané, potom vysielateľ skočí k inému uzlu, ktorý má priame spoľahlivé spojenie so základňovou stanicou. Tento proces bude naďalej kontrolovať spoľahlivé vysielanie z priameho spojenia k 3 odrazovým spojeniam v sieti špecifikované na Obr. 3, kým nebude nájdená spoľahlivá cesta. Preto sa bude stále meniť topológia siete, aby sa získala najlepšia spoľahlivosť pre všetky uzly [6].

Obr. 3 ukazuje príklad algoritmu Multi-hop s Crossbow®. Na Obr. 3 (A) majú všetky uzly priame spojenie so základňovou stanicou ako najspoľahlivejšie spojenie. Avšak, Obr. 3 (B) ukazuje, že ak uzol 3 stratí priame spojenie so základňovou stanicou, vytvorí odrazové spojenie so základňovou stanicou cez uzol 2. Obr. 3 (C), uzol 3 vytvára jedno náhradné spojenie so základňovou stanicou cez uzol 1, keďže uzol 2 je neaktívny kvôli chybe. Smerovanie a topológia siete v Crossbow® reaguje na nestabilné spojenia a chyby uzlov použitím smerovania Multi-hop, ktoré tiež pomôže, ak je dôležitá kvalita služby (QoS).



Obr. 3 Multi-Hop Crossbow

A. Smerovacie algoritmy

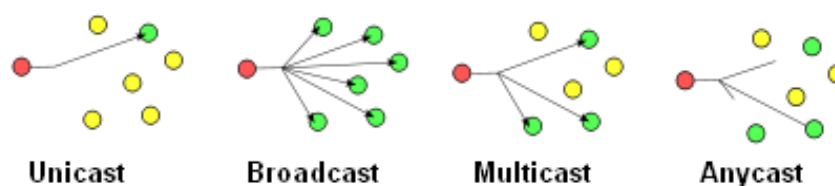
Smerovacie algoritmy sú veľmi dôležité pre každú sieť. V našom prípade, čiže pre senzorové siete, to platí ešte mnohonásobne viac. Celkové vlastnosti siete závisia od voľby vhodného smerovacieho algoritmu. Hlavné vlastnosti dobrého smerovacieho algoritmu sú šetrenie s energiou uzlov a prispôboivosť k meniacim sa okolnostiam (napr. počet a poloha uzlov). Smerovacie protokoly sú tak isto ako všetky protokoly pre senzorové siete tiež zamerané na

úsporu energie. To môže byť urobené tak, že sledujú zostávajúcu energiu v uzloch alebo energiu potrebnú na prenos správ. Energeticky efektívna cesta je vybraná pomocou jedného z nasledujúcich prístupov [7]:

- Cesta s minimálnou energiou - volí sa cesta medzi uzlom a bránou, ktorá spotrebuje minimálne množstvo energie na doručenie paketu.
- Cesta s maximálnou zvyšnou energiou - cesta k bráne, ktorá má maximálnu zvyšnú energiu, t.j. súčet energií uzlov pozdĺž cesty.
- Cesta s minimom skokov (hopov) – algoritmus vyberie cestu s minimálnym počtom skokov. Ak energia potrebná na prenos paketu je rovnaká pre každú cestu, tak metóda s minimálnou energiou a minimom skokov sú ekvivalentné.

Metódy smerovania sa líšia vo forme doručovania [8]:

- Unicast - doručenie správy k jedinému špecifikovanému uzlu;
- Broadcast - doručenie správy pre všetky uzly v sieti;
- Multicast - doručenie správy pre skupinu uzlov, ktoré prejavia záujem o prijímanie správ;
- Anycast - doručenie správy k jednému uzlu zo skupiny, zvyčajne k najbližšiemu uzlu k zdroju.



Obr. 4 Metódy smerovania

V. VYSIELANIE

Senzorové siete môžeme rozdeliť podľa princípu fungovania do troch kategórií:

- (1) Siete, ktoré neustále snímajú prostredie a posielajú dáta do základňovej stanice.
- (2) Siete, ktoré spustia len určité udalosti alebo situácie (napr. dopravné priestupky, požiar).
- (3) Siete, ktoré snímajú podľa pokynov užívateľa.

Na doručenie dát k základňovej stanici sa používajú už spomínané smerovacie algoritmy. V tretej skupine najprv treba zadať sieti úlohu a túto úlohu treba doručiť ku každému uzlu. Na tento účel je použité vysielanie (broadcasting), keď treba doručiť správu k všetkým uzlom v sieti alebo multicasting, keď len pre skupinu uzlov.

VI. VLASTNOSTI

A. Spotreba energie senzorového uzla

Senzorové uzly majú obmedzenú zásobu energie z batérií a metódy na znovu nabitie batérií sú v tomto smere drahé a komplexné. Preto je nutné, aby senzorové uzly používali energiu efektívnym spôsobom. Hlavné úlohy, ktoré potrebujú energiu, sú zisťovanie dát, komunikácia a riadenie. Energia môže byť šetrená navrhnutím protokolov s ohľadom na energiu v aplikácii, spojení a v sieťových vrstvách. Manažment energie vo WSN je veľmi dôležitým problémom, lebo rádiová komunikácia dominuje v spotrebe energie v senzorovom uzle. Sú rôzne protokoly, ktoré šetria energiu limitovaním rádiovkej komunikácie ako sa len dá. Napríklad, senzorové uzly môžu byť „uspaté“ na určitú dobu, keď nie je sledovaný žiaden zaujímavý prípad. Ale je žiaduce, aby senzorové uzly boli schopné zobudiť sa, aby monitorovali kritické prípady, ak sa objavajú.

B. Spoľahlivosť

Sú situácie, kde senzory pracujú vo vzdialenom alebo problémovom prostredí, kde môže byť nemožné oživiť uzly za účelom znovu nabitia batérií. Preto by mala sieť mať určitú životnosť, počas ktorej uzly majú energiu a môžu snímať, spracovávať a vysielat informácie. To znamená, že všetky aspekty musia byť navrhnuté tak, aby boli senzory extrémne efektívne.

C. Bezpečnosť

Bezpečnosť senzorových sietí je problém, ktorému sa zatiaľ venuje málo pozornosti. Je nutné zvoliť správnu rovnováhu medzi použitou pamäťou, spotrebovanou energiou a robustnosťou bezpečnosti. Senzorové siete vyžadujú protokoly s minimálnym dopadom na výkon, ľahkou implementáciou a energetickou nenáročnosťou.

VII. ZÁVER

V tomto článku bola vykonaná analýza použitia senzorov a senzorových sietí v automobiloch. Opísali sme ich vlastnosti, ako aj spôsoby smerovania a vysielania v týchto sieťach.

Bezdrôtové senzorové siete ako technológia budúcnosti má veľké perspektívy. Súčasný stav je taký, že dopyt po tejto technológii sa stále zvyšuje, lebo zákazníci oceňujú jej vysokú kvalitu a flexibilitu.

S istotou sa dá očakávať, že v blízkej dobe sa bezdrôtové senzory a senzorové siete významnou mierou budú podieľať na zvýšení bezpečnosti a spoľahlivosti automobilov, ako na zvýšení bezpečnosti a efektívnosti cestnej premávky.

POĎAKOVANIE

Táto práca bola podporovaná Kultúrnou a edukačnou grantovou agentúrou Ministerstva školstva Slovenskej republiky VEGA na základe projektu č.1/0045/10 (50%) a realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] VLK, FRANTIŠEK: Automobilová elektronika 1: Asistenční a informační systémy. Prvé vydanie, ISBN 80 – 239 – 6462 – 3, Brno, (2006).
- [2] YOO, SEONG-MOO: Sensor Networks for Automobile Tracking and Routing, Electrical and Computer Engineering Department, The University of Alabama, UTCA Report 03303, DTSR 0023424, (2004).
- [3] Global Positioning System, Serving the World. Dostupné na internete: < <http://www.gps.gov/> >
- [4] DANA, PETER H.: Global Positioning System Overview, The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder, (1994) Dostupné na internete: <http://www.colorado.edu/geography/gcraft/notes/gps/gps_f.html >
- [5] NAGY, NORBERT: Architektúry senzorových sietí pre bezpečnú automobilovú prevádzku. Diplomová práca. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, (2009), s. 33.
- [6] HAENSELMANN, THOMAS: Sensornetworks, GFDL Wireless Sensor Network textbook, (2006).
- [7] AKYLDIZ, I. F. – SU, W. – SANKARASUBRAMANIAM, Y. – CAYIRCI, E: A Survey on Sensor Networks, IEEE Communications Magazine, Aug (2002), pages 102–114.
- [8] Network routing in all its purity: Tips, Howto's & Info on BGP, IPv6, TCP/IP. Dostupné na internete: <<http://routingpurity.com/2009/03/22/address-types-unicast-multicast-anycast/> >

Manažment spotreby energie v inteligentných budovách pomocou multi-agentov

Kamil ŠINDLERY, Stanislav MARCHEVSKÝ

Katedra elektroniky a multimediálnych telekomunikácií, Fakulta elektrotechniky
a informatiky, Technická univerzita v Košiciach, Slovenská republika

kamilsindlery@rls-chiptuning.sk, stanislav.marchevsky@tuke.sk

Abstrakt — V článku sa venujeme problematike riadenia napájania a spotreby energie v inteligentných budovách na báze teórie multi-agentov. Článok sa zameriava na riešenie stratégie a najmä na spôsob zjednodušenia riadenia energetických zdrojov v inteligentnom dome. V článku je poukázané na to, že sledovanie a profilovanie správania viac obyvateľov v inteligentnom prostredí je problém, ktorý je vysoko aktuálny z hľadiska ďalšieho výskumu a vývoja riadenia energetických zdrojov v inteligentnom dome s využitím teórie inteligentných agentov a moderných technológií.

Kľúčové slová — Multi-Agent Home Automation System (MAHAS), Multi-Agentové Systémy (MAS), Systém Domácej Automatizácie (SDA)

I. ÚVOD

Svetová populácia dnes rastie o viac ako 80 miliónov osôb ročne, teda spotreba energie sa zvyšuje stále viac a viac. Budova nie je len miestom spotreby elektrickej energie, ale aj potenciálnym miestom decentralizovanej výroby elektriny pomocou zdrojov, ako je veterná, slnečná, geotermálna energia a pod. Stredisko pre obnoviteľné zdroje energie považuje domov za miesto, ktoré predstavuje 47% svetovej spotreby energie [1]. Úloha Systému Domácej Automatizácie (SDA) venovaná riadeniu spotreby je prispôsobenie spotreby energie z dostupných energetických zdrojov, s prihliadnutím na kritériá pohodlia obyvateľov a dosiahnutie kompromisu medzi komfortom obyvateľov a nákladov na prevádzku. Dilemu komfortného a súčasne ekonomického bývania možno vyriešiť pomocou moderných technológií.

A. Inteligentné prostredie

Inteligentné prostredie, inteligentný dom, inteligentná budova to sú viaceré názvy, ktoré pomenovávajú v podstate tú istú vec. Aj keď v dnešnej dobe neexistuje univerzálne prijatá definícia inteligentnej budovy, tak definície v rôznych štúdiách a projektoch a tiež definície firiem, ktoré sa týmto zaoberajú majú spoločné vlastnosti. A to sú:

- úspora energie,
- väčší komfort,
- bezpečnosť obyvateľov.

Európska skupina pre inteligentné budovy založená v Anglicku definovala inteligentnú budovu ako budovu, ktorá „vytvára také vnútorné prostredie, ktoré maximalizuje schopnosť správnej činnosti obyvateľov budovy a zároveň umožňuje účinný manažment zdrojov s minimálnymi nákladmi na zariadenia a vybavenie počas životnosti budovy“ [2].

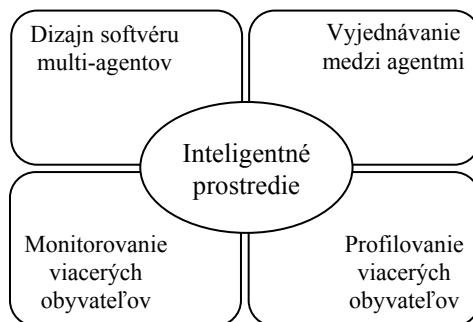
Pojem inteligentný dom je podľa inej definície chápaný ako riadiaci systém integrujúci v sebe všetky spolupracujúce technológie v rodinnom dome. Význam riadenia technológií sa zvyšuje s množstvom systémov, ktoré sa v objekte nachádzajú. Čím rozsiahlejšie sú požiadavky užívateľov na komfort a funkcie objektu, tým zložitejší je celý objekt a tým väčšie nároky sú kladené na riadiaci systém [3].

B. Multi - agenti

Skôr ako budeme skúmať problematiku multi-agentov pre inteligentné prostredie, musíme určiť to, čo myslíme ako "agenta" v súvislosti s inteligentným prostredím. Z jedného pohľadu, sa za agenta môže považovať softvérový komponent, ktorý je autonómny, reaktívny, sociálny a proaktívny, ktorý prispieva k softvérovej architektúre inteligentného prostredia. Z iného hľadiska, môžeme tvrdiť, že inteligentné prostredie bytov je prirodzeným umiestnením multi-agentov. V druhom prípade, obyvatelia sami o sebe môžu byť považovaní za inteligentných agentov. Navyše, softvér musí byť navrhnutý tak, aby komunikoval s každým bydliskom, a tak transformoval softvér sám o sebe do multi-agentových systémov.

Vychádzajúc z týchto rôznych pohľadov na multi-agentové systémy pre softvérové prostredia, sme našli štyri hlavné smery vo výskume multi-agentov pre inteligentné prostredie, ktoré sú v súčasnosti sledované. Tieto smery sú zobrazené na obrázku 1. Obrázok ukazuje, že aktívny výskum prebieha v štyroch hlavných oblastiach. Výskum v prvej oblasti sa zameriava na dizajn softvéru multi-agentov pre inteligentné prostredie. Ďalšie poskytujú riešenia na výzvy, ktoré vznikajú u viacerých obyvateľov v inteligentnom prostredí: monitorovanie viacerých obyvateľov, profilovanie činnosti a správania sa viacerých obyvateľov a vyjednávanie medzi potrebami a preferenciami viacerých agentov [4].

Algoritmy založené na multi-agentových systémoch (MAS) sú v súčasnosti používané v niekoľkých oblastiach, ako je informatika alebo automatizácia. Prvé MAS prístupy k energetickým distribúciám boli prezentované v [5] a [6].



Obr. 1 Súčasné smery vo výskume multi-agentov pre inteligentné prostredie

II. MULTI-AGENTOVÝ DOMÁCI AUTOMATIZAČNÝ SYSTÉM

Návrh riešenia riadenia spotreby energie v inteligentných budovách založených na multi-agentových systémoch, ktoré sú vhodné pre riešenie priestorovo distribuovaných a otvorených problémov, uľahčuje návrh inteligentného Multi-agentového domáceho automatizačného systému (MAHAS).

Hlavné znaky MAHAS sú nasledovné:

- Distribúcia: energetické zdroje a zariadenia sú nezávislé a priestorovo distribuované.
- Flexibilita: niektoré zariadenia môžu akumulovať energiu (rôzne druhy vykurovania), alebo použitie s časovým oneskorením požiadavky obyvateľov (pranie, atď.).
- Otvorenosť: počet pripojených zdrojov a zariadení sa môže meniť s časom, bez toho, aby sa musel kompletne predefinovať riadiaci mechanizmus.
- Možnosť rozšírenia: môžu sa pridať noví agenti určení pre ďalšie služby [1].

MAHAS sa skladá z agentov, z ktorých každý podporuje jeden typ služieb (kúrenie, varenie, atď.) dosiahnutý jedným alebo viacerými zariadeniami.

V multi-agentových systémoch, pojem riadenia zahŕňa operácie ako koordináciu a vyjednávanie medzi agentmi, eliminácia agentov a pridávanie nových agentov v prípade potreby. Štruktúra MAHAS je rozložená do dvoch hlavných mechanizmov: reaktívny mechanizmus a anticipatívny mechanizmus.

A. Reaktívny mechanizmus

Reaktívny mechanizmus je rýchly regulačný mechanizmus, ktorý sa spúšťa, keď úroveň uspokojenia požiadavky agenta klesne pod nedostatočné hodnoty (napr. 10%). Tento mechanizmus, ktorý sa spolieha na vyjednávanie protokolu [7], reaguje rýchlo, aby sa zabránilo

prerušovaniu energie a obmedzeniam a zaručuje dobrú úroveň spokojnosti obyvateľov. Preto, reaktívny mechanizmus nastaví v krátkom čase súbor bodov z predikovaného plánu kritéria, zariadenie je v aktuálnom stave pre obmedzenia a kritéria obyvateľov. Predpovedané nastavené body môžu byť priamo odovzdané zariadeniu alebo upravené v prípade núdze [1].

B. Anticipatívny mechanizmus

Reaktívny mechanizmus je dostatočný, aby sa zabránilo obmedzeniam prerušovania, ale MAHAS možno zlepšiť, aby sa zabránilo častým mimoriadnym situáciám. Toto zlepšenie je dosiahnuté anticipatívnym mechanizmom. Cieľom tohto mechanizmu je výpočet plánov na výrobu a spotrebu služieb v dome. Anticipatívny algoritmus ťaží zo skutočnosti, že na jednej strane sú zariadenia, ktoré sú schopné akumulovať energiu a na druhej strane sú niektoré služby, ktoré majú premenlivý čas ich vykonania; niektoré služby môžu byť buď oneskorené alebo posunuté v čase. Ak MAHAS dokáže predpokladať spotreby energie jednotlivých zariadení, existuje možnosť, že bude riadiť tieto zariadenia efektívnejšie a tým dosiahne nižšiu spotrebu energie.

Tento mechanizmus pracuje v časovom okne, čo zodpovedá vzorkovacej perióde tzv. anticipatívnej perióde. Táto perióda je väčšia ako používa reaktívny mechanizmus. Vzhľadom na veľkú vzorkovaciu periódu, vzorky považuje za priemerné hodnoty energie. To je výhoda, keď uvažujeme predikciu, pretože je ťažké urobiť presné predikcie. Veľká vzorkovacia perióda je tiež dôležitá, aby reaktívny mechanizmus ostal prehľadný. Predikciou nastavené body môžu byť priamo prenášané na zariadenie alebo upravené reaktívnym mechanizmom v prípade obmedzenia porušenia [1].

III. ARCHITEKTÚRA INTELIGENTNÉHO PROSTREDIA MULTI-AGENTOV

Komplexné systémy, ako sú inteligentné prostredia vyžadujú veľké množstvo softvérových komponentov. Každý z týchto komponentov, ak obsahuje určitý stupeň samostatnosti a je vedený v súlade s inteligentným návrhom, možno považovať za inteligentného agenta. S cieľom zabezpečiť, aby kolekcia softvérových agentov bola robustná, jasná a prehľadná, títo vsadení agenti potrebujú vzájomne komunikovať a spolupracovať [4].

Pri navrhovaní multi-agentových prístupov, je preto potrebné výskumom zvážiť:

- (1) predpokladané softvérové úlohy každého agenta,
- (2) typ organizácie, ktorý bude existovať medzi agentmi,
- (2) spôsob, akým agenti budú spolupracovať a zdieľať informácie [4].

Existuje viac prístupov pre definovanie úloh agentov. Definovať úlohy možno podľa funkcie, podľa typu zariadenia, alebo priestorom. Prístup kde sú úlohy agentov definované priestorom bol skúmaný podľa [8]. Priestoroví agenti sú definovaní pre oblasti domu a sú nasadení hierarchicky. Sú pod kontrolou Microsoft Simple Control Protocol a Universal Plug and Play technology. Ako príklad na definovanie úloh agentov podľa funkcie môžeme uviesť [9], kde bola navrhnutá architektúra multi-agentového systému pre inteligentné prostredie, ktoré využíva ZigBee bezdrôtovú senzorovú sieť pre komunikáciu. Pre adresovanie organizačného čísla agenta, sú definované tri vrstvy agentov. Agenti na najnižšej vrstve sú koncové zariadenia pre ZigBee siete na spojenie so senzormi a regulátormi. V strednej vrstve, v ktorej každý agent prispieva špecifickými funkciami, ako je identifikácia správania obyvateľa a preferencie - priority. Napokon, agent na najvyššej vrstve riadi celý systém.

IV. MAHAS MODELOVANIE

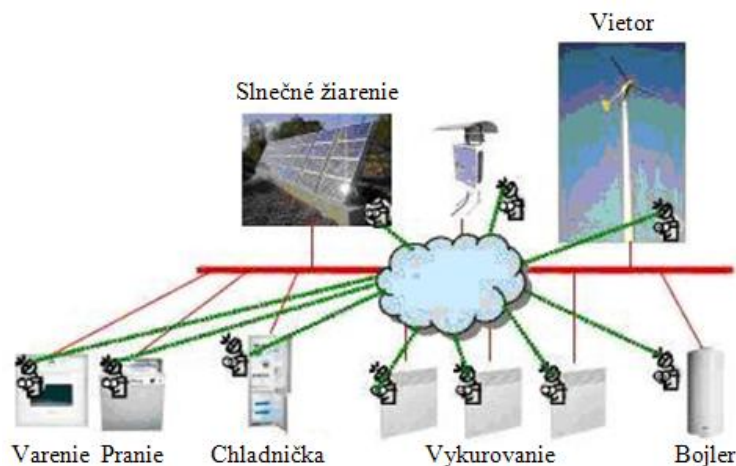
Systém MAHAS je zameraný na dosiahnutie kompromisu medzi požiadavkami obyvateľov na komfort a náklady a zároveň uspokojivým technologickým obmedzením zariadení. MAHAS sa zameriava na minimalizáciu nákladov pri maximalizácii pohodlia užívateľov. Preto MAHAS berie do úvahy rozdiely v cenách, pretože dodávatelia energie môžu účtovať obyvateľom rôzne ceny za energiu v závislosti od času jej spotreby [1].

Jedným z cieľov systému MAHAS je zabezpečiť pohodlie užívateľa. Je to pojem, ktorý môže byť priamo spojený s konceptom spokojnosti funkcie [10]. Spokojnosť funkcie bola definovaná pre energetické zdroje, ako aj pre zariadenia. Spokojnosť funkcie je vyjadrená definovaním funkcie z oblasti intervalu charakteristických premenných [0, 1]. V MAHAS, sa spokojnosť obyvateľov charakterizuje ako pocit zo služby, kde 0 znamená "neprijateľné" a 1 znamená "úplne spokojný".

Spokojnosť funkcie môže byť modelovaná súborom charakteristických bodov. Spokojnosť

funkcie agenta pre energetické zdroje je taktiež vyjadrená funkciou, kde je úroveň spokojnosti definovaná dodávateľom elektrickej energie.

MAHAS je „otvorený“ a „rozšíriteľný“, pretože je tam mnoho druhov služieb. Pre optimalizáciu kritérií, musia agenti komunikovať a spolupracovať pri výmene správ založenej na všeobecne známom modeli. MAHAS je vhodný najmä preto, že agenti dodávaní do zariadenia môžu vložiť vlastné znalosti a vymieňať si správy len na základe zdieľaných znalostí, ktoré môžu byť analyzované všetkými agentami (napríklad: napájací profil služby predstavuje sériu spotrebovanej alebo dodanej energie zodpovedajúcej po sebe idúcim periódam vzorkovania).



Obr. 2 MAHAS

V. MODELOVANIE AGENTOV

V MAHAS, agent podporuje službu $SRV_i \in SRVS$, kde $SRVS$ je súbor všetkých služieb. Služba SRV_i môže byť buď časovaná služba, označená SRV_i^T , alebo permanentná služba, označená SRV_i^P , dosiahnutá pomocou jedného alebo viacerých zariadení s vedomím, že zariadenie môže tiež dosiahnuť viac služieb.

A. Modelovanie agenta časovanej služby

Časovaná služba zodpovedá profilu napájania počas danej periódy, flexibilita služby pochádza z možnosti, že sa môže časovo posúvať. Interné dáta agenta časovanej služby sú modelované podľa [1]:

- charakteristickej premennej, ktorá zodpovedá počiatkovej dobe tejto služby.
- správania modelu, ktorý definuje možnosť profilu napájania potom, čo bola služba spustená.
- spokojnosti funkcie definovanej z množiny možných štartovacích časov do súboru možných hodnôt spokojnosti.

Realizácia pre agenta časovanej služby je zdieľanie poznatkov o časovanej službe. To sa skladá z:

- profilu napájania: $\Pi = [POW(SRV_i)_0, \dots, POW(SRV_i)_n]$, kde $POW(SRV_i)_j \neq 0$ reprezentuje priemerný výkon spotreby / produkcie (negatívne / pozitívne) služby počas k-tej vzorkovacej periódy. Doba trvania služby je stanovená podľa dĺžky profilu napájania $length(\Pi)$.
- hodnoty spokojnosti, ktorá závisí od štartovacieho času služieb.

B. Modelovanie agenta permanentnej služby

Permanentná služba zodpovedá profilom napájania za danú periódu. Flexibilita služby pochádza z možnosti priebežnej úpravy rozdelenia energie. Vnútorne poznatky agenta permanentnej služby sú modelované podľa:

- modelu správania, spájajúceho profil napájania s účinkom na prostredie obyvateľa.
- spokojnosti funkcie, definovanej zo súboru premenných, vypočítaných z modelu správania do súboru možných hodnôt spokojnosti. Parametre spokojnosti funkcie môžu byť zhromaždené v súbore charakteristických bodov. Vzhľadom k KVK (kúrenie, ventilácia a klimatizácia) systému, medzi tieto charakteristické body patrí, obyvateľom požadovaná teplota a obyvateľom neprijateľná teplota.

Realizácia pre agenta permanentnej služby zastupuje zdieľanie poznatkov o časovanej službe. To sa skladá z:

- profilu napájania: $\Pi = [..., \text{POW}(\text{SRV}_i)_k, ...]$, kde $\text{POW}(\text{SRV}_i)_k$ predstavuje priemernú spotrebu energie / produkcie počas k-tej periódy.
- hodnoty spokojnosti. Aj keď realizácia zahŕňa niekoľko períód, je tam len jedna hodnota, z ktorej vyplývajú modely spokojnosti obyvateľa za celú posudzovanú periódu: to sa zvyčajne rovná minimálnej hodnote spokojnosti vypočítanej pre každú periódu.

VI. SLEDOVANIE VIAC OBYVATEĽOV

Potom, čo viac obyvateľov vstúpi do inteligentného prostredia, služby, ktoré poskytuje inteligentné prostredie sú oveľa ťažšie realizovateľné. Obyvatelia sa pohybujú a konajú inak, keď sú v prítomnosti ostatných obyvateľov. Okrem toho, interakcie medzi obyvateľmi sú podstatou im vlastnou, ktoré môžu využívať s funkciami, ktoré inteligentné prostredie môže poskytnúť.

Ťažkosti, ktoré vznikajú v situáciách s viac obyvateľmi sa odrážajú v skutočnosti, že niekoľko výskumníkov už riešilo tieto otázky. Funkcia, ktorá je požadovaná na inteligentnom prostredí je schopnosť nájsť a sledovať obyvateľa v danom priestore. Sledovanie jednotlivého obyvateľa môže byť vykonané tým, že sa pozerá na históriu aktivácie snímača pohybu. Keď viac obyvateľov môže vstúpiť do priestoru, potom už neexistuje priamočiare riešenie. Je potrebné prijať určité riešenie, ktoré by zjednodušilo danú úlohu. Napríklad, by sme mohli požiadať obyvateľov aby nosili zariadenie, ktoré umožňuje ich sledovanie v priestore [11], [12]. Toto partikulárne riešenie je nepraktické v situáciách, v ktorých jednotlivci nechcú nosiť zariadenia, zabudnú nosiť zariadenia, alebo vstupujú a opúšťajú prostredie príliš často.

Zatiaľ čo sledovanie viac obyvateľov je jasne náročný problém a v skutočnosti bolo preukázané, že je NP-Hard [13], výskumníci navrhujú kreatívne prístupy k riešeniu tohto problému. Napríklad v [14], majú vstavané senzory tlaku do podlahy ich inteligentného prostredia, ktoré identifikujú a lokalizujú každého obyvateľa na základe jeho hmotnosti. Táto metóda má však obmedzenie v tom, že nie je schopná rozlišovať medzi jednotlivcami, ak majú blízke hodnoty hmotnosti, alebo sú hneď vedľa seba v priestore. Iní výskumníci spájajú viac zdrojov informácií na sledovanie viacerých obyvateľov v priestore. Ako napríklad, v [15] sa kombinuje zvuk a obraz na sledovanie viacerých ľudí. Informácie z týchto multi-modálnych zdrojov sa používajú na odvodenie pozorovania pravdepodobnosťou.

VII. PROFILOVANIE SPRÁVANIA PRE VIAC OBYVATEĽOV

Za účelom poskytovania služieb, ako je sledovanie zdravotného stavu a pomoc, automatizácie úloh a analýzy činnosti, musí sa inteligentné prostredie naučiť profily správania pre všetky osoby, ktoré obývajú prostredie. Rozpoznávanie aktivity je dobre rozpracovaný výskumný problém. Výskumníci úspešne zvládli riešenie modelov pre rozpoznávanie aktivity z údajov senzora pohybu [14], [16] údajov z prenosných osobných senzorov [17] a ďalších informácií z bezdrôtových senzorov [18], aj keď iba v nastavení prostredia pre jedného obyvateľa. Veľmi málo práce bolo zameranej na učenie profilov správania pre viac obyvateľov.

Počiatočné úsilie bolo nasmerované k týmto cieľom profilovania nastavenia viac obyvateľov v jednom prostredí. Konkrétne Crandall a Cook [19] navrhli prístupy strojového učenia k priradeniu udalosti senzora na osobu v priestore, ktorá spustí udalosť. Majú porovnať Naïvov Bayesov klasifikátor a Markov model prístupu k tejto klasifikácii problému ako predchodca pre výučbu celkovej činnosti profilov. Experimentovanie s dátami z inteligentných bytov ukazuje, že mapovanie udalostí senzorov na obyvateľa je dobre uskutočniteľné. Naïvov Bayesov klasifikátor sa ukázal ako najlepší a dosiahol 98% zaradenie presnosti.

VIII. STRATÉGIA OPTIMALIZÁCIE

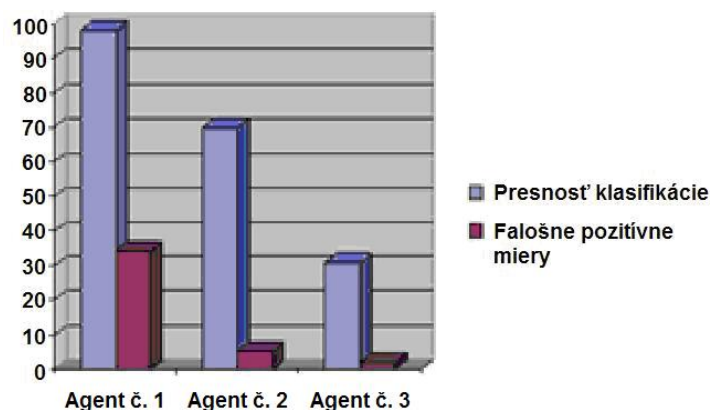
Problém inteligentného riadenia napájania v budovách môže byť rozdelený do sub-problémov zahŕňajúcich rôznych agentov pretože, vo všeobecnosti, obyvatelia nepoužívajú ich zariadenia dlhodobo. Základným princípom je rozdeliť celý problém do samostatného sub-problému, potom riešiť každý sub-problém nezávisle s cieľom nájsť riešenie pre celý problém. Výhodou tejto metódy je zníženie zložitosti celého problému, ktorý závisí od počtu períód pre každý sub-problém a od počtu zariadení. Celý problém je rozdelený na sub-problémy tak, že každý sub-problém nezahŕňa všetky služby. Pretože počet stanovených períód v sub-probléme je nižší ako počet períód celého problému [20], zložitosť riešenia sub-problémov je menšia než zložitosť riešenia celého problému naraz.

IX. KLASIFIKÁCIA

Časová hodnota je dôležitý parameter dát pre algoritmy strojového učenia. Za účelom pohľadu na čas, v ktorom udalosti nastanú boli vytvorené zložitejšie funkcie na zachytenie rozdielov v správaní medzi obyvateľmi. Začlenením časových informácií do parametrov dát, môže tento druh informácie zlepšiť presnosť klasifikátora.

Naivov Bayesov klasifikátor je druh klasifikátora, ktorý bol s veľkou účinnosťou používaný v iných projektoch výskumu inteligentných domov [21].

Obrázok 3 ukazuje správnosť klasifikácie Naivovho Bayesovho klasifikátora pre tri osoby, ktoré sa zúčastnili istého výskumu. V grafe je zobrazená presnosť klasifikácie pre každú osobu, ale tiež aj falošne pozitívne miery.



Obr. 3 Presnosť klasifikácie osôb [22]

Z grafu je vidieť, že Agent č. 1 má najvyššiu presnosť klasifikácie. To je spôsobené tým, že Agent č. 1 bol zodpovedný za väčšinu senzorových udalostí. Z toho vyplýva, že bude veľká pravdepodobnosť, že akákoľvek senzorová udalosť bude mapovaná na Agentu č. 1 a Naivov Bayesov klasifikátor nesprávne priradil hodnoty udalosti Agentov č. 2 a č. 3 Agentovi č. 1.

Za účelom riešenia týchto chýb klasifikácie boli pridané ďalšie opisné vlastnosti do dátového súboru. Typy funkcií ako sú *hodiny dňa*, *dni v týždni*, *časť dňa* a *časť týždňa* môže teraz klasifikátor využiť na rozlíšenie medzi správaním rôznych obyvateľov. Tabuľka I zobrazuje vzorku dát, ktorá bola použitá na klasifikáciu Agentu č. 2.

TABUĽKA I
 VZORKA DÁT POUŽITÁ NA KLASIFIKÁCIU

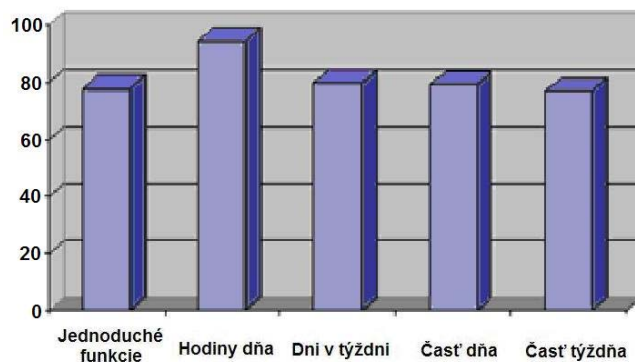
Dátum	Čas	Sériové číslo	Správa	ID
21.12.2007	16:41:41	07.70.eb:1	ON	Agent č. 2
21.12.2007	16:44:36	07.70.eb:1	OFF	Agent č. 2

TABUĽKA II
 TYPY FUNKCIÍ NA REPREZENTÁCIU ČASOVÝCH HODNÔT

Typ funkcie	Príklad
Jednoduché funkcie	07.70.eb : 1#ON
Hodiny dňa	07.70.eb : 1#ON#6
Dni v týždni	07.70.eb : 1#ON#PIA
Časť dňa	07.70.eb : 1#ON#RÁNO
Časť týždňa	07.70.eb : 1#ON#PRACOVNÝ DEŇ

Tabuľka II zobrazuje príklady časových a dátových informácií senzorových udalostí, ktoré môže klasifikátor využiť na lepšie rozlíšenie medzi obyvateľmi.

Pri pohľade na presnosť miery klasifikácie ovplyvnenú výberom typu funkcie, obrázok 4 ukazuje, že použitím funkcie *hodiny dňa* sa výrazne zvyšuje priemerná identifikácia. Navyše, použitím tejto funkcie, klesne falošne pozitívna miera. Ak sú vybrané správne funkcie zo súboru dát, klasifikátor je schopný urobiť lepšiu celkovú klasifikáciu.

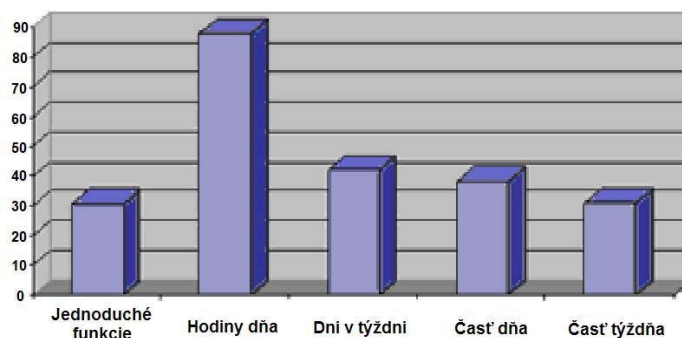


Obr. 4 Priemerná presnosť klasifikácie pre typy funkcií [22]

V tomto súbore dát majú ďalšie funkcie, vrátane *dni v týždni*, *časti dňa* a *časti týždňa* malé zlepšenie, než jednoduché funkcie. So správnosťou viac ako 93% a falošnou pozitívnou mierou nižšou ako 7%, strojová predikcia spoliehajúca sa na tento klasifikátor môže mať vysokú mieru dôveryhodnosti, že to je správny výber vhodných preferencií pre daného obyvateľa.

Preukázanie zlepšenia v presnosti klasifikácie ilustruje Obr. 5. Počiatočná presnosť Agentu č. 3 bola s jednoduchými funkciami 31% (Obr. 3), ale jej hodnota stúpila na 87% opäť pomocou funkcie typu *hodiny dňa* (Obr. 5).

Funkcia *hodiny dňa* preukázala najlepšie zlepšenie presnosti klasifikácie pre všetky osoby. Žiadna z ostatných funkcií nie je taká užitočná ako práve táto funkcia.



Obr. 5 Presnosť klasifikácie Agentu č. 3 pre typy funkcií [22]

X. ZÁVER

V tomto článku boli analyzované vybrané metódy používané pri riešení problémov manažmentu spotreby energie v inteligentných budovách založené na teórii inteligentných multi-agentov. Boli ukázané zásady MAHAS, ktoré umožňujú agentom spolupracovať a koordinovať svoje opatrenia s cieľom nájsť prijateľné optimálne riešenie pre riadenie spotreby. Je tu opísaný princíp modelovania MAHAS-u a tiež aj agentov časovanej a permanentnej služby. Vykonaný prehľad vybraných dosiahnutých výsledkov ukazuje, že riešenie manažmentu spotreby energie v inteligentných budovách na báze inteligentných multi-agentov predstavuje perspektívnu a spoločensky prospešnú oblasť výskumu. To či by MAHAS a multi-agenti, ktorí riadia jednotlivé časti inteligentnej budovy dokázali znížiť spotrebu energie v reálnych podmienkach je veľkou výzvou pre náš ďalší výskum a jeho testovanie nielen v laboratórnych podmienkach.

POĎAKOVANIE

Táto práca bola podporovaná Kultúrnou a edukačnou grantovou agentúrou Ministerstva školstva Slovenskej republiky VEGA na základe projektu č.1/0045/10 (50%) a realizáciou projektu Centrum informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120020) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] ABRAS,S., PESTY, S., PLOIX, S., JACOMINO,M.. An Anticipation Mechanism for Power Management in a Smart Home using Multi-Agent Systems, Grenoble, Saint Martin d Heres, France, (2008).
- [2] ASB, Architektúra Stavebníctvo Biznis, Inteligentné budovy na bývanie: Komfort či obmedzovanie? <http://www.asb.sk/stavebnictvo/technologie/inteligentne-budovy-na-byvanie-komfort-ci-obmedzovanie-1076.html>
- [3] Nizkoenergetické EKO bývanie, Inteligentný dom – mýtus a realita. <http://www.ekobyvanie.sk/priklady-z-praxe/rodinne-domy/inteligentn-dom-mtus-a-realita/>
- [4] DIANE J. COOK. Multi-agent smart environments. *Journal of Ambient Intelligence and Smart Environments* 1, (2009), pages 47–51.
- [5] JENNINGS, N. R. The ARCHON system and its applications. In *Second International Working Conference on Cooperating Knowledge Based Systems (CKBS-94)*, Keele, UK, (1994), pages 13–29.
- [6] H'AGG, S. and YGGE, F. Agent-oriented programming in power distribution automation. PhD thesis, University of Karlskrona/Ronneby, Ronneby, Sweden, (1995).
- [7] ABRAS, S., PLOIX, S., PESTY, S., and JACOMINO, M.. A multi-agent home automation system for power management. In *Proceedings of the Third International Conference in Control, Automation, and Robotics, ICINCO*, Setubal, Portugal, (2006), pages 3–8.
- [8] CHEN,W.,and TSENG,W.: A novel multi-agent framework for the design of home automation. *Proceedings of the International Conference on Information Technology*, (2007), pages 277–281.
- [9] DUAN,P. and LI,H.. Zigbee wireless sensor network based multi-agent architecture in intelligent inhabited environments. *Proceedings of the International Conference on Intelligent Environments*, (2008).
- [10] LUCIDARME, P., SIMONIN,O. and LI'EGEOIS,A.: Implementation and evaluation of a satisfaction/altruism based architecture for multi-robot systems. In: *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, Washington, USA, (2002), pages 1007–1012.
- [11] HIGHTOWER,J. and BORRIELLO,G.: Location systems for ubiquitous computing. *IEEE Computer*, 32(8), August (2001), pages 57–66.
- [12] JYIN,J., YANG,Q. and SHEN,D.. Activity recognition via usertrace segmentation. *ACM Transactions on Sensor Networks*, November (2008).
- [13] ROY,N., ROY,A., and DAS,S.K.: Context-aware resource management in multi-inhabitant smart homes: A Nash h-learning based approach. *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, July (2006), pages 148–158.
- [14] LU, C., HO,Y., and FU,L. Creating robust activity maps using wireless sensor network in a smart home. *Proceedings of the Third Annual Conference on Automation Science and Engineering*, (2007).
- [15] CHECKA,N., WILSON,K., SIRACUSA, M. and T. DARRELL,T.:Multiple person and speaker activity tracking with a particle filter. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (2004), pages 881–884.
- [16] WREN,C.R. and E.M. TAPIA, E.M.: Toward scalable activity recognition for sensor networks. *Proceedings of the International Workshop on Location- and Context-Awareness*, 3987, (2006), pages 168–185.
- [17] MAURER,U., SMAILAGIC,A., SIEWIOREK,D. and M. DEISHER: Activity recognition and monitoring using multiple sensors on different body positions. *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks*, (2006).
- [18] SANCHEZ,D., TENTORI,M. and FAVELA,J.: Activity recognition for the smart hospital. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35, (2005), pages 133–144.
- [19] CRANDALL,A. and D. COOK,D.: Attributing events to individuals in multi-inhabitant environments. *Proceedings of the International Conference on Intelligent Environments*, (2008).
- [20] HABBAS, Z., KRAJECKI,M. and D. SINGER. Decomposition techniques for parallel resolution of constraint satisfaction problems in shared memory: a comparative study. *International Journal of Computational Science and Engineering (IJCSE)* 1(2), (2005), pages 192–206.
- [21] BAHL,P. and PADMANABHAN,V., Radar: An in-building rf-based user location and tracking system. In *Proceedings of IEEE Infocom*, (2000), pages 775–784.
- [22] CRANDALL,A. and COOK,D.J.: Learning Activity Models for Multiple Agents in a Smart Space, (2009).

Formal Description of Behaviour for Object-based Programs

Marián JENČÍK, Valerie NOVITZKÁ

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

marian.jencik@student.tuke.sk, valerie.novitzka@tuke.sk

Abstract – This paper introduces definition of object-based program in category theory. We will define category of software object and category of object-based program. Behaviour of categorically defined object-based program we will describe by coalgebra. Description of behaviour for object-based program by coalgebra is based on observable values which are generated as the results of program execution and captured by polynomial endofunctor. Program execution can be characterized as a code execution of the methods which are looked up by messages sent between software objects. Message is in this case a specific value of data type.

Keywords – software object, coalgebra, category

I. INTRODUCTION

Object-oriented programming is a process, during which, according to the properties and behaviour of real-world objects is created a program that reflects these properties and behaviour [1], [2]. Basic structural unit of object-oriented program is software object, whose structure resembles to structure of real-world object. The software object is thus a model of real-world object.

Software object has two basic features :

- **state of object** - is stored in variables, called *attributes*. Attributes generally represent private encapsulated data and each attribute is considered to be a type;
- **behaviour of object** - is determined by operations working over the states of the object, which are called *methods* (Fig. 1).

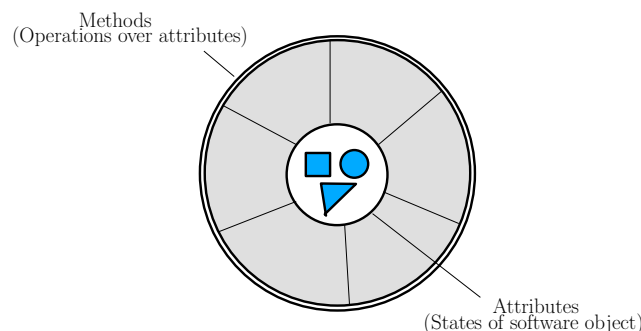


Fig. 1 Software object

The aim of programming, including object-oriented programming is an effort to ensure that computer would perform programmed actions and generate the expected behaviour, that arises in interaction with the environment, where outputs are generated by inputs and outputs are then displayed on output device. This behaviour can be either positive, desired, or negative, undesired, that we want to eliminate.

II. COALGEBRA

Program's behaviour we can theoretically describe by coalgebra. Coalgebras [3], [4], [5] are an original contribution for behaviour of description of running programs, because description is based on observable behaviour. Observable values are generated by program execution and captured by polynomial endofunctor over a suitable signature.

Coalgebra of polynomial endofunctor can be defined as follows :

Definition 2.1: A coalgebra of polynomial endofunctor $G_{Set} : Set \rightarrow Set$ in category of sets and functions is a pair

$$(W_{Set}, w),$$

where

- W_{Set} is an object of category Set , so called *carrierset*
- and w is a morphism of category, so called *coalgebraic operation*, or *dynamics*

$$w : W_{Set} \rightarrow G(W_{Set}).$$

Coalgebraic operation is ordered n-tuple of functions, deconstructors in the form :

$$w = [\![deconstr_1]\!], [\![deconstr_2]\!], \dots, [\![deconstr_n]\!] : W_{Set} \rightarrow G(W_{Set}).$$

□

III. CATEGORY OF SOFTWARE OBJECT

Attributes, as we can see in definition of *software object*, are types. Let's have a finite set of typing term variables $x_i : a_i$, for $1 \leq i \leq n$,

$$Var = \{x_1 : a_1, \dots, x_n : a_n\}, \quad \text{where } \Gamma = \{x_1 : a_1, \dots, x_n : a_n\} \text{ is type context.}$$

Terms are written in the form of sequent

$$\Gamma \vdash t_i : a_{n+1}.$$

If $t_1, \dots, t_n$ are terms of types $a_1, \dots, a_n$ and

$$m_i : a_1 \times \dots \times a_n \rightarrow a_{n+1},$$

for $1, \dots, n$, are methods from product of input types to the final type (**Fig. 2**), then application of method, which can be written as

$$m_i(t_1, \dots, t_n) : a_{n+1},$$

or

$$\dots m_2(m_1(t_1), m_1(t_2)) \dots \text{ of type } a_{n+1}$$

is a term.

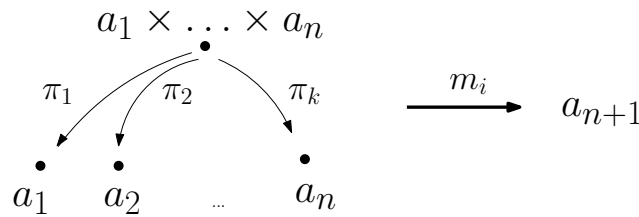


Fig. 2 Method from product of input types to the final type

We can construct the *category of software object* [6] as follows :

Category of software object - $SObject$ is category, in which :

- objects are type contexts $\Gamma = (x_1 : a_1, \dots, x_n : a_n)$, $\Theta = (x'_1 : a'_1, \dots, x'_{n'} : a'_{n'}) \dots$, that can be written as product of types, $\Gamma = (a_1 \times a_2 \times \dots \times a_n)$, $\Theta = (a'_1 \times a'_2 \times \dots \times a'_{n'})$;
- morphisms are maps $[t_1, \dots, t_n] : \Gamma \rightarrow \Theta$, which represent n-tuples of terms such, that :

$$\Gamma \vdash t_1 : a'_1, \dots, \Gamma \vdash t_n : a'_{n'};$$

- identity can be written as $id : \Gamma \rightarrow \Gamma$;
- composition of morphisms is defined by the rule :

$$\frac{\Gamma \xrightarrow{t_1, \dots, t_n} \Theta \quad \Theta \xrightarrow{l_1, \dots, l_k} \Delta}{\Gamma \xrightarrow{v_1, \dots, v_h} \Delta},$$

that it holds

$$v_i = l_j[t_1/x'_1, \dots, t_n/x'_n].$$

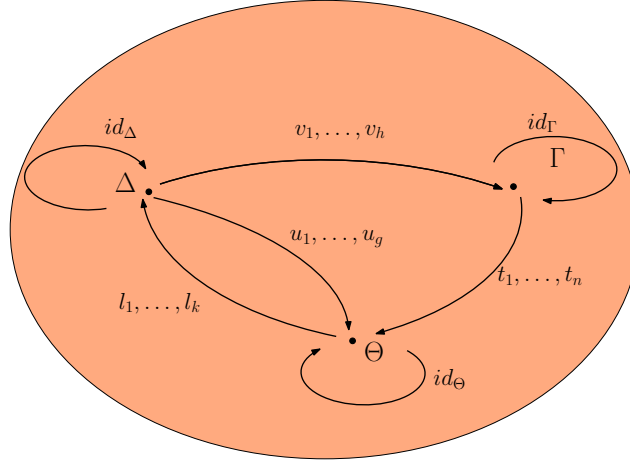


Fig. 3 Category of software object - *SObject*

Category constructed in this manner is :

- *cartesian category*, because it has finite number of products $a_1 \times \dots \times a_n$;
- *closed category*, because it uses functional types;
- *locally small category*, because collection of all morphisms creates a set for all pairs of objects.

IV. CATEGORY OF OBJECT-BASED PROGRAM

Category of object-based program - *PprogOB* [6] can be defined as category of locally small categories, in which :

- objects are particular categories $SObject_1, \dots, SObject_n$;
- morphisms are functors between categories

$$F_i : SObject_i \rightarrow SObject_j;$$

Each functor F_i consists of two components (F_{i_o}, F_{i_m}) :

- F_{i_o} for mapping objects from category $SObject_i$ to the category $SObject_j$

$$F_{i_o} : (SObject_i)_{obj} \rightarrow (SObject_j)_{obj}$$

and for component F_{i_o} definition of functor F_i it holds :

$$F_{i_o}(a_{i_u}) = a'_{j_k};$$

- F_{i_m} for mapping morphisms from category $SObject_i$ to the category $SObject_j$

$$F_{i_m} : (SObject_i)_{morph} \rightarrow (SObject_j)_{morph}$$

and for component F_{i_m} definition of functor F_i it holds :

$$\begin{array}{l} \text{if } \underbrace{m_{i_s}}_{\in SObject_i} : \underbrace{a_{i_u}}_{\in SObject_i} \rightarrow \underbrace{a_{i_v}}_{\in SObject_i}, \\ \text{then } \underbrace{F_{i_m}(m_{i_s})}_{n_{j_l} \in SObject_j} : \underbrace{F_{i_o}(a_{i_u})}_{a'_{j_e} \in SObject_j} \rightarrow \underbrace{F_{i_o}(a_{i_v})}_{a'_{j_f} \in SObject_j}. \end{array}$$

where $a_{i_1}, \dots, a_{i_n}$ are objects (attributes) of the category $SObject_i$, $a'_{j_1}, \dots, a'_{j_p}$ are objects (attributes) of the category $SObject_j$ and $m_{i_1}, \dots, m_{i_k}$, $n_{j_1}, \dots, n_{j_q}$ are methods of the category $SObject_i$, respectively $SObject_j$,

- identity is written as

$$\underbrace{a_{i_j}}_{\in \mathcal{SObject}_i} \xrightarrow{id_{a_{i_j}}} \underbrace{a_{i_j}}_{\in \mathcal{SObject}_i} \mapsto \underbrace{F_{i_o}(a_{i_j})}_{\in \mathcal{SObject}_j} \xrightarrow{id_{F_{i_o}(a_{i_j})}} \underbrace{F_{i_o}(a_{i_j})}_{\in \mathcal{SObject}_j},$$

that is holds $F_{i_m}(id_{\mathcal{SObject}_i}) = id_{F_{i_o}(\mathcal{SObject}_i)}$;

- composition is written as

$$\underbrace{a_{i_j}}_{\in \mathcal{SObject}_i} \xrightarrow{m_{i_e}} \underbrace{a_{i_k}}_{\in \mathcal{SObject}_i} \xrightarrow{m_{i_f}} \underbrace{a_{i_l}}_{\in \mathcal{SObject}_i} \mapsto \underbrace{F_{i_o}(a_{i_j})}_{\in \mathcal{SObject}_j} \xrightarrow{F_{i_m}(m_{i_e})} \underbrace{F_{i_o}(a_{i_k})}_{\in \mathcal{SObject}_j} \xrightarrow{F_{i_m}(m_{i_f})} \underbrace{F_{i_o}(a_{i_l})}_{\in \mathcal{SObject}_j}$$

and for morphism composition it is holds :

$$F_{i_m}(m_{i_e} \circ m_{i_f}) = F_{i_m}(m_{i_e}) \circ F_{i_m}(m_{i_f}).$$

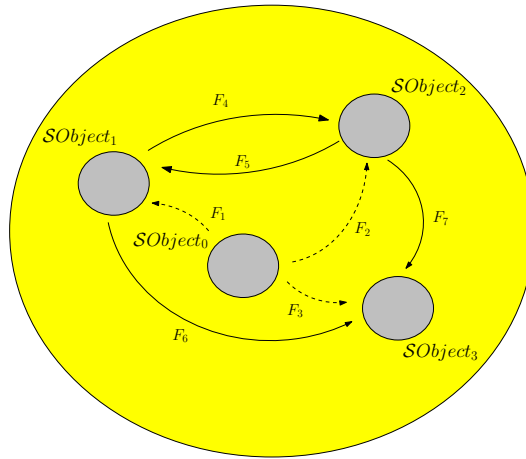


Fig. 4 Category of object-based program - ProgOB

In the category $PprogOB$ we distinguish three kinds of morphisms :

- 1) *Method look up*;
- 2) *Method adding*;
- 3) *Method redefining*.

Definition 4.1 (Method adding): *Method adding* we can define as an extension of the category $\mathcal{SObject}_j$ with method, thus *inclusion functor* is used. Category $\mathcal{SObject}_i$ is a subcategory of the category $\mathcal{SObject}_j$ and for functor F_i it holds :

$$F_i : \mathcal{SObject}_i \hookrightarrow \mathcal{SObject}_j.$$

This functor assigns to attributes and methods of the category $\mathcal{SObject}_i$ corresponding attributes, respectively methods in the category $\mathcal{SObject}_j$ and add new method to the category $\mathcal{SObject}_j$. □

Definition 4.2 (Method redefining): *Method redefining*, compared with method adding does not extend the $\mathcal{SObject}_j$ category, but only modifies the content of the category. Methods are replaced by the change of definition scope, by modifying of attributes. In that attributes are types, the modification of attributes causes the redefinition of these methods, where modified attributes occur.

Let's have methods

$$m_{i_s} : a_{i_1} \times \dots \times a_{i_k} \times \dots \times a_{i_n} \rightarrow a_{i_{n+1}} \quad \text{a} \quad n_{j_l} : a'_{j_1} \times \dots \times a'_{j_l} \times \dots \times a'_{j_p} \rightarrow a'_{j_{p+1}},$$

then for functor F_i it holds :

$$F_i : \mathcal{SObject}_i \rightarrow \mathcal{SObject}_j$$

and functor components are defined as :

$$F_{i_o}(a_{i_k}) = \begin{cases} a'_{j_k} & \text{for } k \neq l, \\ a_{j_l} & \text{for } k = l, \end{cases}$$

$$F_{i_m}(m_{i_s}) = F_{i_o}(a_{i_1}) \times \dots \times \underbrace{F_{i_o}(a_{i_k})}_{a'_{j_l} \in SObject_j} \dots \times F_{i_o}(a_{i_n}) = n_{j_l}.$$

□

Definition 4.3 (Method look up): Method look up we can define as code execution of the methods. The results of code execution will be observable values, thus states (values) are changed by attributes. Method look up is performed as response of the object $SObject_i$ to the message sent $SObject_j$ object. The message is a value of some data type in the category $ProgOB$.

In order to describe the principle of method look up in categorical notions, it is necessary to define:

- special object $\perp$ in category $SObject_j$;
- morphism $(\perp: a'_{j_l} \rightarrow \perp)$ in category $SObject_j$.

Method look up functor we will mark $F_c = (F_{c_o}, F_{c_m})$ and is defined as :

$$F_c : SObject_j \rightarrow SObject_i,$$

that it holds :

$$F_{c_o}(a'_{j_l}) = a_{i_k},$$

$$\underbrace{\perp}_{\in SObject_j} : \underbrace{a'_{j_l}}_{\in SObject_j} \rightarrow \underbrace{\perp}_{\in SObject_j} \mapsto \underbrace{F_{c_m}(\perp)}_{m_{i_s} \in SObject_i}.$$

□

Special object of category $ProgOB$ is an *empty object*, i.e. object, which contains no attributes and no methods. This object is marked as $SObject_0$ and it is an initial object of the category $PprogOB$, because from this object exists exactly one morphism, *method adding*, to all others objects of the category $ProgOB$. In general it can be said, that from empty object we can create any object of the category $ProgOB$ by using morphism method adding.

V. BEHAVIOUR DESCRIPTION OF OBJECT-BASED PROGRAM

For behaviour description of object-based program characterized by category $PprogOB$ we use coalgebra defined in section II. Using coalgebra we can observe the states of particular software objects based on observable values, which are results of code execution by methods look up. Software objects are *black boxes* for us, we are not interesting in their internal structure. Observable values are captured by polynomial endofunctor

$$G_{PprogOB} : ProgOB \rightarrow ProgOB,$$

that is constructed over objects and morphisms of the category $ProgOB$.

Polynomial endofunctor has the form :

$$G_{PprogOB}(SObject_i) = SObject_i \times A,$$

which expresses the action :

$$SObject_i \mapsto SObject_i \times A,$$

where A is the set of all observable values.

Coalgebraic operation

$$(SObject_i, \text{step})$$

of polynomial endofunctor $G_{PprogOB}$ has the form :

$$\text{step} : SObject_i \rightarrow (SObject_i \times A).$$

Carriersets of coalgebraic operations $SObject_i$ are state spaces of particular objects of category $ProgOB$. States of software objects, which are objects of category $ProgOB$, we can observe by observable values. Observable values are obtained by code execution, which is result of sending a message to the object and method look up. Method look up v_i is in this case deconstructor of coalgebraic operation step :

$$\text{step} = \langle \llbracket v_i \rrbracket \rangle : SObject_i \rightarrow G_{ProgOB}(SObject_i).$$

The result of multiple application of the operation step is a sequence :

$$(A_1, A_2, \dots) \in A^{\mathbb{N}},$$

where

$$A_1 = (a_1, \dots, a_n) \quad \text{a} \quad A_2 = (a'_1, \dots, a'_p),$$

represent the states of particular objects $SObject_i$ of the category $ProgOB$, occurred in attributes $a_1, \dots, a_n$, respectively $a'_1, \dots, a'_p$.

Final coalgebra can be expressed as a pair

$$(A^{\mathbb{N}}, \text{step})$$

and final coalgebraic operation of polynomial endofunctor G_{ProgOB} has the form :

$$\text{step} : A^{\mathbb{N}} \rightarrow A^{\mathbb{N}} \times A$$

VI. CONCLUSION

However object-oriented programming is very popular today it can be said that it has no characteristic formal base yet [7]. Basic structural unit of object-oriented program is software object and object-oriented programming can be characterized like a net of software objects communicating to each other. Software objects are black boxes, we do not know their internal structure, but we can observe their behaviour. It is useful to use coalgebra for behaviour description of software objects because description is based on observable variables that are generated as results by code execution of the methods which are looked up by messages sent between software objects.

If we can describe behaviour of object-based program we have to define suitable categorical structure. In our case it is category of object-based program. Category of object-based program is defined as a category of locally small categories, in which objects are particular categories of software objects and morphisms are functors among categories. In the category of object-based program we distinguish three kinds of morphisms : *method look up*, which is defined as a code execution of the methods, *method adding*, which is defined by *inclusion functor* as an extension of the category of software object and finally *method redefining*, which is defined as a change of definition scope of methods caused by attributes modifying.

Behaviour of object-based program is then described by coalgebra whereby observable values are captured by polynomial endofunctor and coalgebraic operation $(SObject_i, \text{step})$ of polynomial endofunctor is form $\text{step} : SObject_i \rightarrow (SObject_i \times A)$.

In our next research we would like to focus on suitable categorical structures for coalgebraic behaviour description of class-based programs and complex program systems based on components. We would like to apply achieved theoretical results to real non trivial object-oriented programs from the area of computer networks or database systems.

ACKNOWLEDGMENT

This work was supported by VEGA Grant No.1/0015/10: Principles and methods of semantic enrichment and adaptation of knowledge-based languages for automatic software development.

REFERENCES

- [1] O. Čada, *Objektové programování - naučte sa pravidla objektového myšlení*. Praha: Grada Publishing, a.s., 2009.
- [2] S. J. Chapman, *Introduction to JAVA*. Prentice Hall, June 1999.
- [3] D. Mihályi and V. Novitzká, *Princípy duality medzi konštruovaním a správaním programov*. Košice: Equilibria, 2010.
- [4] V. Novitzká and V. Slodičák, *Kategorické štruktúry a ich aplikácie v informatike*. Košice: Equilibria, 2010.
- [5] V. Slodičák and M. Ľalová, "Some useful structures for categorical approach for program behavior," in *Proceedings of the 21st Central European Conference on Information and Intelligent Systems, CECIIS 2010 Proceedings*. Varaždin, Croatia: Faculty of Organization and Informatics, University of Zagreb, September 22-24, 2010, pp. 477-484.
- [6] M. Jenčík, "Formálny popis správania objektovo-orientovaných programov," písomná práca k dizertačnej skúške, KPI FEI TUKE Košice, 2010.
- [7] M. Wand, "Type Inferences for Objects with Instance Variable and Inheritance," *Theoretical Aspects of Object-Oriented Programming*, MIT Press, pp. 99-120, 1994.

Integration Framework for Common Knowledge Representation from Models

Maroš TYRPÁK, Iveta TONHAUSEROVÁ, Zdeněk HAVLICE

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

maros.tyrpak@gmail.com, iveta.tonhauserova@tuke.sk, zdenek.havlice@tuke.sk

Abstract — Nowadays software projects have very large dimensions and developers need a lot of information to have good overview in the system they develop. Usually all information are independent and form different sources. This independence can result into inconsistency between these information and lead into unexpected errors.

This thesis describes and design solution which links different information about software project into one integrated unit called architectural knowledge.

Keywords — knowledge, ontology, OWL, model, integration, UML, JAVA

I. INTRODUCTION

Since 1970's software development is in crisis. Range and complexity of software systems grows as fast as requirements of software users. The result of this trend is that development and maintenance is very difficult, complex and expensive.

Documentation and models of software systems are one of most important things that can provide useful information about system and can make development and maintenance process easier. However, models and documentation are very often inaccurate and facts which specified there do not correspond to the reality and implementation.

II. KNOWLEDGE BASED SYSTEM FOR SOFTWARE DEVELOPMENT SUPPORT

Creating of system's design together with documentation is good practice which can reduce costs and errors. As software changes with requirements of customers, initial design does not match system's implementation. Very often new functionality is added to the system or existing functionality is changed, but these changes are not reflected in documentation and models. When new developers joins the project and relates to inaccurate and obsolete documentation the result is that new unexpected errors and bugs are created.

To prevent these problems there is needed to make documentation and models accurate. Rewriting of documentation is very boring, costs lot of time and money and nobody wants to do this work. And sometimes it is very hard to find differences and changes between implementation and design/documentation.

To minimize problems which originate as the result of inconsistency between software system's documentation and implementation there can be created application which would provide all kinds of information about system and its components in structured and integrated way. For example there can be part of class diagram together with information extracted from source codes and vice versa. If there will be some differences between these two parts, developer will know immediately that something is not correct.

On Fig. 1 is displayed architecture of knowledge based system which would provide architectural knowledge about specified software project. It consists from five main parts:

- Repository of software system contains software models, documentation, source codes and binaries
- Processing modules extracts knowledge form parts of software system
- Alignment layer links elements of particular parts

- Querying layer should provide interface for accessing architectural knowledge
- Visualization layer represents browser of architectural knowledge

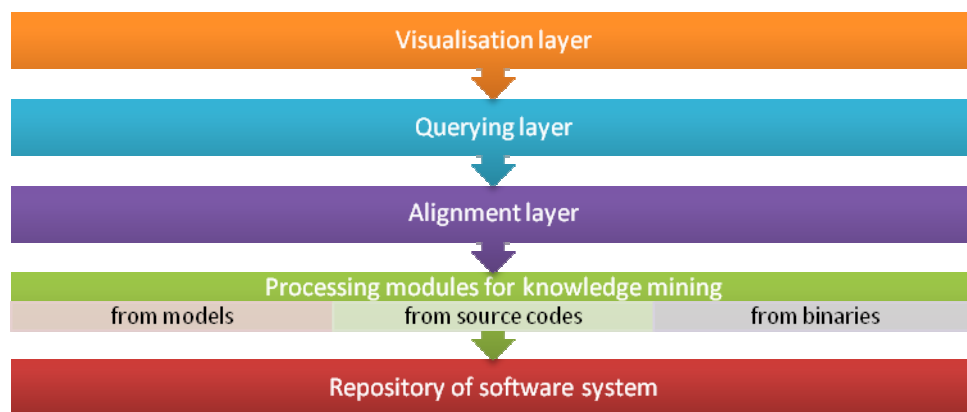


Fig. 1 Architecture of system for software development support

The target of this work is not to create the whole system, but only two parts of it – **alignment layer and querying layer**.

III. ARCHITECTURAL KNOWLEDGE

There are several types of knowledge. They are often separated into groups according to the problem domain which they describe. For example healthcare, automotive, e-commerce, banking, etc. Knowledge which describes software development process and software project is called **architectural knowledge**.

“Architectural knowledge presents a thorough understanding of specific software system. It is defined as the integrated representation of the software architecture, its structure, behavior, the architectural design decisions and the external context/environment.”[3]

Architectural knowledge is not generic. It does not describe several projects. One instance of architectural knowledge should describe just one particular software system because it contains project specific data.

Architectural knowledge should contain this information: [2]

- relations and dependencies between software artifacts
- relations and dependencies between modules and elements of software system
- structure and description of elements and interfaces
- relations with external systems

The exact type of stored data is and the way how it is stored is not specified. It depends on methodology that is used in specified project, because each of them requires specific approach. In case of system that is based on Object-Oriented-Programming (OOP) the most important information are modules or packages of system and classes (with fields and methods definitions, interfaces, comments, etc.).

It is useful when architectural knowledge contains information not only from implementation phase of project but also from earlier phases (specification, analysis, design), for example class diagram, package diagram, requirements model, database schema, etc. This information should be integrated and linked with information from implementation parts (source codes).

IV. INTEGRATION FRAMEWORK DESIGN

Good architectural knowledge contains information from more than model type. To create integration framework which will provide such knowledge must be chosen supported models and their data sources and input formats. Knowledge extracted from these models then will be linked together into one integrated unit.

A. SUPPORTED MODELS

There are many models which can be used as data source for architectural knowledge. These models can be divided into two groups:

- Description models
- Executable models

Description models describe structure or behavior of system, but it cannot be directly transformed into executable format. Usually they are created by analytics or architects and they are used as specification for programmers in implementation phase.

There are several languages which suit to this definition. Probably the best known is Unified Modeling Language (UML), general-purpose graphic semiformal modeling language. It provides 13 types of structure, behavior and interaction diagrams [6]. The most used type of diagram is **class diagram**, so it is primary candidate for support in integration framework.

Executable models describe structure, behavior and interaction of system. They can be directly transformed into executable format (binaries, JAVA bytecode, MSIL, etc).

This group of models includes programming languages. They describe all important parts of software systems (structure, behavior, interaction) in formal way. Representatives of executable models are source codes written in JAVA, C#, C++, PHP and many more. It is very hard to support all of these languages in integration framework because all of them have different syntax, semantics and specific features.

There will be supported just two types of models in initial phase of integration framework development. Class diagram as description model and JAVA source codes as executable model.

B. DATA SOURCES

Effective knowledge-based system which will provide architectural knowledge about software project must have good knowledge base. Quality of that knowledge base depends not only on process of knowledge extraction but very important part is selection of data sources and input data formats.

Class diagrams are stored in CASE system's native formats or in XMI format. XMI (XML Metadata Interchange) was designed for transport of models between different CASE systems. It is more portable so it is better choice as input for integration framework.

For extraction of knowledge from source codes there can be used directly source codes, but there are two disadvantages. Larger software system contains a lot of source files and in source files there is a lot of information which is not needed in architectural knowledge. It should be useful to have some tool which will at first just extract important information in one integrated unit.

There are some tools which extract information from source codes and store it in some format. In 2010 at Department of Computers and Informatics was developed such tool as part of bachelors thesis [8] with focus to be used as part of integration framework. It extracts information from JAVA source codes and stores them as ontology in OWL format.

As data sources will be used **XMI** format for class diagram and **OWL** format for JAVA source codes.

C. EXTRACTION OF ARCHITECTURAL KNOWLEDGE'

Key part of integration framework is to extract partial knowledge from data sources and link elements in these partial knowledge bases. On Fig. 2 is displayed this process.

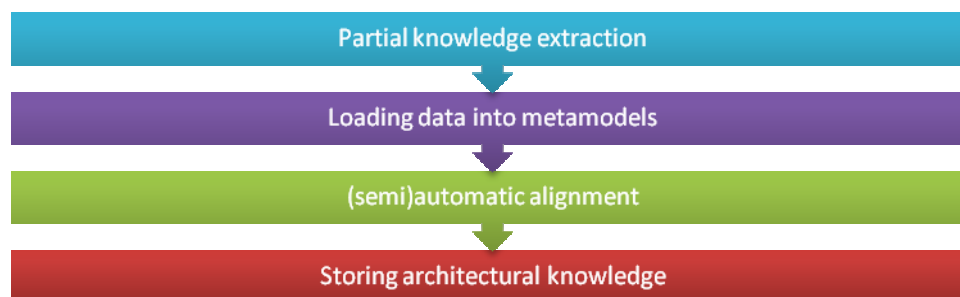


Fig. 2 Process of architectural knowledge extraction

At the beginning partial knowledge must be extracted using tool for getting OWL ontology from JAVA source codes. In the next step everything is loaded into metamodels (OWL ontology a class diagrams from XMI files). Metamodels are used to reduce dependency of alignment process on input formats. In future there can be added support for other formats just by creating other parser which will read input and fill metamodel without need to change alignment algorithm.

The most important part of whole process is alignment. It links classes in class diagrams with

corresponding JAVA classes. If this step will not work correctly, architectural knowledge provided by integration framework will not be proper and useless.

Alignment process must be semiautomatic just in some cases. Not every time is system implements exactly as it is designed. Second reason is developers don't observe naming conventions or packaging. The result is that classes in class diagram do not match implemented classes they have different names, methods, fields, they are in incorrect packages.

Algorithm of alignment uses all of this information to test whether two classes from different data sources match. Approximation strategy is not very appropriate because it can provide incorrect result leading to incorrect architectural knowledge. So if alignment algorithm cannot decide which classes match together it will use user help to finish its work.

Last step is storing architectural knowledge into permanent storage, so it will be accessible after restart of application immediately.

V. CONCLUSION

This paper describes vision and architecture of knowledge based system for software project's support. It explains architectural knowledge concept and presents its basic building parts. There is presented the design of integration framework which should provide architectural knowledge about specified software system using its UML models (class diagram) and source codes (JAVA).

The design of integration framework should be extended in future by adding support of other modeling and programming languages or by including knowledge extracted from binary files of software system.

ACKNOWLEDGMENT

This work was supported by VEGA Grant No.1/0198/11 Use of Knowledge in Software Processes and Architectures.

REFERENCES

- [1] Jetter, A. – Kraaijenbrink, J. – Schroder, H. H. – Wijnhoven, F.: Knowledge Integration. Aachen. A Springer company 2006. 208 s. ISBN 978-3-7908-1586-3
- [2] Adamuščinová, Iveta.: Znalosti, ich reprezentácia a využitie v životnom cykle a architektúrach softvérových systémov, KPI FEI TU Košice, 2008, 112 s.
- [3] Adamuščinová, I. – Révész, M, Havlice, Z.: Using Architectural Knowledge in Process of Software Maintenance, SAMI Proceedings, 2010, s. 83-88, ISBN: 978-1-4244-6423-4
- [4] Adamuščinová, I. – Kovács, A.N.: Architectural Knowledge and the Process of its Acquisition, 10th Scientific Conference of Young Researchers, 2010.
- [5] Keet, Maria C.: Aspects of Ontology Integration. Edinburgh: Napier University, School of Computing, 2004. 87 s.
- [6] Pender, Tom: UML Bible. Wiley Publishing, 2003. 984 s. ISBN 978-0-7645-2604-6
- [7] OWL 2 Web Ontology Language [online] <http://www.w3.org/TR/owl2-overview/>
- [8] BANDRÉ, Marek: Využitie prostredia Eclipse v znalostnom inžinierstve. Bachelor thesis. Košice: KPI FEI TU Košice, 2010. 40 s.

The Process of Expressing Web Services by Terms of Category Theory

Csaba SZABÓ, Viliam SLODIČÁK, Viliam MEDE

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

csaba.szabo@tuke.sk, viliam.slodicak@tuke.sk, viliam.mede@student.tuke.sk

Abstract – This paper contributes to the area of software process formalization and analysis by presenting a methodology for creation of a categorical expression of web services. The aim is, after introducing definitions of terms of category theory and web services, to show the steps of the process of category construction. The presented procedure uses web service implementation in Java language as its source. The output of the process is the category for web service syntax. The paper also shows the future directions of the actual research.

Keywords – category theory, Java, source code analysis, Web services

I. INTRODUCTION

The Web started out supporting human interactions with textual data and graphics. People use the Internet daily to buy consumer goods, and read the latest news. This level of interaction is fine for many purposes. Internet based applications need to be able to find, access, and automatically interact with other Internet-based applications. Web services improve Internet use by enabling program-to-program communication. Through the widespread adoption of Web services, applications at various Internet locations can be directly integrated and interconnected as if they were part of a single, large IT (information technology) system. Web services are Extensible Markup Language (XML) applications mapped to programs, objects, or databases or to comprehensive business functions. Using an XML document created in the form of a message, a program sends a request to a Web service across the network, and, optionally, receives a reply, also in the form of an XML document sent as a message. This technology can be used in many ways [1].

Nowadays, formalization of software engineering tasks and routines becomes emphasized due to the lack of complete proofs of the underlying processes. Software process improvement is the second area, where mathematical models gain on importance.

The goal of current research is to express web services by terms of category theory. The purpose of it is to achieve a highly expressive and precise model for analysis and improvement in the field. This improvement could be achieved by web service implementation automation based on formal definitions of expected behavior.

The organization of this paper is as follows. First, a gentle introduction into category theory is presented, then web services are analyzed. Section IV presents the methodology of creation of a category for web services representation. The last section concludes recent work and shows future work directions.

II. CATEGORY THEORY

Category theory is a mathematical tool for formal description of system behavior. Such a behavior is mainly deducted from discoverable state/output changes of the system or process being analyzed [2]. Formal definition of a category is given below.

Definition 1: A category [3], [4], [5], [6], [7] is a collection of *objects*, and for each pair of objects A, B , a set of $\text{hom}(A, B)$ of *morphisms* between them such that:

- 1) for each object X the set $\text{hom}(X, X)$ contains the *identity morphism*,
- 2) for each triple of objects X, Y, Z and pair of morphisms $f \in \text{hom}(X, Y)$ and $g \in \text{hom}(Y, Z)$, there is a *composite morphism* $(g \circ f) \in \text{hom}(X, Z)$,
- 3) for each pair of objects X, Y and morphism $f \in \text{hom}(X, Y)$, the identity morphisms are *left* and *right units* for composition: $\text{id}_Y \circ f = f = f \circ \text{id}_X$, and
- 4) for each 4-tuple of objects X, Y, Z, W and triple of morphisms $f \in \text{hom}(X, Y)$, $g \in \text{hom}(Y, Z)$ and $h \in \text{hom}(Z, W)$ composition is associative: $h \circ (g \circ f) = (h \circ g) \circ f$.

By other words, a category is a graph with a rule for composing arrows head to tail to give another arrow. This rule is subject to certain condition [5]. Morphism $f \in \text{hom}(X, Y)$ is also written $f : X \rightarrow Y$.

A *functor* is a structure-preserving map between categories. The formal definition is as follows.

Definition 2: A functor [3], [4], [5], [6], [8] $F : C \rightarrow D$ is a pair of functions $F_0 : C_0 \rightarrow D_0$ and $F_1 : C_1 \rightarrow D_1$ for which are the following conditions fulfilled:

- 1) if $f : X \rightarrow Y$ in C , then $F_1(f) : F_0(X) \rightarrow F_0(Y)$ in D ,
- 2) for any object A of category C , then $F_1(id_A) = id_{F_0(A)}$, and
- 3) if $g \circ f$ is defined in C , then $F_1(g) \circ F_1(f)$ is defined in D and $F_1(g \circ f) = F_1(g) \circ F_1(f)$.

A brief summary on known types of categories and their advantages follows. Cartesian closed category and recursive category [5] allow modeling of concurrent, selection and loop structures respectively. Quotient category and slice category can be used to model refining. Constructive theory and free category can be used to check complete and reachable, etc., and generate nodes and edges coverage test codes [3], [4], [8], [9].

III. WEB SERVICES

This section presents web service characteristics as described in [1] and [10]. The presented source code example deals with a simple problem of addition operation of a calculator.

A. Introducing

Web services are Extensible Markup Language (XML) applications mapped to programs, objects, or databases or to comprehensive business functions. Web services can run on desktop and handheld clients to access such Internet applications as reservations systems and order-tracking systems. Web services can also be used for business-to-business (B2B) integration, connecting applications run by various organizations in the same supply chain. Web services can also solve the broader problem of enterprise application integration (EAI).

Programs that interact with one another over the Web must be able to find one another, discover information allowing them to interconnect, figure out what the expected interaction patterns are – a simple request/reply or more complicated process flow – and negotiate such qualities of service as security, reliable messaging, and transactional composition. Some of these qualities of service are covered in existing technologies and proposed standards, but others are not.

B. Technologies

Web services require several related XML-based technologies to transport and to transform data into and out of programs and databases.

1) *XML:* XML (Extensible Markup Language), the basic foundation on which Web services are built provides a base language for defining data and how to process it. XML represents a family of related specifications published and maintained by the World Wide Web Consortium (W3C) and others. XML is a family of technologies: a data markup language, various content models, a linking model, a namespace model, and various transformation mechanisms.

2) *WSDL:* WSDL (Web Services Description Language), an XML based technology, defines Web services interfaces, data and message types, interaction patterns, and protocol mappings. WSDL is at the heart of the Web services framework, providing a common way in which to represent the data types passed in messages, the operations to be performed on the messages, and the mapping of the messages onto network transports. WSDL is divided into three major elements:

- Data type definitions
- Abstract definitions
- Service bindings

Each major element can be specified in a separate XML document fragment and imported in various combinations to create a final Web services description, or they can all be defined together in a single document.

3) *SOAP:* SOAP (Simple Object Access Protocol), a collection of XML-based technologies, defines an envelope for Web services communication–mappable to HTTP and other transports–and provides a serialization format for transmitting XML documents over a network and a convention for representing RPC interactions. SOAP is fundamentally a one-way communication model that ensures that a coherent message is transferred from sender to receiver, potentially including intermediaries that can process part of or add to the message unit. SOAP has several main parts:

- Envelope: Defines the start and the end of the message

- Header: Contains any optional attributes of the message used in processing the message, either at an intermediary point or at the ultimate end point
- Body: Contains the XML data comprising the message being sent
- Attachment: Consists of one or more documents attached to the main message (SOAP with Attachments only)
- RPC interaction: Defines how to model RPC-style interactions with SOAP
- Encoding: Defines how to represent simple and complex data being transmitted in the message

4) **UDDI**: UDDI (Universal Description, Discovery, and Integration), a Web services registry and discovery mechanism, is used for storing and categorizing business information and for retrieving pointers to Web service interfaces. The UDDI framework defines a data model in XML and SOAP application programming interfaces (APIs) for registering and discovering business information, including the Web services a business publishes. Businesses use SOAP to register themselves or others with UDDI; then the registry clients use the query APIs to search registered information to discover a trading partner. An initial query may return several matches from which a single entry is chosen. Once a specific business is chosen, a final API call is made to obtain the specific contact information for a business.

The basic Web services standards are used together. Once the WSDL is obtained from the UDDI or other location, a SOAP message is generated for transmission to the remote site.

C. WS examples

An example web service code for a simple calculator supporting only one operation (i.e. addition) is given below.

Listing 1 Calculator example – Java

```
1 package org.me.calculator;  
2  
3 import javax.jws.WebMethod;  
4 import javax.jws.WebParam;  
5 import javax.jws.WebService;  
6  
7 @WebService()  
8 public class CalculatorWS {  
9     @WebMethod(operationName="add")  
10    public int add(@WebParam(name="i") int i, @WebParam(name="j") int j)  
11    {  
12        return i+j;  
13    }  
14 }
```

The WSDL definition of the same web service is shown on Listing 2.

Listing 2 Calculator example – WSDL

```
1 <?xml version="1.0" encoding="UTF-8"?>  
2 <definitions xmlns:wsu="http://docs.oasis-open.org/wss/2004/01/oasis-  
3 -200401-wss-wssecurity-utility-1.0.xsd" xmlns:wsp="http://schemas.  
4 xmlsoap.org/ws/2004/09/policy" xmlns:soap="http://schemas.xmlsoap.  
5 org/wsdl/soap/" xmlns:tns="http://calculator.me.org/" xmlns:xsd="  
6 http://www.w3.org/2001/XMLSchema" xmlns="http://schemas.xmlsoap.org  
7 /wsdl/" targetNamespace="http://calculator.me.org/" name="  
8 CalculatorWSService">  
9 <types>  
10 <xsd:schema>  
11 <xsd:import namespace="http://calculator.me.org/" schemaLocation="http:  
12 //localhost:8080/CalculatorApp/CalculatorWSService?xsd=1"></  
13 xsd:import>  
14 </xsd:schema>  
15 </types>  
16 <message name="add">  
17 <part name="parameters" element="tns:add"></part>  
18 </message>  
19 <message name="addResponse">  
20 <part name="parameters" element="tns:addResponse"></part>  
21 </message>  
22 <portType name="CalculatorWS">  
23 <operation name="add">  
24 <input message="tns:add"></input>
```

```

17 <output message="tns:addResponse"></output>
18 </operation>
19 </portType>
20 <binding name="CalculatorWSPortBinding" type="tns:CalculatorWS">
21 <soap:binding transport="http://schemas.xmlsoap.org/soap/http" style="
    document"></soap:binding>
22 <operation name="add">
23 <soap:operation soapAction=""></soap:operation>
24 <input>
25 <soap:body use="literal"></soap:body>
26 </input>
27 <output>
28 <soap:body use="literal"></soap:body>
29 </output>
30 </operation>
31 </binding>
32 <service name="CalculatorWSService">
33 <port name="CalculatorWSPort" binding="tns:CalculatorWSPortBinding">
34 <soap:address location="http://localhost:8080/CalculatorApp/
    CalculatorWSService"></soap:address>
35 </port>
36 </service>
37 </definitions>
    
```

IV. THE CATEGORY CREATION PROCESS

A. Related work

In paper [2] several approaches on categorical description of web services were presented. Authors of [8] define the category for web services as follows.

Definition 3: A Web service is a graph $WS = (Sts, Ops, source, target)$ such that:

- 1) $Sts = \{s_1, s_2\}$, called state set, is corresponding to the object set of category,
- 2) $Ops = \{ws\}$, called operation set, is corresponding to the arrow set of category,
- 3) $source(ws) = s_1, target(ws) = s_2$, called *source* and *target functions* of an arrow respectively.

Conclusion from such a web service definition in [8] is that web service composition is a complex path of a category. The path with length equal to 1 or 0 is called an atomic service or empty service respectively and could be considered as identity morphisms on the web service states. Web service composition process is just one of looking for a path from start state to target state.

The approach presented above is very close to the category for component based systems composition that uses components $comp$ as objects and the morphism $conn : comp \rightarrow comp$ for their connection, besides the identity morphism on components. This category gives the basic idea for the WS categories shown below. WS_1 represents component-based construction; WS_2 is based on web service definition based WS construction.

Component-based WS construction category [2] WS_1 uses ws as objects. Morphisms of the WS_1 category are as follow.

$$\begin{aligned}
 id : ws &\rightarrow ws, \\
 conn : ws &\rightarrow ws,
 \end{aligned} \tag{1}$$

where $conn$ is the morphism for connecting two web services to a more complex whole.

Web service definition version uses three main objects that find their meaning in the WS_2 category too: $wsdef, wserv, wsmet$. These objects correspond to WS definition, service and method respectively. Morphisms follow.

$$\begin{aligned}
 id_{wsdef} : wsdef &\rightarrow wsdef, \\
 id_{wserv} : wserv &\rightarrow wserv, \\
 id_{wsmet} : wsmet &\rightarrow wsmet, \\
 incl : wserv &\rightarrow wsmet, \\
 impl : wsdef &\rightarrow wserv, \\
 (incl \circ impl) : wsdef &\rightarrow wsmet,
 \end{aligned} \tag{2}$$

where identity morphisms are obvious, $incl$ identifies method inclusion in service implementations, $impl$ shows the implementation to the definition of WS. Composition presents the relation from WS definition to method implementation.

B. Process proposal

Given the definitions and examples presented above, it is possible to propose a strategy for categorical description of web service syntax and semantics. The proposed strategy follows.

Proposal 1: For a given web service definition in Java and/or WSDL, the process of WS category creation consists of:

- 1) analysis of WS syntax, and
- 2) analysis of WS semantics.

For both types of analysis, the following attributes need to be analyzed:

- 1) services definition,
- 2) methods definition,
- 3) method parameters.

This analysis produces *object* and *morphism candidates* for the WS category that need to be selected.

For syntax expression, a single category is required. To express semantics, a group of categories with non-degenerative functors will be required.

C. The calculator example

The Java code analysis using the process proposal shown above leads to the following results.

1) *Semantics:* At semantic level, web service UDDI universal identifier, WS name, the add method, and its two parameters were identified. Each with the corresponding type constraint.

2) *Syntax:* At syntactical level, three objects were identified: *webservice*, *webmethod*, and *webparam*. The syntactical rules expressed by a basic category follow.

$$\begin{aligned} g &: \text{webmethod} \rightarrow \text{webservice}, \\ f &: \text{webparam} \rightarrow \text{webmethod}, \\ (g \circ f) &: \text{webparam} \rightarrow \text{webservice}, \end{aligned} \quad (3)$$

where f represents the relation "needs set as parameter", and g represents relation "service provides operation."

These facts could be derived from the WSDL definition as well, but this one is less readable than the Java source code. The functor for semantics expression is subject of investigation.

V. CONCLUSION

This paper presented a process proposal for web service explanation by category theory. The proposal is based on observational analysis of web service definitions for syntactical and semantic structures.

There were web services described in detail. Focus was on technological implementation and standards in WS definition. The Java alternative of definition was also discussed. The conclusion from this analysis is that both approaches could be used for WS category creation.

The presented procedure uses web service implementation in Java language as its source. The output of the process is the category for web service syntax.

The proposed process is incomplete, it needs proofs and improvements, but even this early version has its benefits presented in the examples. Process correctness and efficiency needs to be investigated. It also needs to be generalized for use with more complicated WS such as ones using huge databases, or ones that are created as web service compositions.

Secure web services have a kind different definition. This type of WS was not considered during related work. It is also the subject of future work in the area of WS category theory description.

ACKNOWLEDGMENT

This work is the result of the project implementation: Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] Cs. Szabó and V. Mede, "Web services for content management systems," *Electrical Engineering and Informatics, Proceeding of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice*, vol. 1, pp. 342–346, 2010.
- [2] Cs. Szabó and V. Slodičák, "Software engineering tasks instrumentation by category theory," in *SAMI 2011, Proceedings of the 9th IEEE International Symposium on Applied Machine Intelligence and Informatics*. Smolenice, Slovakia: elfa, s.r.o., January 27–29, 2011, pp. 195–199.
- [3] V. Novitzká, D. Mihályi, and A. Verbová, "Coalgebras as models of system's behaviour," in *Proc. of AEI 2008, Int. Conf. on Applied Electrical Engineering and Informatics 2008*, Athens, Greece, 2008, pp. 31–36.

- [4] D. Mihályi and V. Novitzká, "A coalgebra as an intrusion detection system," *Acta Polytechnica Hungarica*, vol. 7, no. 2, pp. 71–79, 2010.
- [5] V. Slodičák and Cs. Szabó, "Recursive coalgebras in mathematical theory of programming," *Abstract of the 8th Joint Conference on Mathematics and Computer Science, MACS 2010*, July 14-17, 2010.
- [6] J. A. Goguen, "A categorical manifesto," *Math. Structures in Comp. Sci.*, vol. 1, no. 1, 1991.
- [7] M. Stay, "Category theory for the java programmer," Nov. 3 2007, <http://reperiendi.wordpress.com/2007/11/03/category-theory-for-the-java-programmer/>.
- [8] X. Zhang, H. Miao, and H. Zeng, "The syntactic and semantic model of web services composition based category," in *Proc. of Int. Conf. on Advanced Comp. Theory and Engineering (ICACTE'08)*, Phuket, Thailand, 2008, pp. 444–449.
- [9] J. Baez and M. Stay, "Physics, topology, logic, and computation: A rosetta stone," *New Structures for Physics, Lecture Notes in Physics 813*, pp. 95–172.
- [10] "Web Services Description Language (WSDL) 1.1," March 15, 2001, <http://www.w3.org/TR/wsdl>.

Action Graphs and Their Properties

Martina ĽALOŤOVÁ, Valerie NOVITZKÁ

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

martina.lalova@tuke.sk, valerie.novitzka@tuke.sk

Abstract – In our paper we clarify the definition of action graphs for a dynamic system, which we explain on a specific example. We focus on some properties of action graphs described by groups. Groups can describe the structure of dynamical systems and compare their actions with action graphs. We discuss their power and weaknesses and propose some opportunities for their use.

Keywords – dynamic systems, π -calculus, action graphs

I. INTRODUCTION

Our research deals with dynamics of program systems. There are many definitions and descriptions of dynamic systems. One of the useful structure representing dynamic systems are action graphs. From our perspective, we can take advantage of the characteristics of this technique to describe the behavior of the system. An action graph is a structure that characterizes the behavior of the dynamic system. Actions can be grouped to sets and such complex action graphs are useful for describing properties of complex dynamic systems.

II. PROCESSES IN π -CALCULUS

One of the known formal method for describing processes is π -calculus. In this section we shortly characterize this method.

The π -calculus is a process calculus as a extension of the process calculus CCS (Calculus of Communicating Systems) [1]. It is able to describe concurrent computations whose configuration may change during the computation. The π -calculus is mathematical formalisms for describing and analyzing properties of concurrent computation and processes.

A *process* is running instance of a program, including all the values of variables and states. The process can have many actions. In our contribution we choose a process sending messages between nodes.

A *message* is defined as an entity designated by its name. π -calculus does not define the data types, or message content.

Messages are described by:

- their names,
- their ports (communication channels) and
- their behavior.

A *structure of a process* is defined by a port name designated by a list of ports specified in brackets after the name of the port and determined by the behaviour of the process, thus:

$$\text{Process name}(\text{list of port}) = \text{behavior}$$

A. Example:

Let A, B, C be the processes:

Then: x, y, z are ports, $\bar{y}$ is output port. Left side of the term represents a process determined by name and a list of ports. We have knowledge of behaviour of these ports.

$$\begin{aligned} A(x, y) &= \bar{y}\langle x \rangle. A'(y) \\ B(y) &= y(z). \bar{z}. B'(y, z) \\ C(x) & \end{aligned}$$

Communication of processes is represented as term, where

- Behavior of $A(x, y)$ - output prefixing $\bar{y}\langle x \rangle. A'(y)$ of this process describes that the name y is emitted on channel x before proceeding as A .

- $B(y)$ with input prefixing $y(z)$. $\bar{z}. B'(y, z)$ of this process is waiting for a message that was sent on a communication channel named y before proceeding as B , binding the name received to the name z .
- $C(x)$ is process C waiting for a message x .

◇

III. ACTION GRAPHS

π -calculus describes dynamic processes in the form of terms. We prefer graphical representation of dynamic processes and action graphs are the best solution. Every term in π -calculus can be translated to a fragment of action graph. Therefore we can say that these two methods are equivalent [2], [3].

Nodes of action graphs are placeholders for processes. Edges of action graphs express connections between processes. Action graphs can be used for recursion description of the behavior [4], [5]. We are able to represent recursive writing in node description, which can be a subsystem described by another action graph. Graph inside the node is an abstraction of internal parameters of the system when node terminates by the way using abstraction.

Certainly, connections can be composed, i.e. we can express a connection of node V_i to V_k through V_j .

We can use action graphs to describe different types of interactions between processes.

Definition of action graph is based on the graph theory [6] and provides a description of the structure of a system and its objects and relations, which allows transparent graphical representation.

An *action graph* is a sextuple

$$(V, H, S, T, L, R),$$

where

- V is a set of *nodes*, representative of processes;
- H is a set of *edges*, representative of connections;
- S is *initial node*, representative of process initialing to make a connection;
- T is a set of *final nodes*, representative of processes, to which they connect;
- L is a set of *labels* that are associated to edges depended on source and target node.
- R is a set of *reaction rules*.

Label assigns to a source node V_i the associated label of edge.

A function Label assigns a description as a pseudo-code, to each edge, function and also a node.

Each node of graph also represents the 'region', each edge represents the function for interaction.

Nodes of action graph are elements of the set N with neighboring defined by actions of elements from S on the set N . We consider continuous action graph, which corresponds with Schreier coset [7].

By comparing group of actions with morphisms, we are able to investigate the relationship between them.

Concepts of action graph we illustrate on the example in section IV.

IV. EXAMPLE

We construct an action graph from example II-A described by π -calculus [8], [9]. We name this action graph as P . Action graph consists of:

- initial node S , which is initial point and also the point of contact with environment,
- set of nodes V as a representative of processes with two elements, and it is nodes $\{V_a, V_b\}$,
- set of edges H with edges $\{a', b', c', d', e', f', g', h'\}$ the duplication of edges had been created to make alternatives more transparent,
- single element set of final nodes T which have a connection with a single element of V_c .
- set of signification L of the function Label denoting the nodes and
- set of reaction rules R .

Action graph has to represent the name of the sent of message from initial node, which is connected with environment, represented by the node V_c .

In system P we define interface ε , context as an action.

Links l, k, r, p and m describe the mobility of the system P illustrated schematically in figure Fig.1. Edges $a', b', c', \dots$ from the set of edges H create a context of node [6], [10]. In our case, they

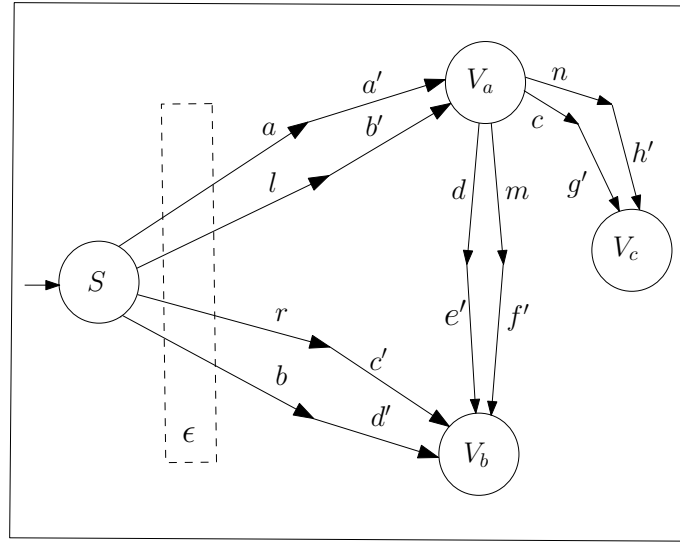


Fig. 1 Action graph of system P

also represent response of a node, to which a connection is made. We describe the action only in one way, a way of receiving the message to node V_c .

Context with domain ε is composed of all used ports, where a, b, d, c are actions of nodes S, V_a, V_b and V_c . Pairs (l, r) and (m, n) are the reaction rules from R , a set of reaction rules - the reaction relation from the initial node S to the final node. For the relation of initial node S we define as:

$$\longrightarrow \stackrel{def}{=} \{(b' \circ l, c' \circ r)(l, r) \in R\}$$

As we can see, there are two possible sessions from the beginning of initial node S , relation l for node V_a and relation r for node V_b . If the name required from node V_b does not come, action b does not start. This choice is repeated in node V_a too, where the chosen relation is n and action c . Message delivered to node V_c had been sent from node S across node V_a .

We write the relation as $S \xrightarrow{b'} V_a \xrightarrow{def} b'S \rightarrow V_a$, where we use an action rule l of our system P .

Thus $S \xrightarrow{b'} V_a$ for pair $(S; V_a)$ is relation l .

Consequently we can define the relation $V_a \xrightarrow{h'} V_c \xrightarrow{def} h'V_a \rightarrow V_c$, where we use action rule n .

Similarly $V_a \xrightarrow{h'} V_c$ is a relation for pair $(V_a; V_c)$ using n .

Message was received by $S \xrightarrow{b'} V_a \xrightarrow{h'} V_c$, by composition of actions $a \circ c$ and therefore we have to choose edges a', g' .

In system P we define the interface of individual nodes, for example ϵ of node S , which is "region" for action a and b . Initial node S is a root of system P , $S \equiv \text{root}(P)$. Marking of the border around the image symbolizes the closeness of system with respect to the environment.

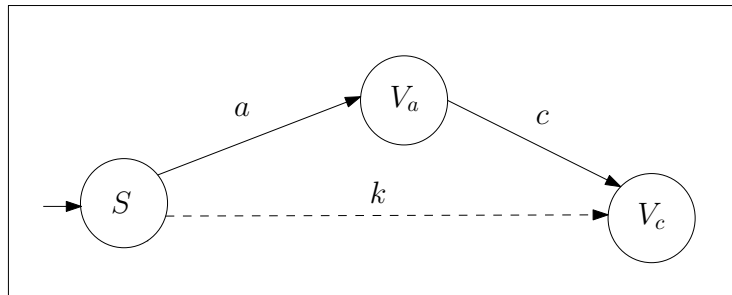


Fig. 2 Simplified action graph of system P

On the Figure 2 we can see message being sent from the beginning of initial node S to the final node V_c .

$$S \longrightarrow V_c \stackrel{def}{=} k : (a \circ c) \quad (1)$$

◇

V. GROUP OF ACTION

We use grouping of attributes described on a mathematical base to describe behaviour of a system and relationship between them. Definitions of single terms of this description are needed.

An action is considered as an act, step execution or a block of program.

An action is considered a minimal and essential element of constructed action graph.

A. Set of action

Let Z be an non-empty set, where its elements are actions. We call Z a set an *set of actions*.

Group $(Z, \circ)$ is a pairs, where Z is non-empty set and $\circ$ is a binary operation on this set, while

- operation $\circ$ is associative on Z , i.e.

$$\forall a, b, c \in G : a \circ (b \circ c) = (a \circ b) \circ c,$$

- there exist a neutral element $e \in G$ with property

$$\forall a \in G : a \circ e = a = e \circ a,$$

- to each element $a \in G$ exists an inverse element $b \in G$ such that

$$a \circ b = e = b \circ a.$$

Symbol $\circ$ interprets an operation of composition. For better understanding of the links and relations in the structure we need to know the properties of individual morphisms.

Morphisms of actions are ordered pairs

$$(\phi, \psi) : (Z, G) \rightarrow (Z', G'), \text{ where}$$

$$\phi : Z \rightarrow Z' \text{ is a function and } \psi : G \rightarrow G' \text{ is a morphism}$$

such that,

$$\phi(u \circ g) = \phi(u) \circ \psi(g).$$

The compositon of morphisms is right associative.

Morphisms in the form of

$$(\phi, id) : (Z, G) \rightarrow (Z', G) \quad (2)$$

are called *equivariants* of actions and morphisms in the form of

$$(id, \psi) : (Z, G) \rightarrow (Z, G') \quad (3)$$

are *invariants* of morphisms of actions.

Invariant morphisms represent an intuitive description of groups, which have the same behaviour on a given set. This completes the basic types of relations between elements of a structure. By investigating other attributes, the possibilities of coverings of actions and morphisms can be used, and also similarity of action graphs and bringing-together classes [7].

VI. COVERING

Property of function or action covering ensures that, the result will be again a function or an action (we can a be grouped objects with this property and use them as a single structure).

We say that an action $(\tilde{Z}, \tilde{G})$ *covers an action* (Z, G) always if a morphism exists

$$(\phi, q) : (\tilde{Z}, \tilde{G}) \rightarrow (Z, G). \quad (4)$$

It can be split to two morphism:

- 1) *equivariants covering*

$$(\phi, id) : (\tilde{Z}, \tilde{G}) \rightarrow (Z, \tilde{G}) \quad (5)$$

- 2) *invariant covering*

$$(id, q) : (Z, \tilde{G}) \rightarrow (Z, G), \quad (6)$$

where action $\tilde{G}$ on Z , is defined by

$$z \circ \tilde{g} = z \circ q(\tilde{g}), \quad (7)$$

and $z \in Z$ a $\tilde{g} \in \tilde{G}$.

A morphism $f : A \rightarrow B$ is *epimorphism*, if for any structure G and any morphisms

$$g, h : B \rightarrow C \quad (8)$$

if

$$f \circ g = f \circ h, \text{ is valid} \quad (9)$$

then $g = h$.

Application of these assumptions to example IV we get:

$$G = \{a, b, c, d, e, f, \dots, z\}$$

$$Z = \{a, b, c, d, k\},$$

where

$$Z \subseteq G \quad \text{and} \quad q : \tilde{G} \rightarrow G$$

The action described in (1) is

$$k = (a \circ c)$$

and

$$(id, q) : (k, \tilde{G}) \rightarrow ((a \circ c), G)$$

is invariant covering using equation (6). Equality

$$(k \circ \tilde{g}) = (k \circ q(\tilde{g}))$$

follows from equation (9), where

$$k \in Z \quad \tilde{g} \in \tilde{G} \quad q(\tilde{g}) \in G.$$

We can use all these properties for simplification of action graph description in such way that we preserve all properties of the system that are interesting for us. This approach allows us to create a more clear action graph.

VII. CONCLUSION

We have shown action graphs that enable graphical description of complex dynamic systems. Also we have outlined that usage of mathematical structures of groups can lead to a more simple and transparent action graphs.

The beauty and diversity of action graphs wells within a fact that, it is past use had been bounded to bio-chemistry, economy and other science disciplines. Their usability in informatics, is not given by a standard frame.

We follow our research with investigating more useful properties of action graphs and working out a method for their construction in various binds to program systems.

ACKNOWLEDGMENT

This work was supported by the project VEGA No. 1/0015/10: Principles and methods of semantic enrichment and adaptation of knowledge-based languages for automatic software development.

REFERENCES

- [1] D. Sangiorgi and D. Walker, *PI-Calculus: A Theory of Mobile Processes*. New York, NY, USA: Cambridge University Press, 2001.
- [2] M. Ľalová, "Akčné grafy pre popis dynamických systémov," 2010.
- [3] C. Szabó and V. Slodičák, "Software engineering tasks instrumentation by category theory," pp. 195–199, 2011.
- [4] M. Ľalová and V. Slodičák, "Some useful structures for categorical approach for program behavior," 2010.
- [5] M. Ľalová, V. Novitzká, and V. Slodičák, "Algebras on the duality principle for program behavior," in *Journal of Information, Control and Management Systems*, 9, 1, 2010, 2010, pp. 61–70.
- [6] A. Phillips, L. Cardelli, and G. Castagna, "A graphical representation for biological processes in the stochastic pi-calculus," *Transactions on Computational Systems Biology*, 2006, to appear.
- [7] A. Malnič, *ACTION GRAPHS AND COVERINGS*, 2000.
- [8] G. L. Cattani, J. J. Leifer, and R. Milner, "Contexts and embeddings for closed shallow action graphs," Tech. Rep., 2000.
- [9] M. Ľalová, "Dynamic systems and their description: Calculus vs. action graphs," in *SCYR 2010 -10th Scientific Conference of Young Researchers FEI TU of Košice*, 2010, p. 270–274.
- [10] B. König, "Description and verification of mobile processes with graph rewriting techniques."

The Resource Character of Linear Logic

Pavol MACKO, Valerie NOVITZKÁ

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

pavol.macko@tuke.sk, valerie.novitzka@tuke.sk

Abstract – Linear logic is logic that can describe the dynamic processes of real world by using resources. This aspect is usable in programming to describe systems that are changing over time. Thus linear logic is very suitable formalism serving as an interface between logic and computer science. This document describes how to introduce time and space as resources into linear logic.

Keywords – linear logic, time and space, resource-oriented programming

I. INTRODUCTION

Linear Logic (LL) was formulated by J.-Y. Girard in 1987 and published in his paper [1]. This theme is still actual and it makes sense to consider it, because it brings more and more new results that are beneficial for the computer sense. Linear logic provides a logical perspective on computational issues such as control of resources. It is able to describe systems that are changed during they are used. Applications of linear logic cover the theory of concurrent processes, logic programming and functional programming. Using Curry-Howard correspondence, propositions of linear logic are interpreted as types [2].

Linear logic is logic that can describe a dynamics of real world processes and therefore it can describe the dynamics in programming too, by explicit handling with time and space.

II. BASIC CONCEPTS OF LINEAR LOGIC

Linear logic is a extension of classical and intuitionistic logic. Instead of emphasizing truth, as in classical logic, or proof, as in intuitionistic logic, linear logic emphasizes the role of formulas as resources, so each *formula* of linear logic is an *action*, or *resource* [1], [3]. It is logic that can describe real world processes. Another innovation that characterizes the linear logic are new linear connectives, which we describe in the next section.

Intuitionistic logic in comparison with linear logic does not know the resource term, because the formula in intuitionistic logic is static, its state is fixed [2]. By contrast, in linear logic formulas are actions or resources, which are performed or consumed, resp. thus vary in time [4]. This property is reflected by the rules of weakening and contraction, which will be explained below. Resource-oriented features of the linear logic reject these rules and the resources consider like as in real world, so after their using are consumed (modified).

On the basis that linear logic is working with resources, it has been founded many applications in computer science. For example, by using this logic we can describe systems that are changing over time.

A. Linear logic vs classical and intuitionistic logics

Linear logic differs from classical and intuitionistic logic in several fundamental ways. Classical logic may be viewed as if it deals with static propositions about the world where each proposition is either true or false. Because of the static nature of propositions in classical logic, one may "duplicate" propositions: P implies $(P \text{ and } P)$. Implicitly we have learned that "one P is as good as $twoP$ ". Also, the propositions: $(P \text{ and } Q)$ implies P is true. Here the proposition Q has been "thrown away". Both of these sentences are valid in classical logic for any P and Q .

In linear logic, these sentences are not valid. A linear logician might ask "Where did the second P come from?" and "Where did the Q go?" Of course these questions are unnecessary in the classical logic, since propositions are assumed to be static, unchanging facts about the world.

Linear propositions has dynamic properties of finite resources [5].

For example, consider the propositions A , C , E , conceived as resources:

$$\begin{aligned} A &= \text{a pack of chips} \\ C &= \text{a pack of candies} \\ E &= \text{one euro} \end{aligned}$$

Consider the following axiomatization of a vending machine:

$$\begin{aligned} E &\text{ implies } A \\ E &\text{ implies } C \end{aligned}$$

Then in classical (or intuitionistic) logic, one is able to deduce

$$E \text{ implies } (A \text{ and } C),$$

which may be read as "With one euro, I may buy both a pack of chips and a pack of candies". Although this deduction is valid in classical logic, it is nonsense in the intended interpretation of proposition as resources: one cannot buy two packs of wares with one euro from the vending machine described.

B. Connectives and constants

Linear logic introduces new logical connectives. One of the most immediately striking thing about linear logics is that it has two forms of conjunction ($\otimes$, $\&$) and two forms of disjunction ($\wp$, $\oplus$). To go with these it also has two forms of true ($\top$, 1) and two forms of false ($\perp$, 0) [1], [3].

The linear negation is the one negative operation of linear logic. It is an unary operation. Linear negation of a formula A is marked as $A^\perp$ and it is read *nil* A . It expresses the consuming of resource A . Also, we can express by negation the duality of status change.

$$\text{action of type } A = \text{reaction of type } A^\perp$$

An important property of linear negation is the rule of double negation (involution negation), which looks like in classical logic: $A^{\perp\perp} \equiv A$.

In linear logic are valid the following De Morgan laws [?]:

$$\begin{aligned} (A \otimes B)^\perp &= A^\perp \wp B^\perp & (A \wp B)^\perp &= A^\perp \otimes B^\perp \\ (A \& B)^\perp &= A^\perp \oplus B^\perp & (A \oplus B)^\perp &= A^\perp \& B^\perp \end{aligned}$$

This formula $A \multimap B$ of linear logic means that it consumes A resource to produce B resource. Thus, formula A is not following its use in linear implications, because the resource A is consumed (like in real world).

The multiplicative conjunction, written $A \otimes B$, expresses that both actions A and B will be done. The additive conjunction, written $A \& B$, expresses that only one of these actions will be performed and we can choose which one. It is called external nondeterministic. Dually, there are two disjunctions. The multiplicative disjunction $A \wp B$. This is one of the connectives that is hardest to explain. One way of looking at it is merely to note that $A \multimap B$ can be defined as $A^\perp \wp B$. That is either you have something that is looking to consume an A resource, or you produce a B resource. We can try to paraphrase A as "if action A isn't executing, then action B is executing and vice versa." The additive disjunction $A \oplus B$ stands for the possibility of either A or B , but we do not know which. This is also known as internal nondeterminism. For each connective exists a neutral element, unit 1 , impossibility 0 , top $\top$, bottom $\perp$ and the following rules stands for them:

$$(A \otimes 1) \equiv A \quad (A \oplus 0) \equiv A \quad (A \& \top) \equiv A \quad (A \wp \perp) \equiv A \quad 0 \equiv \top^\perp \quad 1^\perp \equiv \perp$$

When we summarize the facts, a formula A can be written as:

$$A ::= 0 \mid 1 \mid \perp \mid \top \mid A_1 \multimap A_2 \mid A_1 \otimes A_2 \mid A_1 \& A_2 \mid A_1 \wp A_2 \mid A_1 \oplus A_2 \mid A^\perp$$

C. Exponentials

Girard introduced [1] into linear logic the modal operators. These operators provide multiple use of resources, so we can overwrite a formula of classical logic into the form of linear logic. The modal operators are called the exponentials. They express that we can use a resource repeatedly, as many times as we need. The exponentials have two shapes:

Of course: !

The formula $!A$ means that resource A is always available, so this source is infinite and indicates the potentiality (multiple using of source).

Why not: ?

It is used in some deduction rules for updating the potentials.

Example:

a) a linear implication

$$!A \multimap B$$

says that A has before and after the implication the same truth value. The implication is the transcription of the implications of classical (predicate) logic to linear logic using exponential.

b) To illustrate the use of linear connectives and modal operators we introduce the following example. Suppose for a fixed price 5 a restaurant provides a menu:

Fixed-price menu: 5
 Hamburger
 Coke
 All french fries you can eat
 Onion soup or salad
 Pie or icecream (depending on availability)

Now we can encode this information in the linear logic formula:

$$A = (E \otimes E \otimes E \otimes E \otimes E) \multimap [H \otimes C \otimes !F \otimes (O \& S) \otimes (P \oplus I)]$$

D. Weakening vs contraction

Exponential operator $!$ allows using of restricted form of weakening and contraction. [3] These rules may be used only for modal formulas.

Weakening rule:

$$\frac{\Gamma \vdash B}{\Gamma, !A \vdash B}$$

It allows to extend the context of sequence only by modal formula.

Contraction rule:

$$\frac{\Gamma, !A, !A \vdash B}{\Gamma, !A \vdash B}$$

It allows to exclude the one occurrence of of two same modal formulas from the context.

III. THE SEQUENT OF LINEAR LOGIC

The sequent is a base element of linear deductive system. It is written in the Gentzen's style and in the linear logic it has the following form:

$$\Gamma \vdash A$$

where Γ is the final list $B_1, \dots, B_n$ of a linear formulas and A is also a linear formula. Γ is called context and it consists of assumptions for the derivation of formula A [1], [3]. A context consists of assumptions needed to derive a formula A .

A proof rule sequent calculus consists of a set of hypothesis a horizontal line and a single conclusion below the line, as it is seen in follow:

$$\frac{\text{hypothesis1} \quad \text{hypothesis2}}{\text{conclusion}}$$

Deductive system of linear logic consists of basic and structural rules [6], together with rules for the introduction of linear connectives listed in the previous section.

Basic rule is only one, the axiom of identity

$$\frac{}{A \vdash A} \text{ (id)}$$

Structural rules are two, the cut rule and the exchange rule.

$$\frac{\Gamma, A \vdash B \quad \Delta \vdash A}{\Gamma, \Delta \vdash B} \text{ (cut)} \quad \frac{\Gamma, A, B, \Delta \vdash B}{\Gamma, B, A, \Delta \vdash B} \text{ (exchange)}$$

The cut rule says that if from the context Γ , A is possible to derive a formula B and from the context Δ is possible to derive a formula A , then formula A can be excluded and B is derivable from the context Γ, Δ .

The exchange rule says that the order of a linear formulas in the context is not important.

Logical rules

The sequent calculus of linear logic contains deduction rules, some of which are given.

$$\begin{array}{ccc} \frac{\Gamma_1 \vdash A_1 \quad \Gamma_2, A_2 \vdash B}{\Gamma_1, A_1 \multimap A_2, \Gamma_2 \vdash B} (\multimap_{-l}) & \frac{\Gamma A_1 \vdash A_2}{\Gamma \vdash A_1 \multimap A_2} (\multimap_{-r}) & \\ \\ \frac{\Gamma, A_1, A_2 \vdash B}{\Gamma, A_1 \otimes A_2 \vdash B} (\otimes_{-l}) & \frac{\Gamma_1 \vdash A_1 \quad \Gamma_2 \vdash A_2}{\Gamma_1, \Gamma_2 \vdash A_1 \otimes A_2} (\otimes_{-r}) & \frac{\Gamma \vdash A, B}{\Gamma \vdash A \wp B} (\wp_{-r}) \\ \\ \frac{\Gamma, A_1 \vdash B}{\Gamma, A_1 \& A_2 \vdash B} (\&_{-l_1}) & \frac{\Gamma, A_2 \vdash B}{\Gamma, A_1 \& A_2 \vdash B} (\&_{-l_2}) & \frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \& B} (\&_{-r}) \\ \\ \frac{\Gamma, A_1 \vdash B \quad \Gamma, A_2 \vdash B}{\Gamma, A_1 \oplus A_2 \vdash B} (\oplus_{-l}) & \frac{\Gamma \vdash A}{\Gamma \vdash A \oplus B} (\oplus_{-r_1}) & \frac{\Gamma \vdash B}{\Gamma \vdash A \oplus B} (\oplus_{-r_2}) \end{array}$$

IV. RESOURCE IN LL

Resources in the LL can describe dynamic processes in the real world. We will be more concrete and we take into account the domain of programming. From this perspective, we will consider the resources as time and space, because these characteristics are able to describe the character of the behavior of program in time with variables of certain type. Then we can describe the time and space (memory) of program systems in terms of linear logic.

A. Time in LL

Time into LL introduced Girard [7] by polarization of formulas. Girard says that the change of polarity of formula means the change (increment) of time. For simplicity, we use the term cluster, which marks the number of steps in the proof, which does not change polarity. Thus, each cluster means increment of time.

Polarity

Polarity of formulas are directly related to the connectives, which Girard divided into two classes [4]. Connectives of algebraic style he considered as *positive*: $\otimes$, $\oplus$, 1 , 0 ; and connectives of logical style considered as *negative*: $\&$, $\wp$, $\top$, $\perp$.

Formula of linear logic is positive if the external logical connective is positive. Formula of linear logic is negative if the external logical connective is negative.

Deductive rule of LL, which introduces a *negative* logical connectives is called *invertible deductive rule*.

Cluster of negative formulas is one that contains several (consecutive) steps of proof, which are resulting from the application invertible deductive rules.

Dual property to invertibility is *focalisation* [8], which introduces a *positive* connectives. Cluster of positive formulas is one that contains several (consecutive) steps of proof, which contains only positive formulas.

Each cluster includes the steps of proof, which may be performed simultaneously, i.e. they do not depend on time, because increment of time occurs in the transition to the next cluster. This property shows the use of linear logic in programming with parallel calculations.

The sequent of the shape: $\vdash \Gamma$ is called one-side sequent and we can say that all formulas of Γ are positive. But if have two-side sequent $\Gamma \vdash A, B, C$, which contains two positive formulas B, C and one negative formulas A , we can replace this sequent by the following shape: $A^\perp \vdash B, C$ with only positive formulas [9].

Example: Let $A = ((P^\perp \oplus Q^\perp) \otimes R^\perp) \& S^\perp$ be a LL formula, which it represents some process. Formulas P, Q, R, S are positive. We can construct the following three proofs of A . Proof can be constructed in several ways, we chose only one, the following:

$$\frac{\frac{\frac{\overline{P \vdash \Gamma}}{\vdash P^\perp, \Gamma} id}{\vdash (P^\perp \oplus Q^\perp), \Gamma} \perp_{-r} \quad \frac{\frac{\overline{R \vdash \Delta}}{\vdash R^\perp, \Delta} id}{\vdash S^\perp, \Gamma, \Delta} \perp_{-r} \quad \frac{\overline{S \vdash \Gamma, \Delta} id}{\vdash S^\perp, \Gamma, \Delta} \perp_{-r}}{\vdash ((P^\perp \oplus Q^\perp) \otimes R^\perp), \Gamma, \Delta} \otimes_{-r} \quad \frac{\vdash ((P^\perp \oplus Q^\perp) \otimes R^\perp), \Gamma, \Delta}{\vdash (((P^\perp \oplus Q^\perp) \otimes R^\perp) \& S^\perp), \Gamma, \Delta} \&_{-r}$$

It is seen that the polarity of formulas in the proof tree is changing. From negative to positive and then again to negative. These changes of polarity will be more noticeable, if we construct cluste:

$$\frac{\frac{P \vdash \Gamma}{\vdash (P^\perp \oplus Q^\perp), \Gamma} \quad R \vdash \Delta \quad S \vdash \Gamma, \Delta}{\vdash A, \Gamma, \Delta}$$

There each rule expresses time incrementation.

Using De Morgan rules we can write

$$A = ((P^\perp \oplus Q^\perp) \otimes R^\perp) \& S^\perp = ((P \& Q)^\perp \otimes R^\perp) \& S^\perp = ((P \& Q) \wp R)^\perp \& S^\perp = (((P \& Q) \wp R) \oplus S)^\perp$$

Then the proof tree is:

$$\frac{\frac{\frac{\overline{\vdash P, R, \Lambda} id \quad \overline{\vdash Q, R, \Lambda} id}{\vdash (P \& Q), R, \Lambda} \&_{-r}}{\vdash ((P \& Q) \wp R), \Lambda} \wp_{-r} \quad \frac{\vdash ((P \& Q) \wp R), \Lambda}{\vdash (((P \& Q) \wp R) \oplus S), \Lambda} \oplus_{-l}}{\vdash A \vdash \Lambda} \perp_{-l}$$

Formulas with same polarity we can close to cluster again:

$$\frac{\vdash P, R, \Lambda \quad \vdash Q, R, \Lambda}{\vdash (((P \& Q) \wp R) \oplus S), \Lambda} \quad \frac{\vdash P, R, \Lambda \quad \vdash Q, R, \Lambda}{\vdash A^\perp, \Lambda}}{A \vdash \Lambda} \equiv$$

B. Space in LL

Each program contains variables that have a certain type. Each type represents certain space in memory. Thus, for each variable, before running of program, is reserved certain space, depending on the type of variable. The space must have its place in the memory, must have its address. Let's think about this fact in terms of LL.

If we are dealing with the term of space in linear logic, we can safely forget the content of formulas, and the truth values too. Here it is important a location, i.e. address [9]. Each formula has its location in a proof, which we denote ξ .

Designs

We shall therefore build a system of locations. A design D represents a cut-free proof of some formula A , in which all logical information has been erased, only locations are kept, the locus of formula. This locus is the place where the name of formula is written. If formula A has been

located at the empty sequence $\langle \rangle$, then any formula occurring in the proof is a subformula of A : it has a precise location in the subformula tree of A . We can assume that the number of immediate subformulas of any formula B is at most denumerable, and we can name the various immediate subformulas of B as $B_0, B_1, B_2, \dots$. The natural numbers $0, 1, 2, \dots$, which characterize the immediate subformulas of B , are called *biases* [3].

A bias is a natural number, notation $i, j, k, \dots$. A *ramification* is a finite set of biases, notation $I, J, K, \dots$. A locus, or address, is a sequence $\langle i_1, \dots, i_n \rangle$ of biases, notation $\sigma, \tau, \nu, \xi, \dots$.

A locus denotes a place or spatial location of a formula. If we have a formula A and its proof, then A is in the root of the proof tree. Formulas in this tree are subformulas and they have precise locations (absolute or relative to the root). But if some subformula occurs in the proof tree more times, then its every occurrence need to receive a distinct location. If locus of A is σ and B is a subformula of A with bias i then locus of B is $\sigma * \langle i \rangle$. If we have a locus σ , then $\sigma * \tau$ is a sublocus. Symbol $'*$ ' denotes concatenation. Sublocus $\sigma * \tau$ can be *strict* if τ is non-empty sequence or it can be *immediate* if τ consists of just one bias [9]. For locus and sublocus is valid the ordering property: $\sigma \leq \sigma * \tau$.

We have the sequent $A \vdash \Lambda$. Every formula in this sequent has a locus; a finite sequence of biases. Now, we are considering only locations in sequent, so everything else is ignored, and shape, which we got called *pitchfork*[7],[9]:

$$\{\xi\} \vdash \Lambda$$

$\{\xi\}$ is a singleton, which contains only one locus $\{\xi\}$ and it can be empty set too. Λ is a finite set of loci. A locus on the left side is incomparable with every locus in Λ . A pitchfork $\{\xi\} \vdash \Lambda$ consists of a *handle* $\{\xi\}$ and *tines* (loci) in Λ . A pitchfork can be *positive* if it has no handle and *negative* if it has a handle. A pitchfork is *atomic* if it has one of the following form $\{\xi\} \vdash$, or $\vdash \{\xi\}$ [3], [9].

Example: Now we return to previous example and we forget everything except locations of formulas above. We assume that the locus of A is ξ , and biases for P, Q, R, S are 3, 4, 7, 8. We note that Γ, Δ, Λ are no longer formulas but loci. Now we are ready to rewrite previous rules and we get explicit information about space resources needed for proofs.

$$\frac{\xi 3 \vdash \Gamma}{\vdash \xi 3, \xi 4, \Gamma} \quad \frac{\xi 7 \vdash \Delta \quad \xi 8 \vdash \Gamma, \Delta}{\vdash \xi, \Gamma, \Delta}$$

After using De Morgan rules:

$$\frac{\vdash \xi 3, \xi 4, \Lambda \quad \vdash \xi 4, \xi 7, \Lambda}{\xi \vdash \Lambda}$$

We can see that every sequent in these proofs is a pitchfork. This example reflects the option how to represent any LL cut-free proof as pitchfork rules. This allows us to work in LL with space and time.

V. CONCLUSION

In our contribution we presented how to introduce time and space into LL. Incrementation of time depends on the change of formula's polarity in the proof tree. For better illustration we closed formulas with same polarity to cluster, which directly describes the change over time. Space in LL is presented by locus as it is seen in the example.

This approach creates the new possibilities in logic programming and enables to see logic as a means how to solve real life problems by computers.

ACKNOWLEDGMENT

This paper was supported by project VEGA No. 1/0015/10 Principles and methods of semantics enrichment and adaptation of knowledge-based languages for automatic software development.

REFERENCES

- [1] J. Y. Girard, "Linear Logic," *Theoretical Computer Science*, vol. 50, pp. 1–102, 1987.
- [2] D. Mihályi, V. Novitzká, and V. Slodičák, "From Petri Nets to Linear Logic," *CSE'2008*, pp. 48–56, 2008.
- [3] J. Y. Girard, "Linear Logic: Its Syntax and Semantics," *Advances in Linear Logic*, pp. 1–42, 1995.
- [4] J. Y. Girard, "From Foundations to Ludics," *Bulletin of Symbolic Logic*, vol. 09, pp. 131–168, 2003.

- [5] P. Lincoln, "Linear logic," *ACM SIGACT Notices News*, vol. v.23 n.2, p. 1992, 1992.
- [6] J. Y. Girard, "Proofs and Types," *Cambridge University Press*, pp. 1–175, 2003.
- [7] J. Y. Girard, "Locus Solum," *Mathematical Structures in Computer Science*, vol. 11, pp. 301–506, 2001.
- [8] J. M. Andreoli and R. Pareschi, "Linear objects: logical processes with built-in inheritance," *New Generation Computing*, vol. 9(3-4), p. 445473, 1991.
- [9] D. Mihályi and V. Novitzká, "Resource-oriented programming based on linear logic," *Acta Polytechnica Hungarica*, pp. 157–166, 2007.

3D Buildings Modeling and Visualization

Branislav SOBOTA, František HROZEK, Iveta REPICKÁ

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

branislav.sobota@tuke.sk, frantisek.hrozek@tuke.sk, iveta.repicka@tuke.sk

Abstract — Modeling and visualization of 3D objects is the last time one of the fastest growing sector of computer graphics. This work deals with modeling and subsequent visualization of the visual sights, namely The Spiš Castle. It describes the methods used by selecting suitable modeling means and procedures for the modeling and visualization.

Keywords — 3D modeling, 3D visualization, SketchUp

I. INTRODUCTION

Under visualization we can include any visual sensation received with eyesight. Caveman's have already communicate through the paintings, eg. the visual sensations. 3D modeling and visualization is in recent years a rapidly growing industry. 3D models are used in advertisements, films, or scientific presentations. In this paper is presented modeling and visualization of historical buildings with using Google SketchUp [1] for 3D model creation, SuperEngine for visualization and scripting language RUBY for animation creation.

II. ANALYSIS

Creation of 3D model for visualization needs a lot of effort. Everything begins with collecting of information and analysis. It is necessary to consider various aspects before starting the process of modeling, to prepare the groundwork. This is preparing phase. Also it is needed to track problems in a timely manner and try to avoid them. The actual collection of information can starts after this. After these steps starts modeling phase. A check of model for errors comes after 3D digital model creation. The visualization of the final model is the last step itself. This process is depicted in Fig. 1[2].

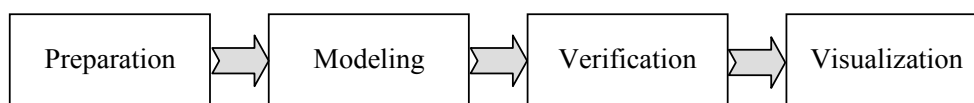


Fig. 1 Modeling and visualization process

A. Modeling tool

For castle modeling it is important to choose right modeling tool. On market are several applications for 3D modeling (Fig. 2). For Spiš Castle was chosen Google SketchUp.

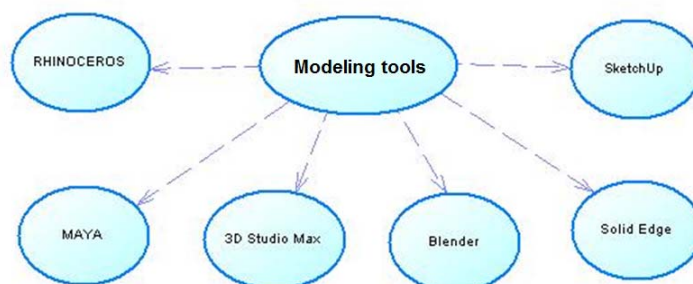


Fig. 2 Modeling tools

B. Visualization software

As visualization software was selected SuperEngine. This visualization software can handle up to 2 times 3 millions of polygons per visualized scene. Also supports network rendering and one visualized scene can be accessed by two or more people from different places through network or internet. SuperEngine supports models with .bin extension. For proper initialization it needs programs ControlCenter, Videorender and glut32.dll library. Scene in SuperEngine can be started by batch file (for example scena.bat) or by Control Console (Fig. 3) program. In this console, scenes can be loaded and triggered on the fly. The main advantage of using Control Console is faster scripting of scenes (SuperEngine use Ruby scripting language for scripting [3]), because no time is wasted with SuperEngine restarting.



Fig. 3 SuperEngine Control Console

C. RUBY scripting language

Ruby is script language which name is derived from ruby crystal. Ruby is programming language with focus on simplicity and productivity. It has an elegant syntax that is natural to read and easy to write. History of Ruby is dated to year 1993. In this year Yukihiro Matsumoto present this simple scripting language. Interpreter of Ruby is written in C language [3]

Example of Ruby script:

```
...
$try = ENGINE::LoadModel("try1.bin");           //loading model named try1.bin
$try.position(30,-50,-1);                        //position of this model in area
$try.visible = 1;Referces:                      //try will be visible in the area
...
def ENGINE::Actions                             //define the begining of action
sequence
$scene_id==0;                                  // the identifier of scene
$scene_count = $scene_count + $FRAMETIME;       //incrementation of the scene count

if ($scene_id==0) then                          //scene count

if(BlinkObject($try)) then                     //if the scene count match object will blink
  $scene_count=0;                             //the scene counter will be set on 0
  $scene_id=1;                                //scene id will be set on 1
end end                                         //end of if's
...

```

III. MODELING AND VISUALIZATION

Work on the model was divided into 3 parts: modeling of the Upper Castle, the lower courtyard and the tower. When 3D model was created, next stage was texture creation. After texturing, the 3D model was ready for visualization (Fig. 4).



Fig. 4 Final 3D model with texture

For visualization was used SuperEngine and animation was scripted with RUBY. Frames from final animation are shown in Fig. 5.



Fig. 5 Frames from final Spiš Castle visualization

For visualization can be also used multiprocessor systems. This system allows doing some visualization stages in parallel, which is real speedup [4], [5]. The immersive visualization of rendered images (real-time computed) is favorable solution. Multi-screen [6], big-screen [7] or stereoscopic display [1] [5] environments increase the quality, factuality and user experiences from castle visualization.

IV. CONCLUSION

The aim of this paper was description of modeling and visualization of Spiš Castle. After analyzing model requirements polygonal modeling appears to be an optimal modeling technique for this kind of project. Modeling tool for 3D model creation was Google SketchUp. As visualization software was used SuperEngine. Final animation of Spiš Castle was created in RUBY. Procedure for model and animation creation presented in this paper can be used for presentation and preservation of historical heritage of Slovakia.

ACKNOWLEDGMENT

This work is supported by VEGA grant project No. 1/0646/09: Tasks solution for large graphical data processing in the environment of parallel, distributed and network computer systems.

REFERENCES

- [1] Homepage of Google SketchUp [online]. [quoted 06.02.2011] Available at: <http://sketchup.google.com/>
- [2] F. Hrozek, B. Sobota, R. Janošo, "Visualization with 3D Interfaces", International Scientific Conference on Computer Science and Engineering, CSE 2010, Proceedings of, vol. 1, no.1, pp. 328-335, September 20-22, 2010, ISBN 978-80-8086-164-3
- [3] Homepage of RUBY scripting language [online]. [quoted 06.02.2011] Available at: <http://www.ruby-lang.org/en/>
- [4] B. Sobota, M. Straka, F. Hlinka, J. Perháč, "Parallel processing of visualization of 3D virtual map project", MOSMIC 2007 - Modelling and Simulation in Management, Informatics and Control, Žilina, 15.10-16.10.2007, Žilina, Fakulta manažmentu, riadenia a informatiky, Žilinská univerzita v Žiline, 2007, pp. 9-14, ISBN 978-80-8070-807-8
- [5] B. Sobota, J. Perháč, Cs. Szabó, Š. Schrötter, "High-resolution visualisation in cluster environment", Grid Computing for Complex Problem 2008, Bratislava, 27.10.-29.10.2008, Bratislava, Slovakia, Ústav Informatiky SAV, 2008, p. 62 - 69, ISBN 978-80-969202-9-7
- [6] O. G. Staadt, J. Walker, C. Nuber, B. Hamann, "A Survey and Performance Analysis of Software Platforms for Interactive Cluster-Based Multi-Screen Rendering", Proceedings of IPT/EGVE 2003, ACM Press, pp. 261-270 (IPT/EGVE 2003, Zurich, May 22-23, 2003).
- [7] Li Z. Liu, R. Samanta, G. Wallace, "Data Distribution Strategies for High Resolution Displays". Computers and Graphics Vol. 25, 811-818, 2001

Stereoscopic Image Processing

Branislav SOBOTA, František HROZEK, Martin FEČKANIN

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

branislav.soota@tuke.sk, frantisek.hrozek@tuke.sk, martin.fečkanin@tuke.sk

Abstract — Presented paper was created to give reader an impression of what is important in production of quality stereoscopic images, both static and dynamic. Paper focuses to identification and measuring differences between stereoscopic images produced by different still cameras taking the same scene. For objective comparison, custom-made test chart for identification of monoscopic image distortions was created, used and measured correction parameters, revealed by this chart, were applied to the processed stereo image pairs. Last part focuses to identification and reviewing of various methods applicable for live stereoscopic video streaming. Solutions were based on Windows Media™ and Adobe® Flash® streaming technologies.

Keywords — stereoscopy, linear parallax, anaglyph

I. INTRODUCTION

The image and video creation and processing in present world should offer something more, than just increased resolution and better compression techniques. For example, there doesn't seem to be much difference between 1080p and 720p video on the same 1080p screen sensed by viewers that would reason twice as much required bitrate, despite the 2,25x higher resolution. Also digital still cameras produce greater physical resolution and bigger files without much impact on perceived quality [1]. It seems that this is not the right way. 3D scene getting using 3D scanner is the other way for an image of real world acquisition [5][7]. Some algorithms for visibility solving are used for the final image computing in this case [6].

One of the possible evolution steps in consumer video and imaging products could be stereoscopy [2]. This paper was created to give reader an impression of what is important in production of quality stereoscopic images, both static and dynamic. First part of the paper focuses to identification and measuring differences between stereoscopic images produced by different still cameras, when shooting the same scene. For objective comparison, custom-made test chart for identification of monoscopic image distortions was created, used and measured correction parameters it revealed were applied to the processed stereo image pairs. Second part focuses to identification and reviewing of various methods applicable for live stereoscopic video streaming. Solutions were based on Windows Media™ and Adobe® Flash® streaming technologies.

II. STEREOSCOPIC IMAGE PROCESSING

Parameters for stereoscopic photography can be divided into two categories: category of classic photography and category of stereoscopic photography. The same state is in movie or in visualization area [8].

Two approaches for creating stereo image pairs were identified – „toe-in“ and parallel method of shooting. Toe-in approach was considered for review, but because of length of the work and lack of time, it had to be dropped in parts utilizing still images and it was used only when stereo video was generated.

There are several formulas for maximum fusible parallax available, but these two were used:

$$B = \frac{P}{L - N} \left(\frac{LN}{F} - \frac{L + N}{2} \right) \quad (1)$$

$$B = \left(\frac{m}{m-1} \right) \frac{N}{30}, \quad m = \frac{L}{N} \quad (2)$$

where B represents computed stereo base, P is maximum allowable on-film deviation (usually taken as $P=F/30$ [1]), L is distance of furthest object in scene, N is distance of closest object in scene and F is focal length [3][4].

Because sensor size was determined as the only factor, that could have an impact to differences between pictures of the same scene taken with different cameras (regardless focal length; this was corrected using a system of ratios), a review of cameras based on their sensor size was created. Five categories were identified: compact, ultra compact, rugged compact, EVF and SLR. Later, categories ultra compact and rugged compact were added to single category called “compact” because of their similar functionality. Review contained 164 different cameras (Fig. 1).

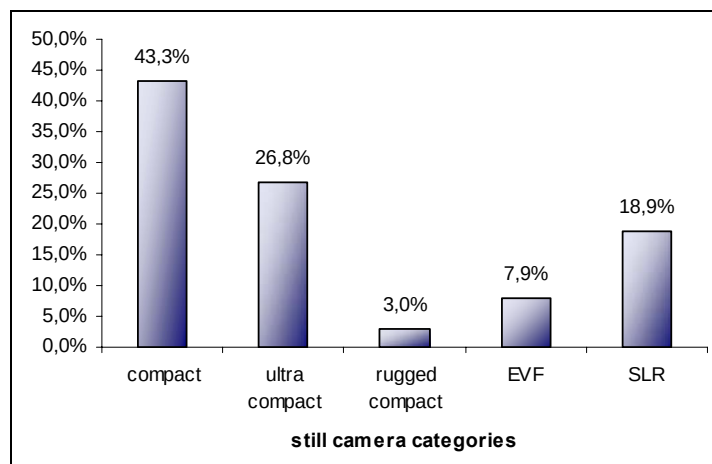


Fig. 1 Camera shares

As expected, most of selling cameras belonged to compact and ultra compact category; and if we assume joined compact category (compact + ultra and rugged compact), it takes 73,1% share.

As further reviewed, different sensor sizes dominate in each category. Smallest sensors (1/2,5“ and 1/2,3“) dominate in joined compact category, 1/2,3“ type is the choice for most EVF cameras and in SLR segment, APS-C standard has the highest share (64,5%).

Next step was to design a test chart that would clearly show the amount of barrel and pincushion distortion, chromatic aberration and sharpness degradation from image center to its corners when testing used cameras. Created chart accomplished all of listed requirements (Fig. 2).

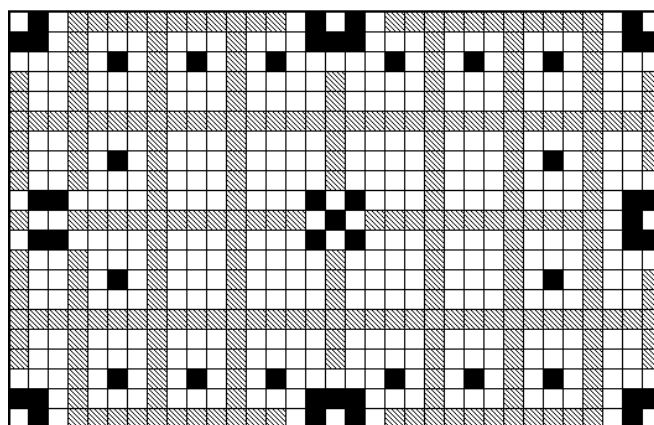


Fig. 2 Designed test chart

Using test chart above (Fig. 2), correction parameters for barrel distortion correction were measured and later applied to produced photos. At wide end, every digital camera used in work needed some degree of barrel distortion correction. Amount of needed correction greatly descended as focal length raised.

Four different scenes were chosen for shooting. First two of them were exterior (one deep, one shallower), the third was interior and last scene was chosen for extreme hyperstereoscopy testing (exterior again). All of the scenes are located in Košice, Slovakia, and all of them

contain mainly architecture and clean lines. This eases measuring objects and corresponding parallaxes.

More than 1 000 image files were generated. This number contains calibration takes, corrected photos, photos of selected scenes and processed stereo images in form of LR image and half-color anaglyph.

Results are confusing. First of all, there sure is a noticeable difference between stereo pairs shot using different cameras, but no connection to the sensor size was found. Even cameras with the same sensor type (1/2.5") and exactly the same focal length (35mm, equiv.) produced different photos. More unexpectedly, changes of aperture size changed size of linear parallax (Fig. 3).

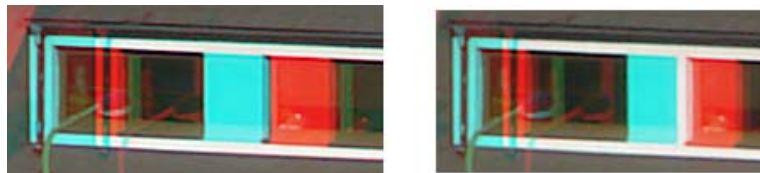


Fig. 3 Linear parallax change; Sony DSC-H1, F=64mm; (left) F5,6, (right) F8

Some degree of surprise created also color changes in anaglyph images. These changes were driven by background color (Fig. 4).

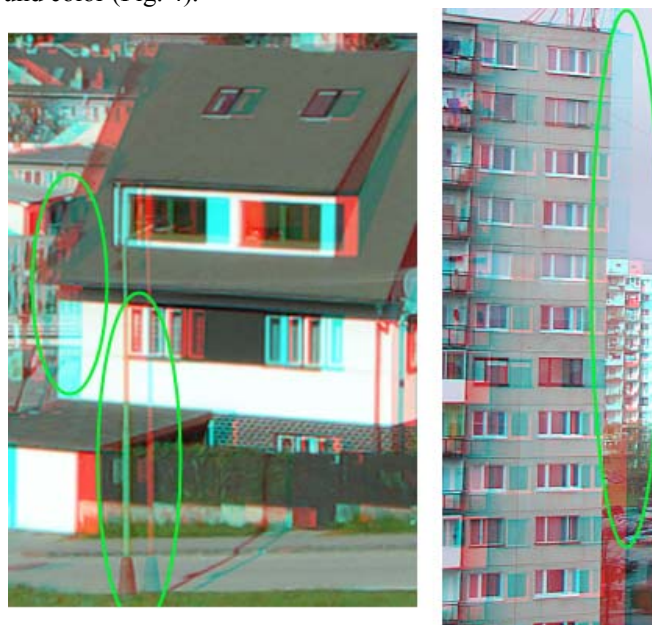


Fig. 4 Anaglyph color changes

Despite of expectations, noise in images did not cause any problems, when stereo images were viewed. Much more disturbing were changes of light, which modified scene contrast, color saturation and also shades intensity. Pictures, where light conditions were significantly different for left and right images, were less comfortable to view, but still contained sensible depth information.

Because differences in pictures were confirmed, but no link to sensor size as determining factor to this has been found, efforts to identification cause of these differences should be made.

III. LIVE STEREOSCOPIC VIDEO STREAMING

Main goal of the second part of the work was to find a solution for broadcasting live stereoscopic video. Two approaches were used – combination of two „independent“ streams from cameras to single side-by-side video stream processed by encoding application and streamed using appropriate streaming server, when the type of the final stereoscopic video (active stereo, color anaglyph, ...) is determined by commercial software player as a first approach; the second approach relies on the combination of two „independent“ streams from cameras to single color or gray anaglyph video stream or single side-by-side video stream processed by encoding application and streamed using appropriate streaming server in a form

suitable for direct viewing at the client side (passive projection using polarizing or Infitec™ filters gaining from side-by-side setup, or using red-cyan goggles for color and gray anaglyph). As compression and streaming solutions were chosen Adobe® Flash™ and Microsoft® Windows Media™ with appropriate applications and streaming servers. QuickTime™ format was also considered, but later dropped. Two schemas were used (Fig. 5 and Fig. 6):

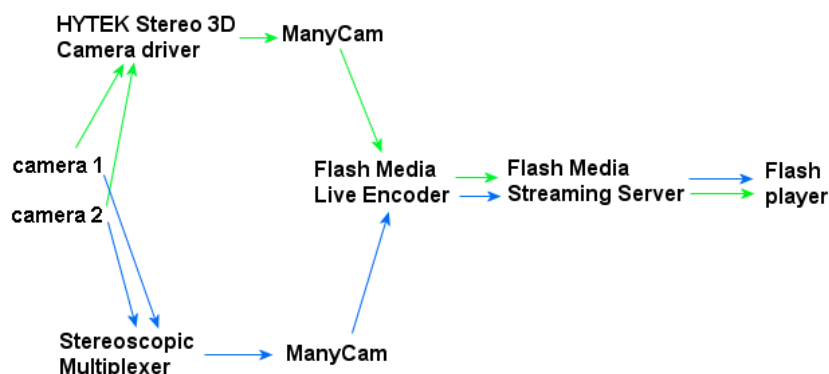


Fig. 5 Live stereoscopic video streaming using Adobe® Flash®

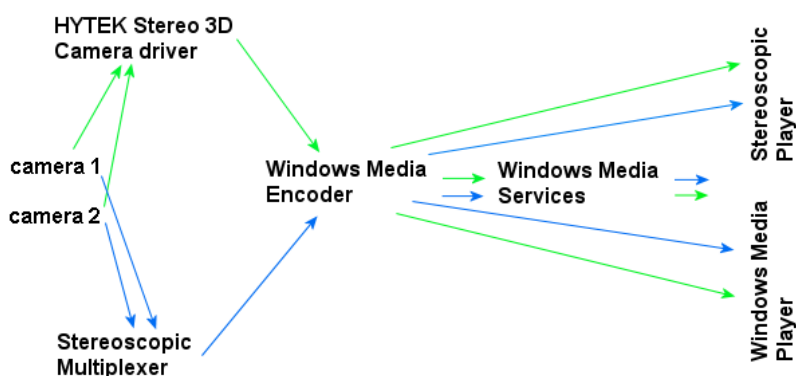


Fig. 6 Live stereoscopic video streaming using Microsoft® Windows Media™

Flash video was selected as a solution for anaglyph video streaming because of its ease of use and because of widespread Adobe Flash player. Everything what the client is required to do is to put on cheap goggles with red and cyan filters and watch the content. More versatile solution depends upon Windows Media streaming technology and is based on transmission of LR video stream, which is composed to the final form at client side. This has also higher bitrate demands, because LR format doubles the horizontal resolution.

When the source of live video is webcam (which often produces noisy images and noise consumes large portions of available bitrate), two setups were identified as balanced (giving reasonable quality at lowest bitrate).

For colour anaglyph and Flash video streaming, H.264 codec and resolution of 320x212 pixels was used (Fig. 7). Video bitrate for optimal image quality was chosen as 436kbps (total stream bitrate was 436kbps video + 64kbps mp3 audio = 500kbps).



Fig. 7 Flash® video, H.264 Baseline, 436kbps, anaglyph

IV. CONCLUSION

Presented paper was created to give reader an impression of what is important in production of quality stereoscopic images, both static and dynamic. First part was focused to identification and measuring differences between stereoscopic images produced by different still cameras taking the same scene. For objective comparison, custom-made test chart for identification of monoscopic image distortions was created. Assumption, that only thing that would have impact on generated photos will be different sensor size of used still cameras, turned out to be incorrect. Generated images were different, but no connection to sensor size was found. Second part was focused on problematic of live stereoscopic video streaming. Tests showed that use of Adobe® Flash® is the best solution for anaglyphic video broadcast.

ACKNOWLEDGMENT

This work is supported by VEGA grant project No. 1/0646/09: Tasks solution for large graphical data processing in the environment of parallel, distributed and network computer systems.

REFERENCES

- [1] Sobota, B., Szabó, Cs., Šuba, S.: "The some problems by construction of multi-screen projection system", Journal of Computer Science and Control Systems, 2, 1, 2009, pp. 57-60, ISSN 1844-6043
- [2] F. Hrozek, B. Sobota, R. Janoš, "Visualization with 3D Interfaces", International Scientific Conference on Computer Science and Engineering, CSE 2010, Proceedings of, vol. 1, no.1, pp. 328-335, September 20-22, 2010, ISBN 978-80-8086-164-3
- [3] J. Bercovitz, "MAOFD Explanation" [online]. [quoted 07.02.2011] Available at: <http://www.angelfire.com/ca2/tech3d/maofd.txt>
- [4] J. Wattie, "Bercovitz and DiMarzio equations". [online] [quoted 08.02.2011]. Available at: <http://www.kiwizone.org/stereo/3dtake/newversionberk.html>
- [5] Kari, P., Abi-rached, H., Duchamp, T., Shapiro, L.G., Stuetzle, W.: Acquisition and Visualization of Colored 3D Objects. In Proceedings of the 14th international Conference on Pattern Recognition, Vol. 1 (August 16 - 20, 1998). ICPR. IEEE Computer Society, Washington, DC, 11.
- [6] Katz, S., Tal, A., Basri, R.: Direct Visibility of Point Sets. In ACM SIGGRAPH 2007 Papers (San Diego, CA, August 05 - 09, 2007). SIGGRAPH '07. ACM, New York, NY, 24.
- [7] Sobota Branislav, Rovňák Maroš, Szabó Csaba: 3D scanner data processing, Journal of Information, Control and Management Systems, Volume 7, 2, 2009, pp. 191-198, ISSN 1336-1716
- [8] Sobota Branislav, Perháč Ján, Szabó Csaba: An application of parallel, distributed and network computer systems to solve computational processes in an area of large graphical data volumes processing., Computer Science and Technology Research Survey, Košice, 27.11.2008, Košice, KPI FEI TU Košice, 2008, 3, pp. 37-42, ISBN 978-80-8086-100-1

An Anomaly-Based Intrusion Detection System

Eva DANKOVÁ, Norbert ÁDÁM, Peter JAKUBČO

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

eva.dankova@tuke.sk, norbert.adam@tuke.sk, peter.jakubco@tuke.sk

Abstract — With the development of wireless networks, is nowadays necessary to ensure that the appropriate level of security is available. For better security in wireless system can be used intrusions detection based on anomalies, which detects network intrusion based on gathering and analysis of information. Paper deals with the design of hierarchical, scalable and flexible architecture of intrusion detection system, to thwart such threats. Dividing anomaly detection in several stages reduces the elements necessary for detection. Architecture can be extended with new anomaly detection methods in the base phases of the detection system. Based on the flexibility properties of proposed system, can be applied in different networks.

Keywords — intrusion detection system, anomaly approach, attack, anomalies

I. INTRODUCTION

Much attention has been focused recently on the security aspects of wireless LAN systems. The rapid growth and deployment of these systems into a wide range of networks and for a wide variety of applications drives the need to support security solutions that meet the requirements of a wide variety of customers.

Wireless networks are exposed to various harmful systems and attacks of ambient environment, such as sniffing, denial of service attacks in the vicinity of a Particular access point. Damage that these attacks can evocate includes leak sensitive information, unauthorized modifying of data and other activities. Therefore, these systems need measures that would enable them to detect attacks and then to take the appropriate measures.

Intrusion detection is a form of detection system, which is based on the assumption that the attack must be different from normal user behavior. To detect such an attack is quite difficult, and mostly fail to reveal it until some damage occurs. Systems based on the detection, must contain a database of behavior, based on which the system is trained. Training is not only time-consuming but can also lead to mistakes.

An alternative approach to the analysis of network behavior is more robust and scalable. The basis of network analysis algorithms is based on anomalies, which is then used to detect attacks. Method based on anomaly is particularly suitable for the detection of worms, malware, acceptable-use policy violations and insider misuse, which requires in comparison with other methods less settings and ongoing updating.

Some intrusion detections based on anomalies is not possible to use in high-speed networks, because it would require a high consumption of elements. This paper deals with intrusion detection system architecture that is hierarchical, extensible and flexible. The fundamental reflection is that the anomaly detection is divided into several stages, thereby is the consumption of resources reduced. The proposed architecture can be extended with new types of anomaly detection methods in the primary phase to the detection system. And due to its flexibility it is possible to use the given architecture in different networks.

Proposed system, which was realized at Department of Computers and Informatics FEEI TUKE, was supported by project APVV-0073-07.

II. ANOMALIES IN WIRELESS NETWORKS

The aim of this paper is to propose an effective wireless intrusion detection system that uses anomaly detection techniques to detect various types of attacks in the wireless network. For this

reason, it should be noted anomalies that may occur.

A. Worm propagations

Network worms are an especially virulent form of abuse that have become an increasing threat to the Internet over the past several years. Unlike their viral predecessors, worms are network-aware and self-propagating, greatly increasing their potential reach and impact. Worm propagation attacks exploit security weaknesses in operating systems or applications to penetrate a system and to propagate themselves to as many other systems as possible.

To detect this type of attack is used the intrusion detection based on the signature. These detection systems are located on the edge of the network, thereby preventing the spread of worms in the network. It is very important to detect worm in its early days and respond in a minimum of time.

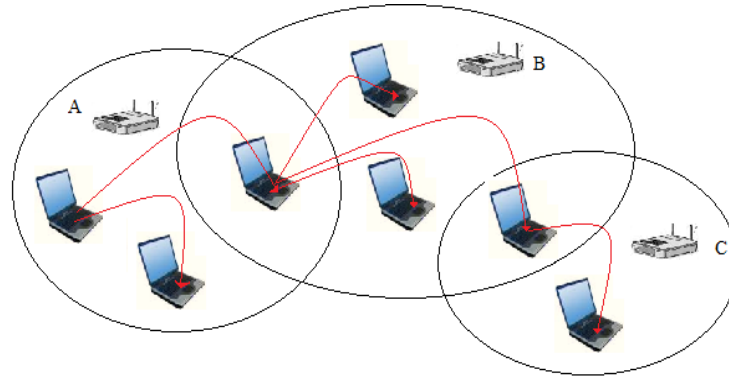


Fig. 1 Worm propagation attack

System based on signature are time-demanding for worm packet analysis, and therefore is advisable to use system based on anomalies, which can increase the speed of detection by monitoring traffic changes. One possibility is matching destination port numbers between incoming and outgoing connections, analyzing the ratio of error messages due to closed ports generated by the scanned systems to the total number of connection requests.

B. Wormhole attack

The wormhole attack is a very serious problem, which is directed against packet routing in networks. This type of attack is very difficult to detect and prevent. In this attack, an adversary receives packets at one location in the network and tunnels them (possibly selectively) to another location in the network, where the packets are resent into the network.

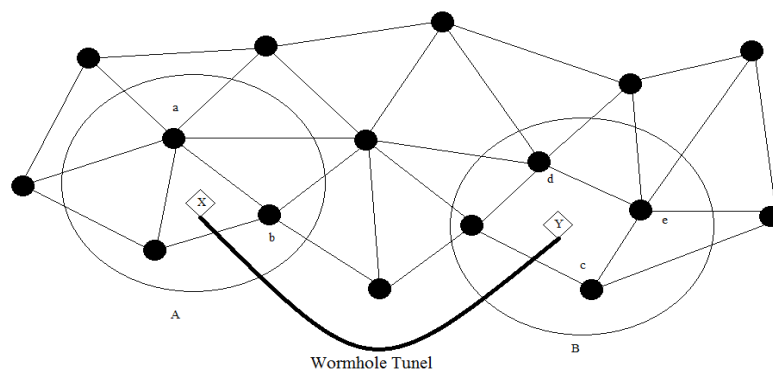


Fig. 2 Wormhole attack

Capabilities of the wireless network allow greater flexibility in network and therefore cannot ensure any knowledge about a participating node. Therefore as a result it is much easier to threaten the routing protocols or certain connections than in wired networks. By establishing a wormhole, an attacker aims at attracting as much traffic as possible to a node controlled by himself by influencing the routing metrics. The attack causes increased volume of traffic near the tunnel endpoints.

C. DDoS attacks

DDoS attacks are among the greatest threats. This type of attack does not utilize a lack of system, but focuses on source overloading. This type of anomalies includes attacks as link capacity or system memory by flooding the system with more traffic than it can process. The aim of DDoS attacks is to prevent user access to network resources.

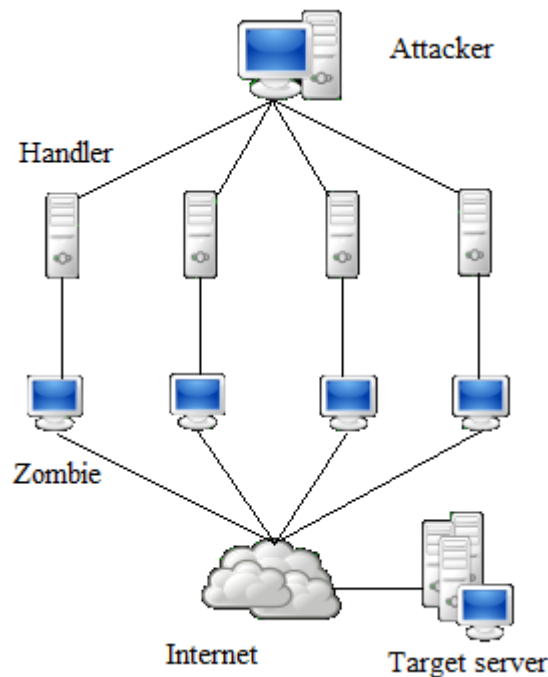


Fig. 3 DDoS Architecture

The most common way of DDoS attack is to flood a network with degenerate or faulty packets, crowding out and causing legitimate traffic systems not to respond. This anomaly would result in heavy traffic volumes.

D. Botnet

Botnet, also known as robotic network, is a group of computers that run the same computer robotic application controlled or manipulated by human operators. They are also called as botmasters for the purpose of doing co-related tasks. The system is organized in a simple hierarchical structure comprised of many computers, also called as drones or zombies, and a command center that controls all of the drones or zombies into action. Botnets can be large (composed of ten to twenty thousand drones) or small (composed of five hundred to a thousand drones). It really depends on the sophistication and complexity of its use. Botnets are very good tools in the hacking world because of their ability to gain other computers that could be used for any purpose.

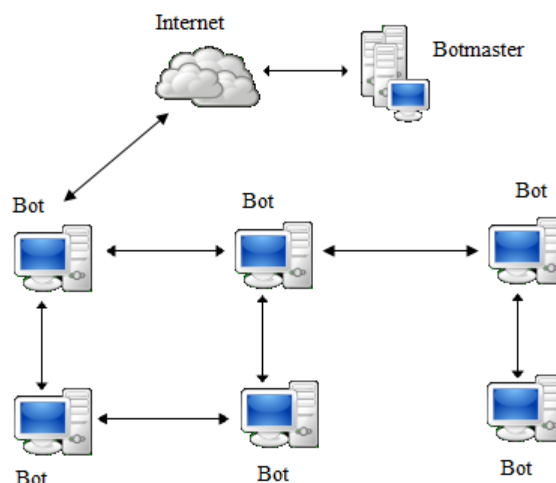


Fig. 4 Botnet attack

A botnet detector can look at exploit patterns, code downloading, bot coordination communications and outbound attack launches. By comparing these known botnet traffic patterns to the traffic flows within the trusted network, it can alert users when it detects flows that are likely to comprise botnet traffic.

III. ATTACK DETECTION APPROACH

In the case of attacks in wireless networks, attacked node allure other nodes to send their traffic to him. The consequence of this attack can be the traffic collision, or disabling access to the source node sending packets to the target amount. As it is evident, so this type of attack influences the traffic in different ways. If a wormhole is set up by an attacker, traffic to some nodes (nodes compromised) will be observed to suddenly increase.

DDoS attack increases number of packet collisions in the network, while worm propagation will demonstrate an increasing amount of traffic related to one node. Therefore, we can detect attacks in wireless networks by monitoring for these anomalies. Intrusion detection system based on anomalies can be highly effective and can be applied to different networks. Nevertheless, it is necessary to adjust some thresholds whenever the application is moved and every network or host has its own peculiarities and without threshold tuning the network administrator is often alerted to events that are normal on some hosts.

IV. PROPOSED ARCHITECTURE

The wireless network covers usually large physical area, and therefore are employed several access points due to provide adequate signals throughout the network. A fundamental aspect of the application of intrusion detection system based on anomalies is to place sensors everywhere where the access points are situated. By providing comprehensive coverage of the physical infrastructure at all sensors with wireless access points locations, the majority of attacks can be detected. Proposed security architecture is based on functionality combination of two agents:

- Node agents
- Root agents

The node agents are used to monitor local security and are specially designed for individual network node. This type of agent functions with no abstraction. All nodes on the network have a Corresponding Nodal Agent, Which is observing packet distribution. This observing is the first step of detection mechanism. In the event of an attack node agent notifies root agent. Using a centralized approach, the root agent is deployed in the access point itself. This is used as central management system to gather and process all data. Its primary goal is to monitor the state of its network infrastructure and Corresponding node agents.

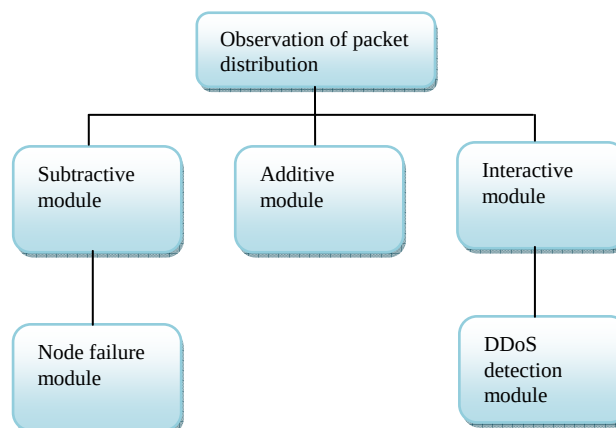


Fig. 5 Architecture of the detection system

The proposal is a hierarchical, extensible and flexible system. The observation of packet distribution is the first stage of the hierarchical detection system, which dedicates only a low level of analysis effort. The next stage is passed in the case, if during the first level an condition of intrusion or anomaly is detected. These stages detect only anomalies based on the information collected in the first stage. Splitting intrusion detection system based on anomalies

into several stages help to save resources. The common functionality of a machine is less affected by a hierarchical system than by a one stage system.

The proposed system can be also extended to include new anomaly detection methods into the basic stage. Another advantage of the proposed system is that it can be applied in different networks and scenarios due to its flexibility. This architecture support high security in a networking scheme.

V. CONCLUSION

An intrusion detection system based on anomaly approach is an important addition to the security in wireless networks. Due to drawbacks to implementing this type of wireless intrusion detection system, the benefits should prove to outweigh the shortcomings. With the capability to detect a variety of attacks, the benefits of a wireless IDS based on anomaly approach can be substantial. In view of the immense rate of wireless adoption, the ever-increasing number of threats to WLANs, and the growing complexity of attacks, a system to identify and report on threat information can greatly enhance the security of a wireless network.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. APVV 0073-07, also the authors are pleased to acknowledge the financial support of the Cultural and Educational Grant Agency KEGA of the Slovak Republic under grant No. 3/7110/09.

REFERENCES

- [1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
- [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [4] K. Elissa, "Title of paper if known," unpublished.
- [5] R. Nicole, "Title of paper with only first word capitalized", *J. Name Stand. Abbrev.*, in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [7] M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.

Aplikačné prostredie pre správu systémov a služieb v počítačových sieťach

Eva DANKOVÁ, Norbert ADÁM, Peter JAKUBČO

Katedra počítačov a informatiky, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika

eva.dankova@tuke.sk, norbert.adam@tuke.sk, peter.jakubco@tuke.sk

Abstrakt — Problematike metód správy počítačových sietí sa venuje v dnešnej dobe pomerne intenzívne. Prostredníctvom tohto článku je možné tieto metódy doplniť. Aplikácia navrhovaného približenia k problematike správy počítačových sietí je demonštrovaná vytvorením počítačovej siete pomocou vopred definovaných pravidiel. Tento článok analyzuje možnosti správy služieb s návrhom modulárneho systému pre potreby ďalšieho rozširovania systémových a komunikačných služieb, ktoré sú v sieti poskytované. Výsledok práce je možné aplikovať v ľubovoľnej rozsiahlej sieti, pričom sieťovým správcom zjednoduší správu služieb a urýchli konfiguračné procesy, ktoré sa pri správe počítačovej siete vykonávajú.

Kľúčové slová — počítačové siete, správa systémov, OSI model

I. ÚVOD

Základom ľubovoľnej komunikácie je informácia, ktorá je pragmaticky chápaná ako akákoľvek relevantná správa alebo údaj. Prirodzeným aspektom ľudskej povahy je potreba komunikácie, avšak ak sa vzdialenosť medzi komunikujúcimi osobami zväčší, objavia sa prvé problémy. Na tomto mieste vzniká potreba použitia rôznych techník pre prenos informácií, ktoré umožňujú ich prenos na väčšiu vzdialenosť. Začiatkom šesťdesiatych rokov minulého storočia sa začali v Spojených štátoch amerických objavovať myšlienky na vytvorenie rozsiahleho komunikačného média, ktoré by umožňovalo výmenu informácií – počítačovú sieť. Táto sieť mala spájať strategické vojenské a vládne miesta.

Prvotné použitie teda počítačová sieť našla v sfére obrany štátu vo vtedy nepokojnej dobe, ktorá sa dnes označuje ako obdobie studenej vojny. Vďaka tomu bol položený základ koncepcie počítačovej siete tak, ako je známy dodnes. Pokrok vedy a techniky od tej doby v mnohom pokročil, avšak využitie počítačových sietí ostalo rovnaké. Ich základnou úlohou je naďalej prenos informácií, avšak pojem informácia naberá čoraz širší význam. Najčastejšou úlohou sieťových špecialistov pri správe počítačových sietí je vkladanie rovnakých informácií na viacero počítačov, ktoré príslušné služby využívajú. Medzi základné problémy, ktoré riešia sieťový špecialisti patrí registrácia používateľa a pridelenie služieb, ktoré môže využívať. Administrátori v prípade, že je komplex využívaných služieb v rámci spravovanej siete príliš veľký, väčšinou vkladajú údaje o používateľoch do nezávislých systémov poskytujúcich jednotlivé služby. V takomto prípade je výhodné vytvoriť akýsi systém hromadného vkladania informácií, ktorý sa správa ako samostatná služba. Článok sa zaoberá možnosťou hromadnej správy služieb v počítačových sieťach na základe vlastných pravidiel, ktoré slúžia na konfiguráciu jednotlivých služieb. Umožňuje administrátorom počítačových sietí okamžité aplikovanie nastavení poskytovaných služieb.

II. OSI MODEL

Pre pochopenie a následný návrh aplikačného prostredia pre manažment počítačových sietí je potrebné najprv pochopiť ich fungovanie a základný model. Základný model Open System Interconnection (OSI), ktorý vyvinula Štandardizačná Organizácia (ISO) rieši problém inkompatibility. Znalosť tejto architektúry je základným predpokladom pochopenia funkcie počítačových sietí, spôsobu prenosu dát a nadväzných technológií.

Tento model rozdeľuje protokoly podľa funkcie do siedmich vrstiev (Fig. 1). Každá vrstva používa iba funkcie vrstvy, ktorá sa nachádza pod ňou a poskytuje funkcionality nadradenej vrstve.

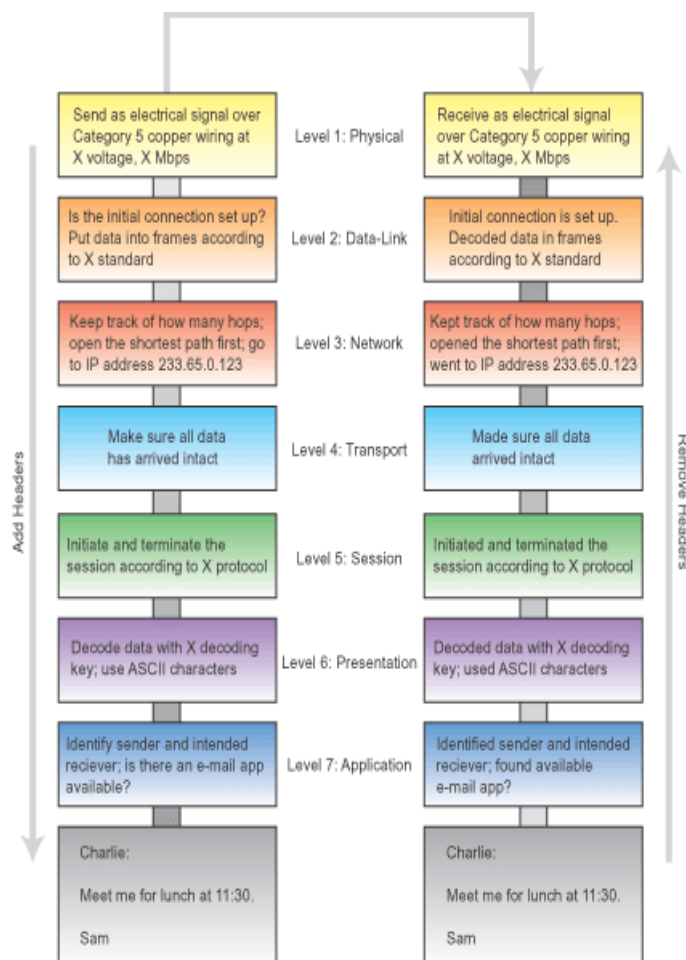


Fig. 1 Referenčný model OSI

Všetky systémy, ktoré implementujú tento protokol pozostávajúci z vrstiev sa nazývajú protocol stack (napríklad TCP/IP stack). Stack je možné implementovať softvérovo, hardvérovo alebo ich kombináciou. Vďaka tomu vrstva, ktorú implementoval jeden výrobca, spolupracuje s inou vrstvou, implementovanou druhým výrobcom (za predpokladu správne implementovanej špecifikácie).

Medzi používateľskou a sieťovou časťou je 4. vrstva, ktorá je chápaná ako rozhranie. Zatiaľ čo prvé štyri vrstvy sú pomerne exaktne definované, zvyšné tri vrstvy nemusia byť striktné použité tak, ako sú definované podľa tohto modelu. Teoreticky každá vrstva pridáva na začiatok balíku dát vlastnú hlavičku s údajmi tejto vrstvy a na záver kontrolný súčet alebo informácie o ukončení dát vrstvy. Okrem základného modelu je potrebné ozrejmiť si aj adresovanie v sieťach.

III. ADRESOVANIE

Zatiaľ čo pre užívateľov je vhodné označovanie jednotlivých sieťových zariadení menami, komunikačné systémy používajú adresy (najčastejšie číselné, vyjadrené desiatkovom, dvojkovom či šestnástkovom tvare). Protokol IP je nespojovaný datagramový protokol, ktorý je predovšetkým zodpovedný za adresáciu a smerovanie paketov medzi hostiteľmi. „Nespojovaný“ znamená, že pred výmenou dát nie je ustanovená relácia a taktiež nie je garantované doručenie paketov. Každý hostiteľ TCP/IP je identifikovaný pomocou logickej adresy IP, ktorá musí byť jedinečná pre každého hostiteľa či sieťový komponent v danej sieti. Každá adresa IP obsahuje ID hostiteľa a ID siete:

- ID hostiteľa (adresa hostiteľa) identifikuje pracovnú stanicu, server, smerovač alebo iného TCP/IP hostiteľa v sieti. ID každého hostiteľa v danej sieti musí byť jedinečné a odlišné od ID siete.

- ID siete (sieťová adresa) identifikuje systémy, ktoré sú na rovnakej fyzickej sieti ohraničenej IP smerovačmi. Všetky systémy na rovnakej sieti musia mať rovnaké ID siete a to musí byť medzi sieťami jedinečné.

IP adresa pozostáva z 32 bitov, spravidla sa používa členenie na 8-bitové segmenty nazývané oktety. Každý oktet sa udáva ako desiatkové číslo v rozsahu 0 až 255 a oddeľuje sa bodkou, takže adresa má tvar „w.x.y.z“. V rámci siete internet je definovaných 5 adresných tried, ktoré náležia sieťam rôznych veľkostí. Trieda siete určuje, ktoré bity sú použité pre ID hostiteľa a ktoré pre ID siete, a teda aj možné množstvo hostiteľov a množstvo sietí.

- Trieda A – adresy sú pridelované sieťam s veľmi veľkým počtom hostiteľov. Najvýznamnejší bit je v adrese triedy A vždy nastavený na nulu. Nasledujúcich 7 bitov (dokončujúcich prvý oktet) tvorí ID siete. Zostávajúcich 24 bitov reprezentujú ID hostiteľa.
- Trieda B – adresy sú pridelované stredne veľkým až veľkým sieťam. Dva najvýznamnejšie bity sú v adrese vždy nastavené na hodnotu 10. Nasledujúcich 14 bitov tvoria ID siete. Zostávajúcich 16 bitov reprezentujú ID hostiteľa.
- Trieda C – adresy sú pridelované malým sieťam. Tri najvýznamnejšie bity sú v adrese triedy C vždy nastavené na hodnotu 110. Nasledujúcich 21 bitov tvoria ID siete. Zostávajúcich 8 bitov reprezentuje ID hostiteľa.
- Trieda D – adresy triedy D sú vyhradené adresám IP pre viacsmerné vysielanie. Štyri najvýznamnejšie bity sú v adrese triedy D vždy nastavené na hodnotu 1110. Zostávajúce bity sú vyhradené pre adresu, ktorú rozpoznávajú zainteresovaní hostitelia.
- Trieda E – trieda E je vyhradená pre budúce použitie. Najvýznamnejšie bity sú nastavené na hodnotu 1111.

IV. KOMPARATÍVNA ANALÝZA EXISTUJÚCICH RIEŠENÍ

V oblasti manažmentu počítačových sietí existuje veľké množstvo produktov zaoberajúcich sa touto problematikou. Analýza je zameraná na aplikácie slúžiace výlučne na manažment sietí a sú zamerané na riešenie špecifických problémov.

A. Telconi Terminal

Telconi Terminal je komerčná aplikácia určená pre sieťový manažment produktov firmy Cisco. Táto aplikácia slúži sieťovým administrátorom pri konfigurácii počítačových sietí a riešení každodenných problémov. Spája využívanie príkazového riadku s grafickým rozhraním vytvoreným pre túto aplikáciu. Telconi

Terminal je veľmi prehľadný a užívateľsky prístupný nástroj pre správu počítačových sietí. Na rozdiel od ostatných produktov nie je závislý na protokole SNMP, ale nepodporuje možnosť konfigurácie sieťových zariadení prostredníctvom sériového portu, čo predstavuje veľkú nevýhodu pri práci s novými zariadeniami bez základnej konfigurácie. Nevýhodou tohto produktu je však potreba profesionálnej znalosti konfiguračných príkazov podporovaných sieťových zariadení.

B. Kiwi CatTools

Kiwi CatTools je voľne dostupný systém, ktorý je vhodným riešením pre manažment a zálohovanie sieťových zariadení. Prednosťami tohto systému sú notifikácia prostredníctvom elektronickej pošty (e-mail) a zvyrazňovanie rozdielnosti v reportoch od jednotlivých zariadení. Taktiež podporuje protokoly Telnet, SSH, TFTP a SNMP.

Kiwi CatTools predstavuje užitočnú variantu pri výbere systému pre správu veľkej počítačovej siete, ktorá obsahuje veľké množstvo manažovateľných sieťových prvkov. Jeho výhodou je možnosť súčasného pripojenia na viacero zariadení v rovnakom čase. Okrem toho ponúka funkciu automatického zálohovania konfigurácie manažovateľných zariadení. Nevýhodou tohto systému je absencia pripojenia prostredníctvom sériového portu a malá prehľadnosť grafického rozhrania tohto systému.

C. Cisco SNMP Tool

Cisco SNMP Tool patrí do kategórie voľne šíriteľných aplikácií, ktorý okrem možnosti manažovania zariadení od firmy Cisco poskytuje aj funkciu zisťovania topológie manažovanej

siete. Výsledkom typologickej analýzy je grafické zobrazenie aktívnych sieťových prvkov v danej sieti. Aplikácia Cisco SNMP Tool poskytuje možnosť zisťovania topológie počítačovej siete, avšak iba v prostredí zariadení, ktoré podporujú protokol CDP a majú túto podporu zapnutú. Potrebuje čítať prístup cez protokol SNMP na všetky zariadenia pripojené v sieti. Táto aplikácia môže byť použitá ako nástroj pre zisťovanie topológie LAN a WAN sietí, ale pri sieťach s väčším počtom uzlov nastávajú problémy s vizualizáciou. Na základe uskutočnenej analýzy je možné realizovať návrh vlastného aplikačného prostredia.

V. NÁVRH REALIZÁCIE APLIKAČNÉHO PROSTREDIA

Hlavnou myšlienkou návrhu aplikačného prostredia je rozdelenie celého projektu na jednotlivé vrstvy, ktoré v budúcnosti umožnia jeho rozšírenie a doplnenie o rôzne nové komponenty slúžiace pre správu a komunikáciu so sieťovými prvkami.

Aby bolo možné neskôršie rozšírenie systému o ďalšie prvky a možnosti, je potrebné definovať procesy pripojenia, riadenia, odosielania údajov a vytvárania jednotlivých makier, na základe ktorých môžu v budúcnosti vzniknúť doplnkové vrstvy. Prostredie aplikácie je objektovo orientovaný jazyk C#. Objektovo orientovaný prístup tohto jazyka je výhodný kvôli potrebe definovania jednotlivých tried a členských metód jednotlivých vrstiev, ktoré spolu realizujú funkčnosť daného systému. Pre definovanie vrstiev platia nasledujúce vlastnosti:

- Každá vrstva je uložená vo vlastnom adresári, ktorého názov je identický s názvom danej vrstvy. V tomto adresári sa nachádza súbor s príponou .csproj, ktorý obsahuje definície tried, ktoré vrstva využíva a komplexné informácie o danej vrstve, ktorých štruktúra je uložená vo formáte XML.
- Členské triedy sú uložené v adresári danej vrstvy a ich názov zároveň reprezentuje ich funkciu.

A. Application Layer

Táto vrstva zabezpečuje základnú funkčnosť celého aplikačného prostredia nakoľko inicializuje základné premenné, ktoré sa využívajú pre komunikáciu, vytváranie konfiguračných súborov a definovanie makier. Obsahuje aj triedu pre zaznamenávanie udalostí, čo zabezpečuje bezpečnosť fungovania jednotlivých vrstiev.

Súčasťou tejto vrstvy je aj systém pre odchytyvanie jednotlivých chybových hlásení, ktoré môžu nastať počas realizácie a behu tohto aplikačného prostredia. Je zameraný na chyby, ktoré sa môžu vyskytnúť pri realizácii a následnej funkčnosti metód, slúžiacich na komunikáciu so sieťovými zariadeniami a na zaznamenávaní udalostí počas behu programu.

B. Data Layer

Vrstva DataLayer slúži pri správe údajov, ktoré aplikácia využíva počas vykonávania všetkých funkcií. Táto vrstva definuje základné vlastnosti makier, ktoré slúžia pri ich ukladaní a vyhľadávaní do kolekcie, ktorá sa v tejto vrstve inicializuje. Súčasťou tejto vrstvy je aj systém pre definovanie komponentov makier, ktoré sú potrebné pre vykonanie príkazov. Všetky komponenty sú uložené ako enumeráty kvôli možnosti neskôršieho rozšírenia využiteľných vlastností. Táto vrstva obsahuje aj dve statické triedy, ktoré aplikácia využíva pri ukladaní a načítavaní údajov. Tieto triedy vytvárajú systém pre serializáciu a deserializáciu údajov do jazyka XML.

C. Presentation Layer

Vrstva PresentationLayer realizuje grafickú vizualizáciu aplikačného prostredia, ktorá je rozdelená na tri časti:

- MainWizard predstavuje grafickú vizualizáciu počiatočných nastavení, na základe ktorých sa uskutočňuje spojenie so sieťovým zariadením (Obr. 15). Jednotlivé možnosti nastavenia sú automaticky generované od závislosti na spôsobe komunikácie so zariadením.
- Makro slúži ako súbor grafických komponentov, ktoré sa využívajú pri grafickej realizácii makier. Každé makro môže obsahovať ľubovoľný počet komponentov, v závislosti od potreby a zložitosti príkazov, ktoré vykonáva.
- Poslednú časť tvorí trieda MainForm.cs, umiestnená v koreni vrstvy PresentationLayer. MainForm je trieda, ktorá reprezentuje grafickú vizualizáciu hlavného okna celej aplikácie. Zobrazenie tohto okna je podmienené úspešným nadviazaním spojenia so sieťovým zariadením.

D. *SerialportLib a TelnetLib*

Tieto vrstvy sa využívajú pri realizácii spojenia aplikácie so sieťovým zariadením pomocou sériového portu alebo komunikačného protokolu Telnet. Pri výbere komunikačného protokolu bola hlavným kritériom zložitosť, od ktorej sa priamo odvíja možnosť implementovať daný protokol do programovej realizácie. Napriek možnosti použitia knižníc, ktoré sú voľne šíriteľné pre obidva spôsoby pripojenia, bol pri programovej realizácii zvolený variant vytvorenia vlastných komunikačných vrstiev. Obsahom týchto vrstiev sú iba dve triedy, ktoré zabezpečujú nasledujúce funkcie:

Tabuľka 1
TRIEDA COMPORTS.CS

Názov metódy	Funkcia
CloseComPort	Ukončenie spojenia v prípade predošlého používania
ComPorts	Otvorenie spojenia na zvolenom sériovom porte
connect	Implementácie metód triedy Iconnection
disconnect	Implementácie metód triedy Iconnection
sendString	Implementácie metód triedy Iconnection
WriteCompleted	Ukončenie úspešného zápisu cez sériový port
WriteToComPort	Zápis reťazca znakov

Tabuľka 2
TRIEDA TELNETCONNECTION.CS

Názov metódy	Funkcia
ArrangeReplay	Pripájanie identifikátorov paketov pri odosielaní dát
connect	Implementácie metód triedy Iconnection
ConnectCallback	Vytvorenie asynchrónneho spojenia na sokete
disconnect	Implementácie metód triedy Iconnection
DispatchMessage	Transformácia odosielaného reťazca na pole znakov
initialize	Inicializácia premenných potrebných pre spojenie
OnReceivedData	Príjem dát a následné spracovanie
ProcessOption	Identifikácia prijatých paketov
RespondToOption	Výber parametru pri vyjednávaní počas komunikácie
sendString	Odosielanie spracovaných dát

A. *Test Macro*

Možnosť generovania vlastných makier patrí medzi jeden z najväčších prínosov. Vrstva testMacro patrí medzi nepovinné vrstvy samotnej aplikácie. Jej funkciou je priblížiť postupnosť procesov, ktoré sú vykonávané pri vytváraní makier. Obsah vrstvy testMacro tvorí trieda Program.cs, ktorá počas vykonávania svojho kódu vytvára objekty tried, uložených vo vrstve DataLayer. Samotný proces generácie (Fig. 2) pozostáva z troch základných častí:

- vytvorenie nového makra
- uloženie nového makra v kolekcii makier
- definovanie vlastností nového makra

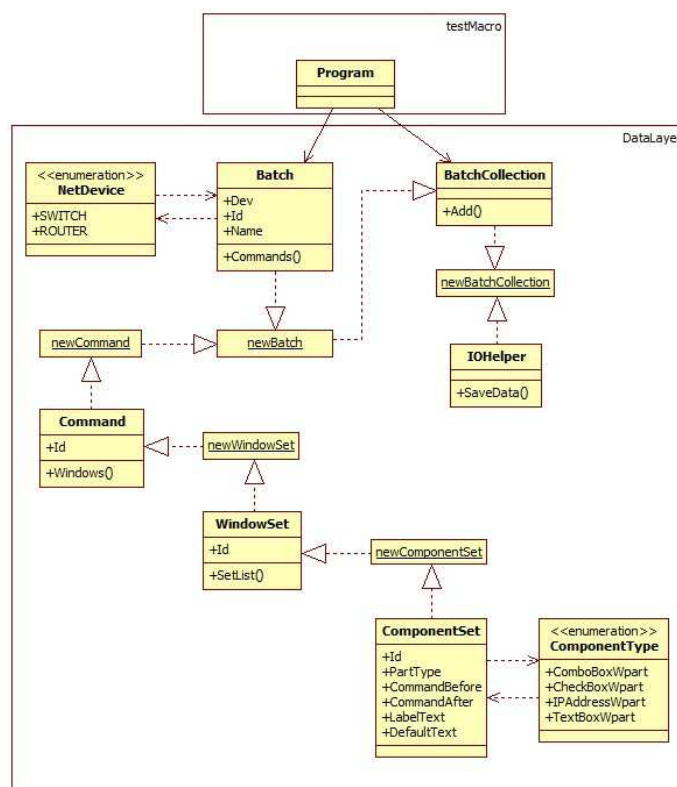


Fig. 2 Proces generácie makier

VI. OVERENIE

Overenie funkčnosti aplikačného prostredia je realizované vytvorením a následným použitím makier na sieťových zariadeniach.

Tieto zariadenia sú zapojené do siete, ktorej topológia je znázornená na Fig. 3. Funkciou týchto makier je konfigurácia protokolu DHCP (Dynamic Host Control Protocol), ktorý slúži na automatické pridelenie IP adres DHCP klientom zo zoznamu dostupných IP adres a pridelenie portov na prepínači do konkrétnych VLAN (Virtual Lan) sietí. Toto overenie je realizované na sieťových zariadeniach od spoločnosti Cisco. Výstupom tohto overenia je výpis aktuálnej konfigurácie zariadení po použití makier.

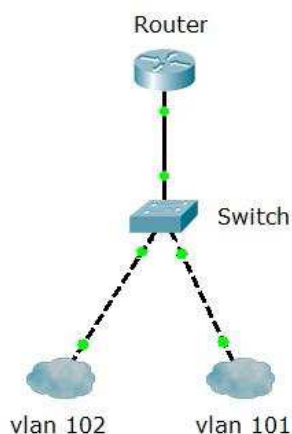


Fig. 3 Topológia konfigurovanej siete

A. Vytvorenie makier

V prípade konfigurácie protokolu DHCP a zaradenia portov do konkrétnej vlan siete na sieťových zariadeniach od spoločnosti Cisco, sú obidva tieto procesy realizované ako séria príkazov, ktoré špecifikujú potrebné parametre. Na generovanie obsahu makier do formátu jazyka XML slúži trieda Program.cs vrstvy testMacro.

B. Použitie makier

Po úspešnom vytvorení makier je možné ich okamžité použitie v aplikačnom prostredí. Proces ich použitia je znázornený na Fig. 4.



Fig. 4 Ukážka použitia makra pre konfiguráciu DHCP a rozsahu protokolov

Návrhu výslednej aplikácie predchádzala analýza existujúcich riešení, ktoré ukázali niektoré nedostatky špecifikované v realizovanom projekte. Výstupom práce je reálna aplikácia, ktorá bola testovaná v rámci rozsiahlej siete. Aplikácia sa ukázala ako výhodná, intuitívna a bezproblémová v spolupráci so sieťovými zariadeniami od spoločnosti Cisco, ktoré sú využívané v testovaných sieťach.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] D.E. Comer, "Internetworking with TCP/IP: Principles, protocols, and architecture." Volume 1, Pearson Prentice Hall, 2006.
- [2] L. M. Surhone, M. T. Timpelton, S. F. Marseken, "OSI Model: Protocol (computing), Open Systems Interconnection, International Organization for Standardization." Betascript Publishers, 2009, 6130324510, 80 pages.
- [3] G. Held, "Ethernet networks: design, implementation, operation, management." British Library Cataloguing in Publication Data, ISBN 0-470-84476-0, 2003.
- [4] Bruce Potter, Bob Fleck, "Securing Wireless Networks, O'Reilly Media, Inc., 2003.
- [5] Rob Allen, Nick Lo, Steven Brown, "Zend Framework in Action", Manning Publication, 1.edition, 2008.
- [6] Andy McCullough, "Designing a wireless network", Syngress, 2001.

Simplification of the Real-Time Network Traffic Monitoring

Adrián PEKÁR, Juraj GIERTL, Martin RÉVÉS, Peter FECILÁK

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

adrian.pekar@gmail.com, juraj.giertl@tuke.sk, martin.reves@tuke.sk,
peter.fecilak@tuke.sk

Abstract – This paper deals with real-time monitoring of network traffic, it introduces the Analyzer-Collector Protocol and its Application Programmable Interface. Particular attention is given to the BasicMeter measuring tool and the explanation, why are tools like BasicMeter so important in modern networks.

Keywords – real-time monitoring, network traffic, QoS, API, ACP, BasicMeter, ACPapi, exporter, collector, analyzer

I. INTRODUCTION

Today's computer networks, which are based on the best-effort traffic model, are designed to transmit video, sound and data (or some combination of these traffic types) at the same time. This kind of transmission is provided by converged networks, which instead of setting own separate links for data and voice, make do with only one converged link. However, in these networks, to get the desired quality of the applications sensible to latency (e.g. Voice over Internet Protocol - VoIP, Video on Demand - VoD), we need to ensure a given level of Quality of Services (QoS).

The phrase QoS is binded to a set of parameters (packet rate, packet loss, one way delay, jitter, round-trip time, etc.), which are used to define the characteristics of computer network traffic. With measuring these parameters, we can ensure the functionality of the above mentioned real-time applications, the fulfillment of conditions specified in the Service Level Agreement (SLA), prevent network attacks, find out dominate traffic sources, etc. Therefore, real-time monitoring of QoS parameters has a significant role in the administration of converged networks and their services.

By means of Analyzer-Collector Protocol [1], the BasicMeter tool [2] was designed for measuring network parameters and their real-time computing. While the implementation of real-time data collection and computation is in some cases really problematic, there was a need for an Application Programmable Interface for servicing the Analyzer-Collector Protocol (ACPapi). ACPapi was designed to simplify the work of the programmers during the development of real-time data dependent applications. Its main purpose is to serve the ACP, which gives opportunities for communication between the parts of the BM tool, whereby the programmers do not have to care for the implementation details of this communication.

In the following sections an alternative method for real-time network traffic monitoring is presented, starting with a simple introduction of the BasicMeter tool; following with a detailed description of the Analyzer-Collector Protocol and its communication principles; up to the presentation of the ACPapi and its basic implementation details.

II. THE BASICMETER MEASURING TOOL

The architecture of the tool was developed in the Computer Network Laboratory at the Technical University of Kosice within three sub-projects:

- BEEM – Basicmeter Exporting and Measuring process
- JXColl – Java XML Collector
- BM Analyzer – BasicMeter Analyzer

Regarding to the roles of these projects (BEEM, JXColl, BM Analyzer), in the following they will be specified as the exporter, collector and analyzer.

The concept of the BasicMeter tool, as described in Figure 1, is in conformance to the IPFIX architecture [3].

Analyzer is a front-end interface, which cares about both, the visualization of the computed data and the control of the architecture's lower parts. However, the analyzer itself is not a part

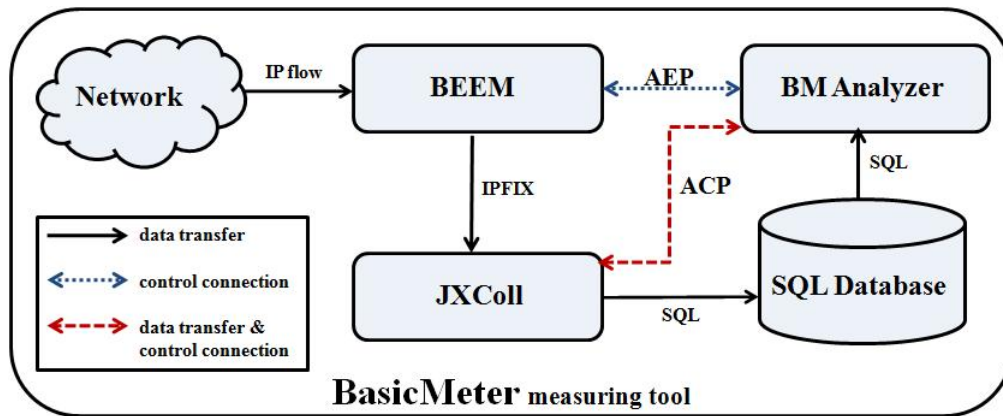


Fig. 1 The architecture of the BasicMeter measuring tool

of the IPFIX architecture, therefore we had to draft and implement our own network protocols for servicing the communication between the analyzer and the other parts of the measuring tool. These protocols are the Analyzer-Collector Protocol (ACP) and the Analyzer-Exporter Protocol (AEP).

AEP was proposed to control the connection between the analyzer and the exporter. With AEP, it is possible to operate the behavior of the export process directly from the customers' application.

In the primary concept of the tool, the measured data should be stored in a database for later analyzes. However, in the case of real-time data computation this method is absolutely inapt. With the growth of data in the database, the processing of the queries is becoming more and more time-consuming. For this reason there was a need for choosing an other way to gain data about network traffic without the ineffective access to the database. Ad hoc was developed the ACP, which allows effective computation of data between the collector and the analyzer. The collector is permanently storing data in the database and if needed, it is simultaneously sending exactly the same data (or a part of it) to the analyzer by the means of ACP.

III. ANALYZER-COLLECTOR PROTOCOL

ACP is a binary, application layer protocol. TCP is used for communication, but it could be easily replaced with any other reliable, connection-oriented protocol. The communication is bi-directional and works on the base of client-server model, while the client side is represented by analyzer and the server side is represented by the collector. As it was mentioned in the previous section, the primary function of the protocol is to gain network traffic data with avoiding the ineffective database queries. To achieve this, besides data, control messages are also transmitted, which are dedicated for format and accuracy checkout of received data by the analyzer.

Communication Principle

The communication is based on sending queries by the analyzer and sending replies with data by the collector. This type of communication could be easily illustrated by a Finite State Machine (FSM) of both sides. The particular communication phases are represented with machine's states at which every message has its own unique ID.

Figure 2a and 2b depict the state transitions of Analyzer-Collector Protocol. State graph in Figure 2a is representing a Mealy type Finite State Machine. Figure 2b is describing a Moore Machine. For the reason that the outputs of the Moore Machine are representing the particular activities of the Analyzer (e.g. receiving data), they are not present in Figure 2b. For easier understanding of state transitions from old states to new states, the messages that invoke these transitions are also included in Table 1 and 2.

ACP is forwarding:

- data (0)
- control messages (1)

Types of the control messages:

- (A) – authentication
- (0) – setting template
- (1) – setting filter
- (2) – suspending data transmission

- (3) – resuming data transmission
- (4) – setting data transmission type
- (5) – acknowledging received data

States of the ACP:

- S_1 – start
- S_2 – waiting
- S_3 – receiving
- S_4 – transmitting
- S_5 – suspended
- S_6 – stop

Messages sent by analyzer:

- A_0 (A_1) – incorrect (correct) authentication data
- 0_0 (0_1) – incorrect (correct) template
- 1_0 (1_1) – incorrect (correct) filter
- 2_0 (2_1) – data transmission unsuccessfully (successfully) suspended
- 3_0 (3_1) – data transmission unsuccessfully (successfully) resumed
- 4_0 (4_1) – not supported (accepted) data transmission type
- 5 – acknowledging the received data

Messages sent by collector:

- 0 – unsuccessful authentication in state S_1 (after established connection), the requested template was rejected in state S_1
- 1 – successful authentication in state S_2 (after established connection), the requested template was accepted in state S_2
- 10 (11) – filter not accepted (accepted)
- 20 (21) – suspension of data transmission was rejected (accepted)
- 30 (31) – resume of data transmission was rejected (accepted)
- 40 (41) – not supported data transmission type (data transmission type is supported and also accepted)

Table 1 State transitions - collector side

OldState	IncomMsgID	OutgoMsgID	NewState
S_1	A_0	0	S_6
	A_1	1	S_2
S_2	0_0	0	S_2
	0_1	1	S_4
S_4	0_0	0	S_4
	0_1	1	S_4
	1_0	10	S_4
	1_1	11	S_4
	2_0	20	S_4
	2_1	21	S_5
	4_0	40	S_4
	4_1	41	S_4
S_5	5	-	S_4
	3_0	30	S_5
S_6	3_1	31	S_4
	-	-	-

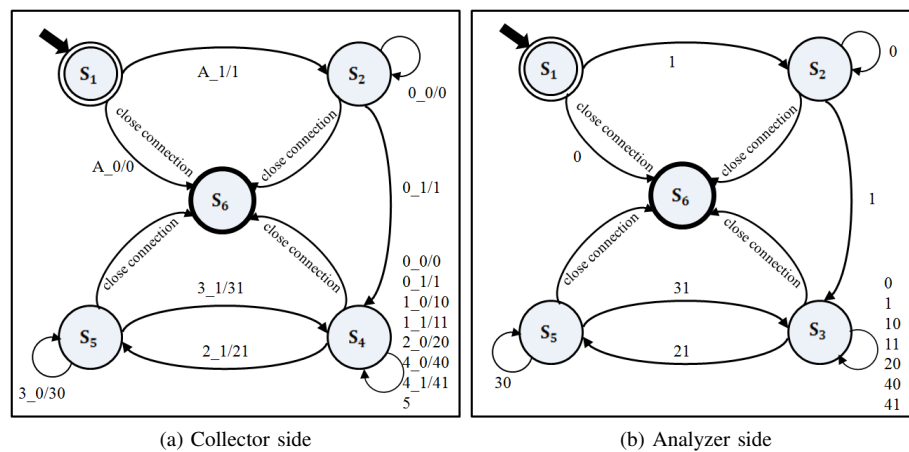
Table 2 State transitions - analyzer side

OldState	OutgoMsgID	IncomMsgID	NewState
S_1	A_0	0	S_6
	A_1	1	S_2
S_2	0_0	0	S_2
	0_1	1	S_3
S_3	0_0	0	S_3
	0_1	1	S_3
	1_0	10	S_3
	1_1	11	S_3
	2_0	20	S_3
	2_1	21	S_5
	4_0	40	S_3
	4_1	41	S_3
S_5	5	-	S_3
	3_0	30	S_5
S_6	3_1	31	S_3
	-	-	-

Description of the Communication

The FSM graphs (Figure 2a and 2b) and the tables (Table 1 and 2) clearly describe the communication via ACP's control messages between the collector and the analyzer. Control messages sent by collector act only as replies to requests generated by the analyzer. However data transmission is running through the same connection, from one point of view it represents the collector's special autonomous activity, which is initiated by the receive of IP flow information from the exporter(s). Regarding to desired simplicity and efficiency of the protocol, the header of the control messages is differing from the header of the data in one byte (value 1 for the control message and value 0 for data).

At the initial state, the collector is awaiting the analyzer's connection at a pre-agreed port (default value for collector is 2138). If there are no problems during the connection establishment, both sides traverse to their S_1 states. Before the bi-directional communication can be opened, the analyzer has to send accurate authentication information to the collector. Therefore as soon as the connection is established, the analyzer sends its login and MD5 password ciphers to the collector. Successful authentication brings both sides to their S_2 states, but if the authentication



is unsuccessful, the collector terminates the connection and the analyzer have to reconnect for another try.

In S_2 state, the collector is awaiting the analyzer's setting of the template message (message 0), which informs the collector about the desired format of the data. At the current state of the ACP, templates are sent as arrays containing the IDs of the information elements, but in the future we count with replacing the arrays with a binary description of these elements. After receiving the template the collector replies with its acceptance (message 1) or rejection (message 0). If the template was rejected, the analyzer has to send another template. During this phase, unless the collector accepts the template and replies with accept message, the communication is suspended. If the template is accepted, the collector traverses to its S_4 state and starts with the data transmission in the format of the accepted template. On the other hand, after the analyzer receives the template accepted message, it immediately traverses to its S_3 state and starts to receive the data from the collector. In this phase the template could be changed at any time, and if it is accepted by the collector, the data would be sent in the format of the new template. Real data, which were received/sent by the analyzer/collector, besides control messages are sent only in states S_3 and S_4 .

By default, collector exports all the traffic information obtained from the exporter(s). However, in a case of need, by setting the filter (message 1) it is also possible to receive only specific information. This is an optional step which is accepted by the collector only in its S_4 state, while only one filter can be set in a given time. Current flow filtering is supported upon the next criteria:

- IP address of the measuring point.
 - Format of the criterion: (A.B.C.D/n), where $0 \leq n \leq 32$;
- Source and destination IP address/mask – with the combination of IP address and mask it is possible to specify not only one specific host, but also a whole subnet. It is also possible to specify more subnets or hosts.
 - Format of the criterion: (A.B.C.D/n; A.B.C.D/n - A.B.C.D/n; A.B.C.D/n , A.B.C.D/n), where $0 \leq n \leq 32$;
- Source and destination port – this criterion is allowing to specify upon the transport layer port number also a specific type of network traffic. It is also possible to set more values and range of values.
 - Format of the criterion: (N; N - N; N, N), where N can be a well known port, registered port and dynamic or private port;
- Protocol – currently the recognition is supported by numbers and like in the case of port criterion, it is possible to set random combination of values and range of values.
 - Format of the criterion: (N; N - N; N, N), where N can be any of the known protocol numbers;

The collector is using message 11 to confirm the accepted filter, and message 10 to inform about the filters rejection. The filter can be canceled or replaced with empty filter rule in the collectors S_4 state.

Control message with ID 4 is used to set the type of the data transmission, which brings more customizability and convenience for the developers using the ACPapi to communicate with the collector. Currently ACP supports two types of data transmission:

- sending the data by collector in N-tuples (message 1), where N is the number of the information elements in obtained template
- sending the data by collector one-by-one (message 2)

By default, the collector sends the data in N-tuples, ergo during one cycle of data reception the analyzer receives N values. This type of data transmission can be changed optionally in the collector's S_4 state. With second type of data transmission the analyzer receives one value during one data reception cycle.

There are another two control messages, which are dedicated for suspending and resuming the data transmission. In S_4 state the collector receives the data transmission suspension message (message 2). The receive of this message causes that the collector sends every data from the currently processed packet, suspends the data transmission and replies with message ID 21, which informs the analyzer about the transfer suspension. Both sides of the communication traverse to their S_5 states. From here on, the analyzer won't receive any data. The only method to continue with data transmission is to send the collector the data transfer resume message (message 3), which brings both sides to their previous states. After this message, the collector replies with acceptance message (ID 31) and starts to transmit data from the actual packet received from the exporter(s). Before this message, the analyzer will not continue with receiving data. In general, there is no reason for the collector to discard the transfer suspension (message 20) or resume (message 30) request.

The last control message is destined for synchronizing the data sending/receive by the collector/analyzer. When the collector sends the data, it awaits acceptance from the other side. Unless this acceptance is obtained, the collector will not send any other data. With this method we can prevent receiving inappropriate data (from experience, without this control message the collector sent in some cases concatenated values, which could easily led to data computation errors).

IV. APPLICATION PROGRAMMABLE INTERFACE FOR SERVICING THE ANALYZER-COLLECTOR PROTOCOL (ACPAPI)

The language chosen was the Java language, which main advantages are in modularity, stability, perspicuity of the source codes, ability of the documentation auto-generation, etc. Except these advantages, Java was chosen also for the reason that the collector was written in Java too. Some of the methods will be used by both parts of the communication, hence of that the implementation in the same language is more preferable to the implementation in different languages. As mentioned before, the ACPapi has to be usable in the case, when the programmers do not want to waste time with designing and implementing their own classes or methods for the communication based on ACP. For this reason, the whole ACPapi was designed as an easy to understand, implement and use API, besides during its creation was also given a tone to its exact documentation and to the annotation of the source code.

An API can be written by means of classes or interfaces. According to [5] the most profound feature of Java interfaces is multiple inheritance. With multiple inheritance, you need just one object to implement an unlimited number of interfaces from an API. In the case of classes, you would need to create one subclass for each API class, and if these classes are related to one another, join their instances by means of delegation. This can significantly increase the amount of occupied memory.

On the basis of this knowledge the ACPapi was written by the means of interfaces, which main advantage expresses in the performance of the final application using ACPapi. On the contrary, with interfaces we also have to count with adding new methods only by filling in the source code.

Because of the above mentioned dual-usability of some methods, the ACPapi was divided into two packages:

- ACPapi package – contains the API interface itself with declarations of the methods, classes with definitions of the methods and other classes
- commonly used package – contains commonly used classes of the analyzer and the collector (classes for working with templates and filters)

Functions defined by Analyzer-Collector Protocol (sending authentication data, templates, filters, etc.) are named upon their functionality. These functions with short description (detailed description could be found in the *Communication Description* subsection) are the following:

- getMd5Digest – method for cipher login and password data
- connectToCollector – method for establishing the connection with the collector
- sendTemplate – method for setting the desired format of the received data
- sendFilter – method for setting filter rules
- SendPause, sendUnPause – methods for suspending or resuming the data transmission
- readCollectorAnswers – method for computing the data sent by the collector

- quit – method for correct termination of the ACP process
- sendTransferType – method for setting the type of the data transmission

Main function of the ACPapi is its data providing for almost real-time computation. The word 'almost' points to the fact, that receiving data by ACPapi brings a certain delay with itself, which results from the data computation by the exporter, their transmission over the network, further computation by the collector and sending them by ACP. Despite of that, from the point of real-time monitoring of network traffic, this delay is absolutely affordable.

Obtaining the data with ACPapi could be implemented in some ways. One method could be storing the data in some data structure (e.g. arrays). Another method could be reading the incoming messages and data in a cycle. Even though that the first method is more comfortable, from the point of real-time data computation the second method was chosen. By reason of proper customizability this method is not part of the ACPapi and is provided only as a design pattern (the source code of the design pattern could be found at the API's website - <http://wiki.cnl.sk/Monica/ACPapi>). Of course, the data could be obtained by other methods, but all of them need a thorough study of the ACP and its API.

Forasmuch as the protocol is defined for communication of both sides, it has to be supported by the collector too. However, a more detailed description of ACP support by collector is out of the scope of this paper. In a case of need, one can find more information about this on the projects website.

V. CONCLUSION

Actual researches and developments on the field of monitoring tools are leading to the improvement of their ability of measuring the main characteristics of network traffic. Besides the main effort is put on the maximization of their real-time data monitoring ability. An important feature of these tools is also their flexibility and ability of supporting a wide scale of complex applications. The BasicMeter measuring tool is focusing to the fulfillment of these requirements, which above all is served for measuring QoS parameters.

In the previous sections the main concept of the BasicMeter tool and its method of getting data for (almost) real-time computation by ACPapi was presented. The Application Programmable Interface for servicing the Analyzer-Collector Protocol fulfilled its goals. Its implementation brought its awaited results. Thanks to the proposed design pattern and some improvements to the ACP (setting transfer type or acknowledging the received data) the ACPapi gives reliable, easy to use and implement application interface for the developers of the BasicMeter measuring tool.

Future work should be aimed at the replacement of the actual design pattern of ACPapi and to the optimization of the templates by replacing the arrays with a common binary description of the information elements.

ACKNOWLEDGMENT

This work is the result of the project implementation: Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] D. Calvetti and L. Reichel, "On the evaluation of polynomial coefficients," *Numerical Algorithms*, vol. 33, pp. 153–161, Aug. 2003.
- [2] G. T. Herman, *Image reconstruction from projections*. Berlin, Germany: Springer-Verlag, Oct. 1980.
- [3] E. M. Johansson and J. E. Mast, "Three-dimensional ground-penetrating radar imaging using synthetic aperture time-domain focusing," *Advanced Microwave and Millimeter Wave Detectors, Proceedings of SPIE*, vol. 2275, pp. 205–214, Sep. 1994.
- [4] D. Miller, M. Oristaglio, and G. Beylkin, "A new slant on seismic imaging," *Migration and integral geometry: Geophysics*, vol. 52, pp. 943–964, July 1987.
- [5] V. Y. Pan, "Solving a polynomial equation: some history and recent progress," *SIAM Review*, vol. 39, pp. 187–220, June 1997.

Adaptive Platform for Video Streaming to Mobile Devices

Miroslav MICHALKO, Miroslav BIŇAS, František JAKAB

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

miroslav.michalko@tuke.sk, miroslav.binas@tuke.sk, frantisek.jakab@tuke.sk

Abstract — Fast detection and recovery of transmission failures and anomalies in data transfer over wireless networks are crucial for multimedia content delivery. To overcome problems appearing in constantly changing environment several techniques were developed and implemented. Due to better utilization of the available bandwidth video streaming servers offers adaptive streaming specially designed for wireless networks. This article deals with concept of platform for adaptive streaming to wireless end devices with consideration of conditions on transmission medium and quality of received content. Paper presents a feedback based mechanism that provides video adaptation based on networks parameters with usage of fuzzy based techniques. Results presented here are based on research in Computer Networks Laboratory at Dept. of Computers and Informatics.

Keywords — mobile devices, network, streaming, video, wireless

I. INTRODUCTION

With the advance of modern technology in wired and wireless networks there is higher need for multimedia content delivery everywhere. Mobile devices are capable to process and display video in high qualities what brings new challenges to network providers to guarantee satisfactory services. As mobile and wireless networks conditions are pretty hard to predict there is need to improve existing protocols to better face problems like packet loss, delay and jitter or service interruption due to vertical handover failures. During last year many solution were offered but demand for higher quality and better responsiveness makes this solutions only partially usable. Most of multimedia applications are highly sensitive to delay and loss. It is acceptable to tolerate certain amount of loss but depends what type of data are going to be lost, for example in video streaming loss of I-frame causes serious degradation visible to end user.

Market shows need to develop special per device mechanism which will detect device parameters and accommodate video resolution to maximum possible quality. There is also need to follow available network conditions and content priority. This is mostly done by operator or administrator setup and cannot be altered by clients or content provider. By providing feedbacks to streaming server in real time best quality could be achieved. Of course this special treatment requires more utilization on server side. In order to improve streaming functionality available techniques needs to be used and altered to mobile streaming needs.

Experiments within Computer Networks Laboratory were done using own developed Videoserver platform based on flash streaming server with mobile extension and open source streaming server like FFserver [1] and VLC [2] modified as needed. Also several simulation tools were used. As primary encoder H.264 presented by Fraunhofer institute is used [3].

II. QUALITY OF SERVICES IN WIRELESS LAN

A. Requirements

In order to deliver distinct content over network, requirements for quality of services per each content type needs to be defined. For instance, for web data content download high reliability is required but delay is not some crucial parameter. In voice over IP medium reliability is required but delay and jitter have to be minimal as possible (for jitter less normally less than 90 miliseconds). Table 1 shows applications and corresponding QoS requirements.

Table 1
 APPLICATIONS AND CORRESPONDING QOS REQUIREMENTS [4]

Application	Reliability	Delay	Jitter	Bandwidth
E-Mail	High	Low	Low	Low
File Transfer Protocol (FTP)	High	Low	Low	Medium
Web access	High	Medium	Low	Medium
Remote log-in	High	Medium	Medium	Low
Streaming audio	Low	Low	High	Medium
Streaming video	Low	Low	High	High
VoIP	Medium	High	High	Low
Videoconferencing	Low	High	High	High
Mobile TV	High	High	High	High

B. Multimedia content encoding

Actually available coding algorithms and schemes offer mobile encoding features. Main parameters for mobile streaming are low complexity for decoding, reliable transmission over lossy channel and high efficiency in compression. Compression is the use of coding techniques to reduce the amount of data used to convey information. Compression algorithm can reduce data by quantizing the image more crudely, either by [5]:

- Using fewer bits to represent each color.
- Reducing the number of pixels in the image.
- Reducing the amount the value of a pixel can change compared to its nearest neighbors.
- Reducing the number of frames per second.

For streaming scenarios video is encoded at server side and then delivered to mobile device where decoding process is in progress. Decoding consumes a lot of battery power so a provider prefers lower video quality in order to support longer transmissions. Also it helps to overcome problem with variety of display resolutions at mobile devices, smaller quality will work also on older devices.

International Standards Organization (ISO) Motion Picture Experts Groups (MPEG) and International Telecommunication Union Standardization Sector (ITU-T) recommends for mobile streaming H.263 and H.264. Best known implementations uses H.264 named as AVC or MPEG-4 part 10. Fig. 1 shows typical H.264 video codec scheme.

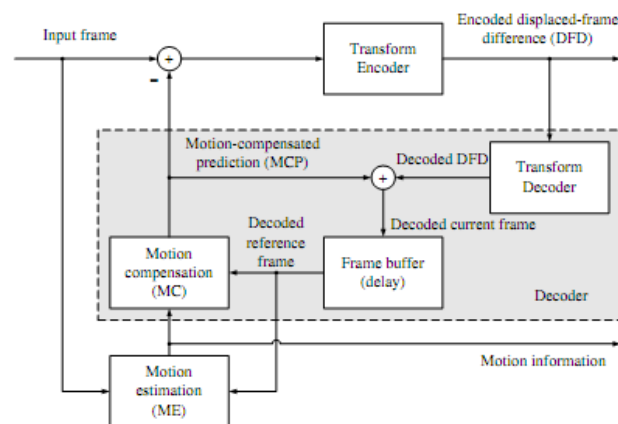


Fig. 1 Scheme of typical video codec [6]

Intraframe compression techniques are used to reduce redundancy within one frame, interframe compression exploits the correlation among subsequent frames to reduce amount of data to be transmitted over wireless channel. In typical implementations such as MPEG or H.263 and H.264, the encoder is five to ten times more complex than the decoder [7]. For testing purposes Fraunhofer Institute shares H.264/AVC reference software [3] which is great for testing implementations and allows changes in many key parameters of coding.

III. STREAMING ARCHITECTURE CONCEPT

Based on study on existing streaming technologies, sound compressions and techniques for transmission in mobile and wireless networks new streaming architecture concept was created. The architecture can be divided into four separate functional blocks:

- Selection of the format of source,
- Video and audio encoder,
- Decoder of transmission (end mobile device),
- Evaluation of transmission line parameters and quality of content.

The selection of a suitable format is done by Manager of sources. Based on identification of end device Manager will choose appropriate format of requested content. Selected video and audio source serves as input to the encoder. The choice of encoder is driven by the content type selected or by the administrator setting. Input encoder parameters for video and audio encoding are set by Manager of coding which evaluates inputs from the Analyzer of transmission line state and the Analyzer of A/V quality. The output stream is combined and synchronized in the MUX and then transmitted over the wireless or mobile channel. Transmission state analysis will work on any type of network it is able to carry for example standard protocol RTP/RTCP.

Device receiving the data stream via video and audio decoder provides correct display of received flows on the user's screen. Built-in analyzer and flow identifier provide feedback to the streaming server, which is serves for better adaptation of broadcasted content to the end device. Transmitted media stream is analyzed also in terms of picture and sound quality. Objective methods are available, such as PSNR for video and PEAQ for audio. Output quality analysis can be used as input for coders to avoid excessive degradation of image and sound transmission.

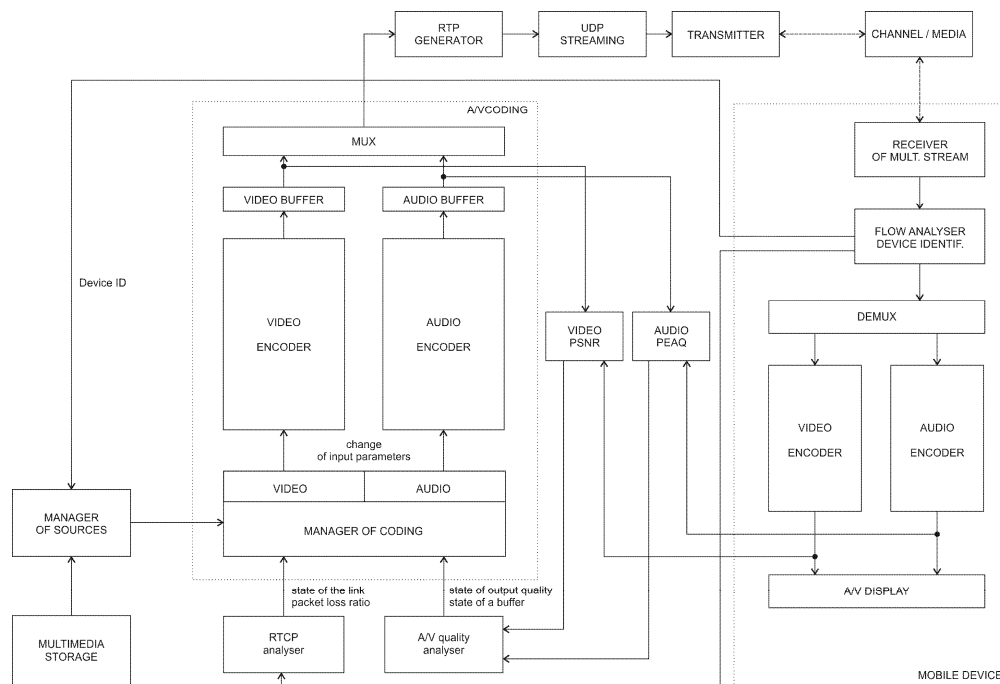


Fig. 2 Streaming architecture concept with enhanced feedback support

The versatility of the proposed architecture is due to possibility to use encoder of a content based on the type of end device. Manager of sources has an ability to select not only format of video and audio but also to select best encoder for it. Feedback obtained from the end mobile device makes it possible to adapt the content to real conditions and device in terms of quality of the image and audio, thus there is no commitment to particular type of mobile device.

IV. ADAPTIVE STREAMING BASED OF FUZZY CONTROLLER

Adaptive streaming can be implemented based on more publicly available models. This work is based on research presented by authors Rezaei, Hannuksela and Gabbouj named as Low-complexity fuzzy video rate controller for streaming [8]. This model was adopted, modified and locally implemented. This model was several times updated and published, but for reason of

this work older model was selected for its properties in video quality encoder part. Based on the need to take into account the transmission conditions in the network reference model was extended with feedback system in order to evaluate network traffic in real time. It is done by new functional blocks, RTCP and Fuzzy analyser of transmission link. The main task of the RTCP analyser is to assess the current parameters of transmission link, namely the degree of packet loss and congestion at the state line. The role of fuzzy link analyser is to evaluate and influence the process of encoding video, which will provide better video output parameters given the state line between streaming server and client end.

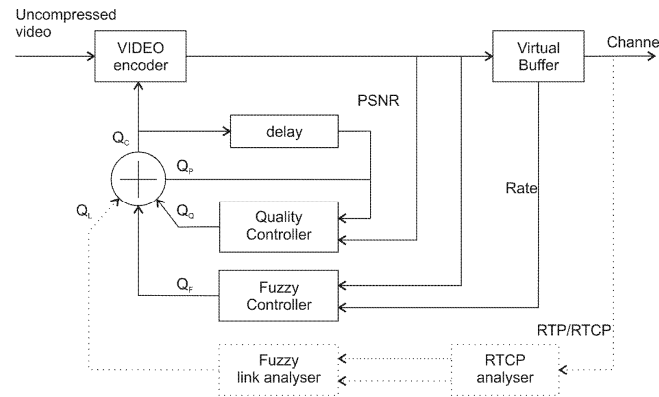


Fig. 3 Block diagram of the fuzzy video rate controller [8] enhanced by Fuzzy link analyser (dotted line)

RTCP analyzer based on RR packets periodically calculates the value of packet loss and link congestion. Similar parameters were presented in Advis algorithm [9] but this model presents expansion and brings different way of integration to fuzzy video rate controller. Feedback mechanism added to reference model is shown on Fig. 4.

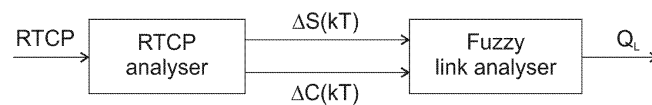


Fig. 4 RTCP analyser and Fuzzy link analyser

Linguistic variable ΔS indicates increasing or decreasing trend of packet loss in transmission channel. Linguistic variable ΔC indicates increasing or decreasing trend of congestion in transmission channel based on the RR evaluation in RTCP packet ECN bit congestion indicator can be taken into account. Range of trend is the range of $\langle -1, 1 \rangle$

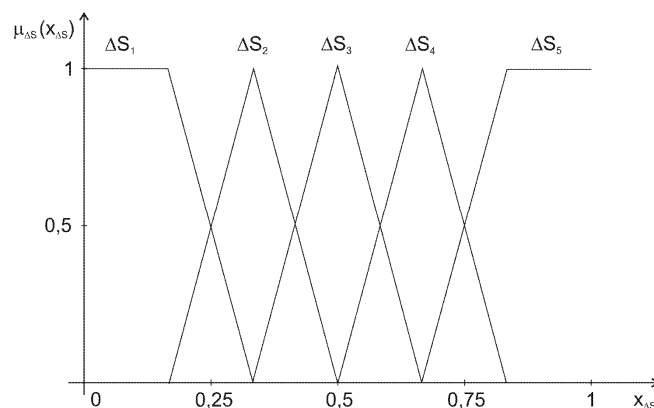


Fig. 5 Membership function of the linguistic variables, $n=5$, ΔS

The proposed solution is based on fuzzy controller of Sugeno-type because it is easy to implement in real terms and the output of the process of inference can be given directly to a sharp value. Membership function of the linguistic variables is presented on Fig. 5 and Fig. 6. Linguistic table with empirically given values is shown at Tab. 2.

Output of Fuzzy link analyser is linked as partial source of information to video encoder and influence quality of video output delivered to end device [10].

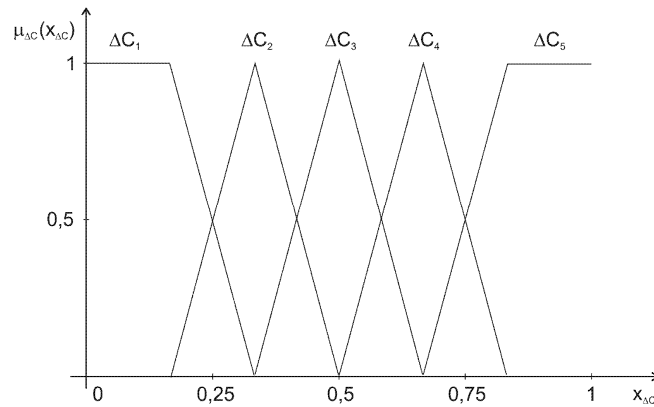


Fig. 6 Membership function of the linguistic variables, $n=5$, ΔC

Table 2
 LINGUISTIC TABLE BASED ON EMPIRICALLY MEASURED VALUES [4]

$Q_L(kT)$		$\Delta S(kT)$				
		ΔS_1	ΔS_2	ΔS_3	ΔS_4	ΔS_5
$\Delta C(kT)$	ΔC_1	1,5	1,3	1,1	1,0	0,9
	ΔC_2	1,3	1,2	1,0	0,9	0,9
	ΔC_3	1,1	1,0	1,0	0,8	0,8
	ΔC_4	1,1	0,9	0,9	0,6	0,7
	ΔC_5	0,7	0,8	0,9	0,7	0,5

The proposed solution has been programmed and verified experimentally. For implementation the source code of H.264/AVC encoder [3] was used and rate model Fuzzy Controller for video streaming implemented locally [8]. Own income is in the form of fuzzy logic and RTCP analyser which was implemented in C++. The proposed experiments were partially verified before their practical implementation in the program Matlab. Video quality is calculated using PSNR calculator.

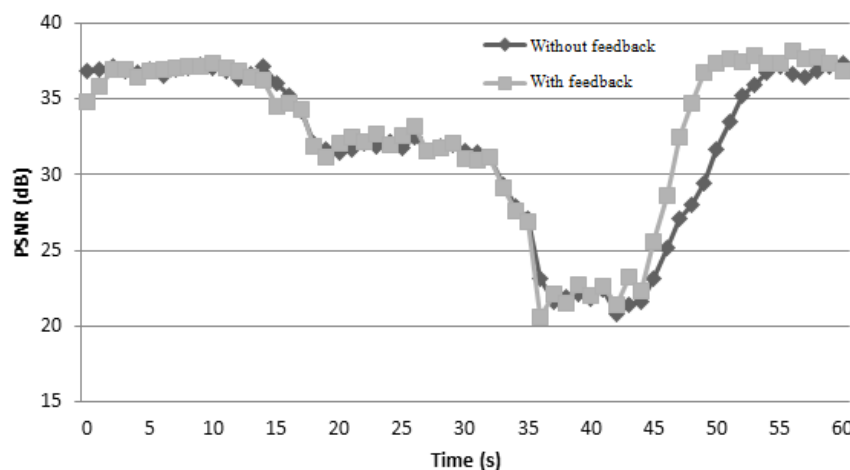


Fig. 7 PSNR in time, 0%, 5% a 10% packet loss, faster recovery to better picture quality

V. CONCLUSION

This paper presents concept of adaptive streaming to mobile devices in constantly changing network environment. Practical implementation actually proves the validity of the use of fuzzy controller for evaluating video quality, which resulted in higher levels of PSNR. Extending the

reference model of the fuzzy line analyser has brought a slight improvement of the quality of the video parameters. Main advantage is simple implementation of the scheme. Disadvantages are high coding resources and limited video rates in real-time coding. Future development will focus on streaming server based on Wyner-Ziv coding scheme and UI based learning and evaluation of network behavior.

ACKNOWLEDGMENT

This work was supported by the Slovak Cultural and Educational Grant Agency of Ministry of Education of Slovak Republic (KEGA) under the contract No. 3/7245/09 (60%). This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research& Development Operational Program funded by the ERDF (40%).

REFERENCES

- [1] FFmpeg /FFserver, FFmpeg foundation, GNU/GPL software. cit. [2011-01-11] Available at internet: <<http://www.ffmpeg.org/legal.html>>, 2011.
- [2] VideoLAN VLC, École Centrale Paris, GNU/GPL software. cit. [2011-01-13] Available at internet: <<http://www.videolan.org/>>, 2011.
- [3] H.264/AVC Reference Software, Telecommunications, Fraunhofer Institute. cit. [2010-08-11] Available at internet: <<http://iphome.hhi.de/suehring/tml/>> Berlin, 2010.
- [4] TAO, Z., et col.. Multimedia Quality-of-Service Support in IEEE 802.11 Standards. [aut.] Yan ZHANG, et col. Broadband mobile multimedia. Boca Raton: Auerbach Publications, 2008.
- [5] Topic, M., Streaming Media Demystified. McGraw-Hill Telecom. 2002, pp. 61, ISBN 978-0071388771.
- [6] All-Mualla, M., Canagarajah, C., Bull D., Video Coding for Mobile Communications: Efficiency, Complexity and Resilience (Signal Processing and its Applications). Academic Press; 1 edition, 2002, pp. 4, ISBN 978-0120530793.
- [7] Girod, B., Aaron, S., Rane, S., Rebollo-Monedero, D., Distributed video coding. Proceeding of the IEEE, 93/1, 2005, pp. 71-82.
- [8] Rezaei, M. - Hannuksela, M. - Gabbouj, M. Low-Complexity Fuzzy Video Rate Controller for Streaming. ICASSP 2006 Proceedings. 2006 IEEE International Conference. 2006, Zv. 2, ISSN 1520-6149.
- [9] Antoniou, P. - Vassiliou, V. - Pitsillides, A. ADIVIS: A Novel Adaptive Algorithm for Video Streaming over the Internet. Athens: PIMRC 2007, 2007. ISBN: 978-1-4244-1144-3.
- [10] Michalko, M. Creation of videostreamings for mobile devices with required. Dissertation. Technical University of Kosice, 2010.

Distributed Emulation Using GPGPU

Peter JAKUBČO, Eva DANKOVÁ

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

peter.jakubco@tuke.sk, eva.dankova@tuke.sk

Abstract – This paper presents new idea about constructing an emulator of a processor using GPGPU technology. The key concept is to utilize graphics card as heterogeneous computing power source. The emulator consists of two parts that communicate together in distributed manner. There are also some real experimental results that show that this idea is worth of evolving.

Keywords – emulation, GPGPU, OpenCL, stream processing

I. INTRODUCTION

Computer emulators have long history. In science, mostly simulators are used. However, the difference is not big and the boundaries are not well-defined; it can be said that emulators deals with some native properties of a host computer, since simulators imitate each part.

The most used techniques for both simulation/emulation of a processor are interpretation and binary translation. The binary translation is more powerful than interpretation. In our experiment implementing Random Access Machine (RAM) emulator we achieved with interpretation about 9.2 MIPS, with dynamic translation we achieved about 97 MIPS.

Since even the best emulation techniques have their power boundaries when emulating too complex computers, the aim is to overcome classic single-platform approach and to build new emulation algorithms and their implementations, based on a high level of parallelism, concurrency and heterogeneity.

The aim is to utilize stream processing, enabled by GPUs, for processor emulation. The goal is to disburden the host CPU of performing computing jobs within emulation, so more power would be available either for more sophisticated emulation, or for another activities.

II. GPGPU AND EMULATION?

The GPGPU is related to stream processing very closely [1]. However, stream processing is a computing model very suitable for computational tasks. According to [2], when the stream processing is applied on generic processors such as standard CPU, only a 1.5x speedup can be reached. By contrast, ad-hoc stream processors easily reach over 10x performance, mainly attributed to the more efficient memory access and higher levels of parallel processing.

In practical speaking, within GPUs, the performance improvement is achieved by minimizing inputs and outputs and a try to make each computation local, spending there that much time as it is possible. We would be able to say that it is worth of developing a GPGPU emulator, if the emulator can satisfy all the three conditions. The summarization:

- 1) Good-performance GPGPU applications are local;
- 2) Good-performance GPGPU applications minimize inputs and outputs;
- 3) Good-performance GPGPU applications work long time (longer that the time needed for their enqueueing and loading);

What we know about CPUs¹? For now, we know only that CPU executes instructions in a loop. That satisfies the condition no. 3, because the loop is infinite.

In order to satisfy condition no. 1, it must be possible to construct a single-kernel CPU emulator. GPUs have relative small memory for storing kernels. But this memory (usually 256 MB) is not that small for keeping the whole emulator. It is assumed (according to their current size) that emulators GPGPU machine code would not take more than a few KBs of memory. And finally, our experiment implementation of RAM emulator proves that. This satisfies the condition no. 1.

For satisfying the condition no. 2, deeper considerations must be taken into account. The number of inputs and outputs in a CPU depends on several conditions:

¹For simplicity, only von-Neumann computers are considered.

- The program that is executed on CPU;
- The events that might happen outside of the CPU (e.g. interrupts);
- The events that the CPU might generate.

The inputs are data that rest in the environment (e.g. peripheral devices). Outputs are data that are generated inside of a CPU and need to be transferred to output device.

Some CPUs ask for input manually (using so-called polling, or busy waiting) – in that point of view, CPU emulators generate event requests for inputs and waits for it. Another option is that devices request for service by interrupt requests (IRQs). CPU (using spooling) reacts to these requests only, by what the CPU is disburden of waiting for event completion. The model of proposed structure in Fig. 2 explains the idea using GPGPU.

Some may think that in order to satisfy the condition no. 2, it would be needed to know the frequency of input/output instructions in ordinary applications and if the number is low, that it is worth to expect that inputs/outputs are not intensive. According to [3], up to 90.5% of Windows kernel and applications instructions do not represent input, output, or interrupt instructions. The memory instructions are not taken as input/output instructions, because, as it will be explained below, memory will be emulated in the graphics card, too, therefore all memory operations are considered to be local.

Seeing that we write to keyboard and move the mouse very often, it just cannot be 0.5% of all operations. The answer is: this measurement measure only static instructions occurrence. It is not possible to determine how often particular instructions are performed, because this depends only on the program that is executed on CPU. Therefore, the condition no. 2 remains open.

III. PROPOSED EMULATOR STRUCTURE

The structure of the distributed emulator is depicted in Fig. 1.

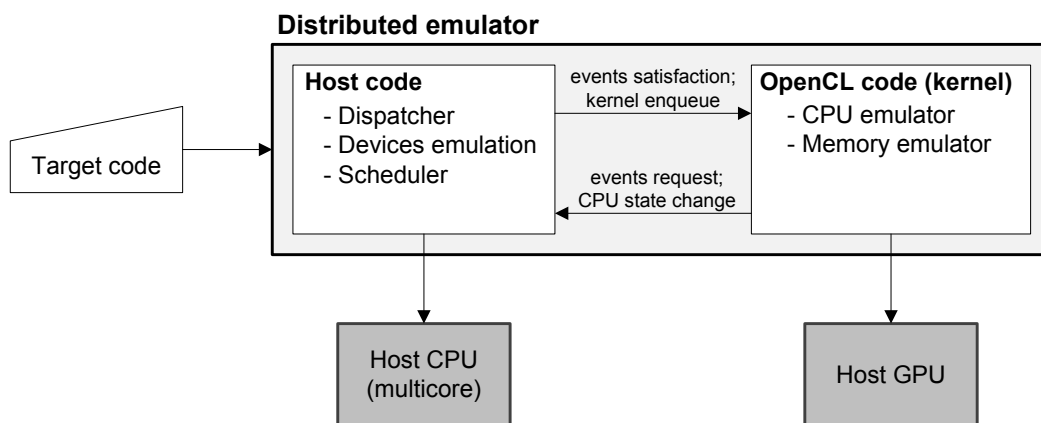


Fig. 1 Structure of distributed computer emulator using GPGPU

The emulator consists of two main parts that communicate in the distributed manner. The idea is that GPU would execute CPU emulator that would be implemented as stream kernel. The host part will be acting like scheduler or dispatcher that would catch events either from the CPU emulator (from GPU), or from other emulated components.

The communication media between CPU and GPU is PCI Express channel. Transfer performance is CPU and chipset dependent. Transfers from the system to the GPU compute device are done either by the command processor or by the DMA engine. The GPU compute device also can read and write system memory directly from the compute unit through kernel instructions over the PCIe bus.

The host part will ensure the communication among user, CPU emulator and other emulated components. The example of events requesting and satisfaction is shown in Fig. 2.

IV. APPROPRIATE STREAMING LANGUAGE

Currently, three options can be considered: CUDA, OpenCL [4] and DirectCompute. For further work, we chose the OpenCL language. The reason is that it is an open standard and can work with CPUs, if GPUs are not available. OpenCL is supported by many vendors and there exist free SDK. The standard is still under development, so today has several restrictions. The most critical one is that within ATI SDK [5], independent kernels cannot be executed concurrently. This restricts the parallelism usable in the emulator. In the present time, CUDA supports maximum 4 independent kernels that can run concurrently (on Fermi cards).

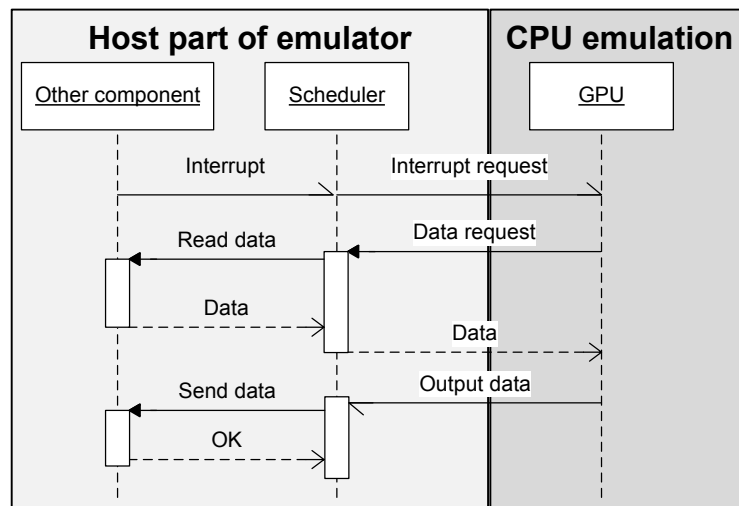


Fig. 2 The example communication between host and GPU – events

V. EMULATION TECHNIQUE WITHIN GPGPU

GPGPU is a technique that uses shader languages for general-purpose computation. However, shading languages are enclosed (their specification is not public) and they differ among GPU models. The access is done directly via drivers, or via higher level languages (OpenCL, CUDA, DirectCompute; older approaches include OpenGL and DirectX).

The portability problems connected with closed shading languages specifications (assemblers) make impossible of performing other technique than interpretation. Another critical problem is, that kernels cannot use function pointers. Therefore, only decode-and-dispatch interpreters is possible to construct.

VI. CONCLUSION

The usage of simulators and emulators is indisputable – it is the validation model. The advantages of full system simulation/emulation are several. Software development can take place before the hardware is ready, therefore validation is faster. Validation is less costly and faster because many engineers can run validation tests on a PC instead of sharing a few HW prototypes. Some things can be done with simulation that can hardly be done with hardware, like verifying correct hardware initialization, simulating defective hardware, internal observations, etc. And the last thing, simulation tools can be connected with formal methods tools.

Stream processing and GPGPU brought a new potential of making high-performance software. They allow it by limited form of parallelism. Applications thus are divided into two parts – host part and client part, where host part is implemented in any classic programming language as we know today and it is executed on CPU. Client part is written in specialized stream language which extends the language of parallelism and on the other side, has some restrictions. The client part is executed on GPU or another compatible device, which disburdens the CPU of computational work that can be parallelized.

This work points at utilizing this new approach in another direction, as was suggested. The intent is to develop simulators/emulators of CPUs that will run as client parts of the application. However, some restrictions are put to the emulator, caused by restrictions of the language. The last chapter proves that despite of all issues, it is worth to study and discover new possibilities of this approach. The following section discusses all objectives for further doctor thesis.

There are still some open issues that were not analyzed, but they require more research. The most important are:

- Preservation of the CPU speed – cycle timing accuracy;
- Device emulation;
- Effective communication between devices and CPU.

ACKNOWLEDGMENT

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] S. Venkatasubramanian, "The graphics card as a streaming computer," *CoRR*, vol. cs.GR/0310002, 2003.
- [2] B. Khailany, W. J. Dally, S. Rixner, U. J. Kapasi, J. D. Owens, and B. Towles, "Exploring the vlsi scalability of stream processors," in *Proceedings of the 9th International Symposium on High-Performance Computer Architecture*, ser. HPCA '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 153–. [Online]. Available: <http://portal.acm.org/citation.cfm?id=822080.822826>
- [3] I.-J. Huang and T.-C. Peng, "Analysis of x86 instruction set usage for dos/windows applications and its implication on superscalar design," in *Proceedings of the International Conference on Computer Design*. Washington, DC, USA: IEEE Computer Society, 1998, pp. 566–. [Online]. Available: <http://portal.acm.org/citation.cfm?id=846214.846668>
- [4] *The OpenCL Specification*, 2010, revision 36. Available at <http://www.khronos.org/registry/cl/specs/opencl-1.1.pdf>. [Online]. Available: <http://www.khronos.org/registry/cl/specs/opencl-1.1.pdf>
- [5] *AMD Accelerated Parallel Processing OpenCL*, 2010, programming Guide. Available at http://developer.amd.com/gpu/ATIStreamSDK/assets/AMD_Accelerated_Parallel_Processing_OpenCL_Programming_Guide.pdf. [Online]. Available: http://developer.amd.com/gpu/ATIStreamSDK/assets/AMD_Accelerated_Parallel_Processing_OpenCL_Programming_Guide.pdf

SIP Based Multimedia Mobility in IP Networks

Jozef JANITOR, František JAKAB

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

jozef.janitor@tuke.sk, frantisek.jakab@tuke.sk

Abstract – Mobile operators are starting to offer fast, low latency mobile Internet (e.g. IP) connections over 3G or 4G technologies, while WiFi is getting deployed and available at more and more places. All that enables a possibility to access IP networks from almost any place and drives multimedia based realtime communications. Users are getting more and more dependent on either a voice over IP or video over IP communication, while generating a huge demand on access to these services any time, from any place, even when in move. That fact opens new challenges in mobile communications - Mobility. Mobility enables a wireless mobile device to change an attachment point to the network at any time. This change can however interrupt active connections that are carrying realtime communications between the communicating nodes. This paper deals with the mobility of multimedia sessions that were using the SIP signalization protocol. The paper includes outputs of measurements of SIP based mobility that were generated with a model application that was used to simulate a vertical handover between different layer 2 technologies.

Keywords – 3G, 4G, Handover, IP networks, Mobility, SIP

I. INTRODUCTION

In the last couple of years, there has been a rapid growth of the number of mobile devices such as notebooks, PDAs or smartphones. Users of these devices want to use them for real-time voice or video communication over cheap and accessible IP networks. IP networks were originally not designed to support a mobility of terminal devices. The mobility support within the IP networks has to be added either by extending several network protocols, or by designing applications with embedded support for mobility. These two approaches can be called mobility management using *intelligent network*, or *intelligent application* [1].

Intelligent networks use network devices and network protocols to detect and manage roaming of the mobile devices. The most commonly used macro-mobility protocol in intelligent networks is the Mobile IP protocol [2]. The Mobile IP protocol uses a tunnel between a Home Agent and a Mobile Node (optionally a Foreign Agent), so that the Mobile Node can be reached by one static IP address all the time. The mobile node must support the Mobile IP protocol on the operating system level.

Intelligent applications must either detect roaming or handover events by themselves, or the operating system can inform the application that the network settings have been changed. Intelligent applications use application layer protocols to manage the roaming mechanisms by implementing the ISO/OSI's Session layer and thus replacing the session identifier from the original session identifier S=(Source IP, Destination IP, Transport Protocol, Source Port, Destination Port), to a new unique session id.

II. MOBILITY OF MULTIMEDIA SESSION USING THE SIP PROTOCOL

Some properties of the SIP signalization protocol can be used to achieve mobility of the VoIP calls. According to [9] the mobility with SIP is divided into four options: *Personal mobility*, *Session mobility*, *Terminal mobility* and *Service mobility*.

1) *Session mobility/Mid-call mobility*: Mid-call mobility allows the terminal (node) to change its IP address even during an active call without losing the active communication session. After the detection of an IP address change, the SIP multimedia application (e.g. VoIP) has to inform the correspondent node about its new IP address. The application has to send a `re-INVITE` request with the existing session id. The session id with SIP corresponds to the value of the `Call-ID`: header which was used in the initial call establishment. In the body of this request,

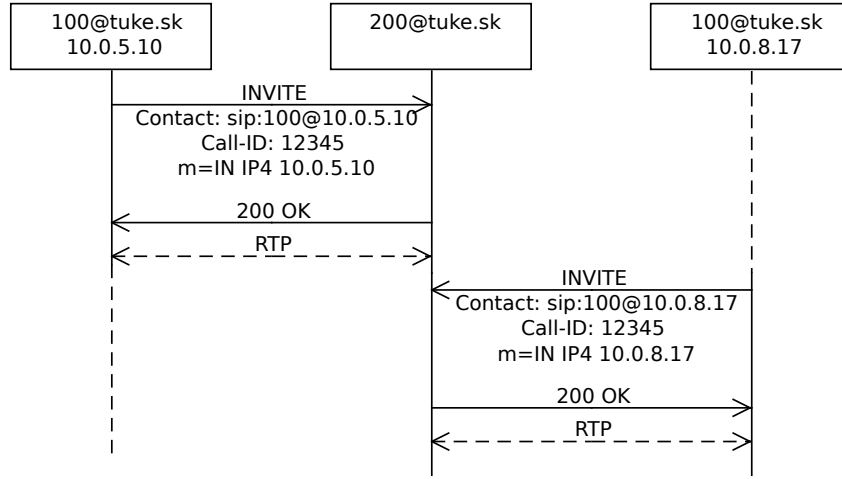


Fig. 1 Mid-call mobility. The mobile node has IP address 10.0.5.10 before roaming and 10.0.8.17 after roaming.

a new SDP message contains the new IP address of the terminal. The message sequence diagram of the roaming client is on the Fig. 1.

The time of the gap in VoIP call during a roaming can be expressed by the equation:

$$T = T_1 + T_2 + T_3 + T_4 \quad (1)$$

where T_1 is the reconnection time, T_2 is the time until the IP address is configured (usually by DHCP), T_3 is the time in which the SIP multimedia application detects the IP address change and T_4 is the time between the re-INVITE is sent and RTP packets are received again.

III. EXPERIMENTAL MODEL

To demonstrate the mobility of the SIP based VoIP calls, an experimental application was built. It is based on the proposed model of mid-call mobility. The application has the ability to detect a change of its own IP address. Part of the application which monitors the current IP address of the terminal is called "IP address change detector". It has an adjustable timer to control the periodicity of the IP address change detection. The value of this timer has influence on the time T_3 from the equation (1), but also on the host's CPU load.

Time T_1 and time T_2 from equation (1) can be omitted if the Mobile Node in the time of disconnection from the old network is already connected to the new network and already has its new IP address configured. Time T_4 depends on more factors, for example, a VoIP user agent on the Correspondent Node, or the complexity of the network.

In the experiment the Mobile Node was connected to an old network using an Ethernet cable and at the same time it was connected to a new network using a WiFi connection. The Ethernet connection had higher priority than the WiFi connection, thus VoIP application used the IP address from the old network prior to the IP address from the new network. A call was established between the Mobile Node and the Correspondent Node. During the call the Ethernet cable was unplugged from the Mobile Node. After a small amount of time the application detected the change of the IP address and sent a SIP INVITE request with a new SDP message to the Correspondent Node. The Correspondent Node received this message and then started to send RTP packets to the new IP address of the Mobile Node.

Packets exchanged during the experiment were captured and from sequence numbers in RTP packets the number of lost packets was determined. From number of lost RTP packet it is easy to calculate the gap in the voice call because each RTP packet contains a voice sample of 20ms:

$$T_x = n_x \times 20\text{ms} \quad (2)$$

where T_x is time of the gap and n_x is the number of lost RTP packets. In the experiment also a time between the last RTP packet in the old network and first RTP packet in the new network was measured. This time was denoted ΔT . The equation for ΔT is:

$$\Delta T = (n_x \times 20\text{ms}) + 20\text{ms} + T_j \quad (3)$$

where T_j is variable time caused by network jitter.

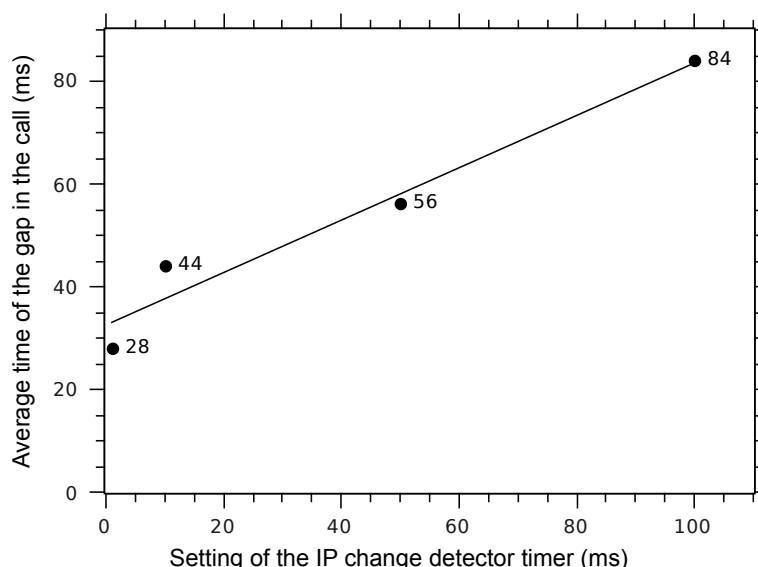


Fig. 2 Gap in a VoIP call during a handover for different values of the IP address change detector timer.

IV. CONCLUSION

This paper presents a concept of a mobility for VoIP calls using the SIP signaling protocol. Using this concept it is possible to build a VoIP application that supports pre-call and mid-call mobility of VoIP calls. This model does not rely on any extra functionality of the networking protocols or network devices. Also no special functionality has to be supported on the Correspondent Node.

The number of lost voice RTP packets depends on more factors, such as the complexity of the network, type of the VoIP user agent on the Correspondent Node, current load of the network, or the type of a SIP Proxy server.

To decrease the CPU load and simultaneously increase the IP address change detection speed, it is possible to use an approach when the operating system informs the application of the roaming event. In this case the application does not have to periodically monitor the current active IP address of the mobile device. On the other hand, this approach makes application more dependent on the underlying operating system, as closer cooperation is needed between the operating system and the application.

ACKNOWLEDGMENT

This work was supported by the Slovak Cultural and Educational Grant Agency of Ministry of Education of Slovak Republic (KEGA) under the contract No. 3/7245/09(60%).

This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF(40%).

REFERENCES

- [1] J. Janitor, P. Fecilak: *Mobility management in VoIP networks: Intelligent networks vs. Intelligent applications*. SCYR 2009, 9th Scientific Conference of Young Researchers, Kosice, May 2009, Kosice, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, 2009, 1. Edition, pp. 154-156.
- [2] C. Perkins, *IP Mobility Support for IPv4*. 2002. RFC 3344
- [3] A. Campbell, J. Castellanos, *IP Micro-mobility Protocols*. SIGMOBILE Mob. Comput. Commun. Rev. 4, 4 (Oct. 2000), 45-53.
- [4] E. Fogelstroem, A. Jonsson, C. Perkins, *Mobile IPv4 Regional Registration*. 2007. RFC 4857
- [5] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, *SIP: Session Initiation Protocol*. 2002. RFC 3261
- [6] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*. 2003. RFC 3550
- [7] R. Sparks, *The Session Initiation Protocol (SIP) Refer Method*. 2003. RFC 3515
- [8] A. Johnston, S. Donovan, R. Sparks, C. Cunningham, K. Summers, *Session Initiation Protocol (SIP) Basic Call Flow Examples*. 2003. RFC 3665.
- [9] H. Schulzrinne, E. Wedlund, *Application Layer Mobility Using SIP* SIGMOBILE Mob. Comput. Commun. Rev. 4, 3 (Jul. 2000), 47-57

Approaches to a Problem of Anomaly Detection in User Behaviour in a Smart Home Environment

Marek NOVÁK, František JAKAB

Dept. of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

marek.novak@tuke.sk, frantisek.jakab@tuke.sk

Abstract – In the paper we discuss a problem of detecting anomalies in user behaviour in a smart home environment. The focus is taken mainly on elderly people, since there is an assumption of stereotypes in their day routine. Process of abnormal behaviour detection is usually divided into two main steps - activity recognition and making decision whether an event belongs to normal state or does not. We introduce several approaches used for activity classification and anomaly detection and identify the drawbacks of current techniques.

Keywords – anomaly detection, action classification, machine learning, user behaviour

I. INTRODUCTION

The number of elderly people in European Union is quickly growing [1]. A nursing home is many times the only option for people who are no longer able to take care about themselves. Many of these seniors prefer to maintain their independence for as long as possible. One of the most notable problem is, when a senior becomes immobilized because of fall and is not able to get up or summon assistance [2]. He may be not able to call the assistance because a mobile phone is located too far or he has faint. People may be left in such situation for a long period of time until a neighbour or a caregiver becomes aware of it and helps.

One way how caregivers try to prevent the similar situation are numerous calls to beneficiaries during a day to control their health state. This might be bothersome for both caregivers and beneficiaries. Another way of dealing with this problem is usage of activation buttons, which the user should push in an emergency situation, for example a CareIP [3]. However, the user has to carry a device, press the button or there is needed some other intervention with the carer.

In a next section the problem of anomaly detection is described and there is introduced a question of the expected system's reaction to abnormal user behaviour. The third section deals with activities classification techniques and the fourth introduces several approaches to anomaly detection of the user behaviour.

II. ANOMALIES IN USER BEHAVIOUR

A. Anomaly detection problem

Anomaly detection refers to the problem of detecting patterns in data which do not conform to normal or expected behaviour. These techniques are extensively used in various applications such as: health care, intrusion detection in security systems, fault detection in safety critical systems, etc.

A standard process used in anomaly detection comprises of two main steps. The first is defining a region that represents normal behaviour and the second is making decision whether an observation belongs to the normal region or does not. Factors which make the decision process a complex task are [4]:

- boundary between normal and anomalous behaviour is often not precise;
- in many domains normal behaviour keeps evolving, therefore the region dedicated as normal could change in the future;
- small deviation from normal behaviour could be an anomaly in some domains, whereas in other domains they might be normal;
- labelled data for training model for anomaly recognition are often not available;
- data may often contains noise which is difficult to distinguish and remove;

B. How smart home should react to a particular behaviour?

When speaking about a smart home environment the big stress is given on the "smartness" of such a system. However, the question is what is meant by the ubiquitous intelligence, how a smart home should react to a particular behaviour of their inhabitants, etc. An C. Tran et al. have addressed several questions regarding the abnormal behaviour detection in their work [5]:

- What type of abnormal behaviour may occur?
- How does the system reason about its inhabitant's behaviour?
- How should it react to abnormal behaviour?
- What information should be involved in abnormality?

To find answers to these questions it is needed to get along with accessible data, to understand user activities from these data and to provide adequate response. If the system will bother the carer every hour with some notification about what elderly is doing, the satisfaction with the system will obviously be very low. On the other hand, if the system will not provide an alert in every situation when elderly is in danger, it may have unwanted consequences.

C. User behaviour Use Cases

An C. Tran et al. have stimulated the discussion of the desired behaviour recognition needed by an intelligent smart home by providing some Use Cases for abnormal behaviour identification [5]. Currently the editable repository consists of 19 Use Cases available at <http://muse.massey.ac.nz/shmuc>.

Table 1
 USE CASE - A VERY LONG SHOWER

Goal:	To detect an unusually extended activity.
Initial state:	Mary was at home alone.
Description:	Mary woke up at 08:00. She began her morning shower at 08:10, as usual, but at 08:40 the motion sensor in the bathroom still indicated movement, and the shower tap was still on, so her shower had lasted for 30 minutes.
Norm:	Mary normally showers for 10 to 20 minutes.
Outcome:	An alert message was sent to Debbie, who called Carita. Carita discovered Mary confused and cold in the shower, having forgotten what she was doing.
System design implications:	An excessively long activity may put the smart home inhabitant at risk. This poses the following questions: When does a shower become longer than usual? 1 minute over the average? 5? 10? Or is the amount of overrun relative (1%? 5%? 10%)? Perhaps a more sophisticated statistic is appropriate.

Use Cases are built on a story of Mary who lives alone and has dementia. Debbie, daughter of Mary, and carer Carita are looking after Mary. The system is able to produce only a small set of outputs: do nothing and wait for further data, raise an alarm, record message or send an urgent alert to Carita. An example of Use Case is shown in Table 1. Among other presented Use Cases are:

- **Abnormality in duration** - abnormal activity takes longer or shorter as usual.
- **Abnormality in time of occurrence** - an inappropriate activity start time could imply illness or even dementia, but a forgetful person may only need a reminder.
- **Inappropriate actions** - the additional behaviour does not conform to the norm. Some actions have been taken, which are usually not.
- **Wrong places** - performing an action in the wrong place (e.g., jumping on the bed or lying on the kitchen floor) may be dangerous or signify that something has gone wrong.

III. ACTIVITY DETECTION AND CLASSIFICATION

Sensors installed in a household may gather a huge number of data. There has to be added a meaning to discrete values measured by the sensors. Data has to be categorized in the sense of finding some time sequences or activities performed by the user. This is the prerequisite for abnormal behaviour detection, when a system will incorporate some context-awareness.

A. Approaches using predefined set of activities

There are several works related to activity classification using a predefined set of activities [6]. The disadvantage of this approach is the necessity of mapping the real behaviour of the user to a

predefined set of actions. Therefore only one main action at a time is analysed, not the day-long behaviour of the user. The causality among the actions is missing in this approach.

Other works are focused only on some types of activities, as provided by O. Brdiczka, M. Langet, J. Maisonnasse and J. L. Crowley [7]. They detect the posture of the user: sitting, standing, lying down using a 3D video tracking system with speech activity detection. This work is closely oriented on some postures, which do not provide an overall view of the behaviour of the user in a household. Fabian et al. make use of wearable devices consisting of a gyroscope and tri-axial accelerometer [8]. Three of these devices are placed on the subject's hip, dominant ankle and wrist. They are able to recognize a set of 7 activities. Although these approaches typically perform well for a small set of activities, they generally do not catch all activities a person performs in a household.

B. Sequential patterns in human behaviour

Guralnik V. and Haigh K.Z. have proposed a machine learning approach to model human sequential patterns in human behaviour [9]. They define sequential pattern as a list of sensor firings ordered in time. Examples of sequential patterns with time intervals and observation frequencies are:

- pattern of motion sensors: bedroom (6:45am–7:45am) → bathroom (7:00am–8:00am) was observed in 75% times;
- pattern of motion sensors: kitchen (6:00pm–6:20pm) → living room (6:20pm–7:00pm) → bedroom (9:00pm–10:00pm) was observed in 60% times

The objective of the model was to found out if it is possible to learn patterns that can explain the lifestyle of people living in household and if no important patterns are missed.

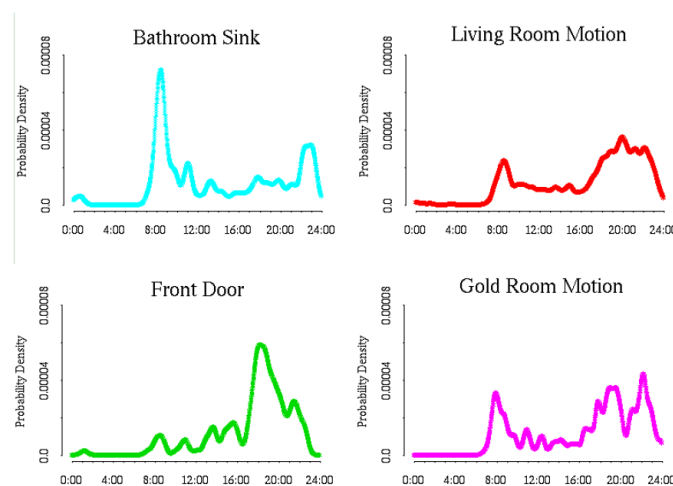


Fig. 1 Individual probability density functions for four sensors

Authors in their approach uses predetermined time intervals before algorithm of learning sequential patterns is started. The time intervals are based on the distribution of sensor firings during the day. Probability density function is used to make histogram of sensor firings smoother. In Fig. 1 are shown probability density functions for four sensors in house (bathroom sink, living room, front door, gold room). From histograms one can see, that bathroom sink is highly occupied in morning hours. Let t_1 and t_2 be points on time axis, then area under the curve in interval t_1 to t_2 is the probability that a sensor will fire in this interval. Time intervals are defined as successive local minimas of the probability density function. They are calculated for each sensor separately.

Some additional clustering techniques are needed to finalize pattern observation. Sequential pattern algorithm generates large number of patterns (above 5300). Nevertheless the usage of optimizing filters the number of patterns (182) was still high. Reasonable number to be analysed by the caregiver is 50.

IV. TECHNIQUES FOR SPECIFIC ANOMALY DETECTION

A. Approaches oriented on one anomaly

Fall is considered as one of the most probable unwanted event that may occur to elderly. There are many works related to fall detection [10], [11]. Fall detection in these works is realized by

using different types of accelerometers, mobile devices or sensors attached on the user's clothes. The drawback of this approach is the necessity of wearing some additional device on a body, which is not very accepted by the users. In addition, wearable devices rely on liability of the user, since he must not forget to attach the device in the morning. However, this task is easy to forget by the user, especially when speaking about elderly and disabled people.

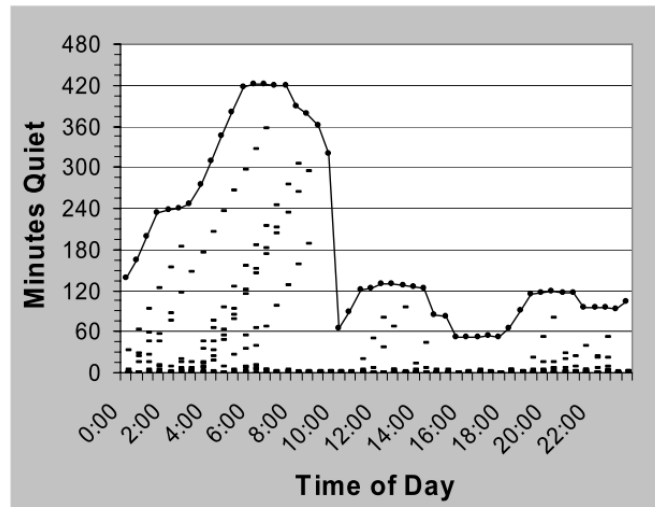


Fig. 2 Inactivity data with alert line

Another anomaly, on which research is focused is detection of inactivity. Cuddihy P., Weisenberg J., Graichen C. and Ganesh M. [12] in their work present an algorithm to automatically detect abnormally long periods of inactivity at home. The result of the work is Automatic Inactivity Detection (AID) system that automatically adjusts to the activity patterns of the resident. The AID algorithm is based on inactivity duration data from the home over the past weeks and months. The elapsed inactivity data are used to create a threshold for acceptable elapsed inactive time for each time of day, in 30-minute increments. If the resident is inactive longer than the determined time interval, then a notification should be sent to a carer. Threshold is depicted in Fig. 2, where x-axis represents time during the day and y-axis represents time interval of detected inactivity. In the Fig. 2 the threshold is created by joining observations of the longest inactivity times. To make algorithm more robust the alert line is changed by adding a proportional parameter to the size of the inactive data.

B. Comparison of different approaches

In Table 2 are introduced three different approaches to anomaly detection. In many works Machine learning techniques such as Hidden Hierarchical Markov Chain, Neural Networks or Support Vector Machines are used to detect abnormal user behaviour. The limitation of all current approaches is orientation on one or few specific scenarios and the the information about the context of the user is missing.

V. CONCLUSION

Current research is oriented on activity classification and techniques that determine whether observed behaviour falls into predefined sets of activities. Other research is focused on finding anomalies in predefined scenarios [7]. However, both of them are limited to a small set of specific activities, which do not correspond to real behaviour of the user at home.

Techniques like automatic inactivity detection [12] algorithm lack some additional information about the context of the user. For example when the user goes to sleep late, it is assumed that he will sleep longer, though this information is not incorporated in this technique.

Our aim is to create a dynamical model that will adapt to changing conditions in the user behaviour by the knowledge of some dependencies between the activities. For this purpose a Bayesian networks seems to be a good solution. In comparison with aforementioned models, we would like to focus not only on one specific abnormal activity, but to create a model from which more abnormalities may be acquired.

Table 2
 COMPARISON OF SELECTED ANOMALY DETECTION APPROACHES

Technique	Authors	Mathematical model	Approach description	Disadvantage
Automatic inactivity detection algorithm	Cuddihy, Weisenberg, Graichen, Ganesh [12]	Inactivity threshold line	System observes the inactivity of the user in 10 days and creates parametrized inactivity threshold. If the inactivity exceeds the threshold an alarm is raised.	detecting only inactivity, information like time or context of the user are not incorporated
Temporal data mining	Jakulla, Cook [13]	Machine learning	Temporal logic of actions (before, after, meets, overlaps, starts ...) is used to identify temporal relations between event. Anomaly is detected by calculating the probability of a given event occurring or not occurring).	long learning phase (59 days), events grouped together are only from the same device (on/off)
Wearable sensor based abnormal activity detection	Yin, Yang, Junfeng Pang [14]	Support Vector Machines (SVM), Kernel Nonlinear Regression (KNLR)	In the first phase, all of the normal traces are directed via a module to compute significant features, which are then passed through a one-class SVM to eliminate the traces that have a high probability of being normal. The rest, which are suspicious traces, are then passed on to a secondary phase for further detection.	generates a large number of abnormal models, user carries out only single type of activities (interleaving activities are not considered)

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6) under grant agreement MonAMI IST-5-0535147 - Mainstreaming on Ambient Intelligence (60%). This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (40%).

REFERENCES

- [1] A. Zaidi, "Features and Challenges of Population Ageing: The European Perspective," March 2008. [Online]. Available: http://www.euro.centre.org/data/1204800003_27721.pdf
- [2] M. E. Tinetti, W.-L. Liu, and E. B. Claus, "Predictors and prognosis of inability to get up after falls among elderly persons," *JAMA The Journal of the American Medical Association*, vol. 269, pp. 65–70, Jan. 1993.
- [3] CareTech, "Digital Technology For Peace of Mind." [Online]. Available: <http://www.careip.com.au/news/1-news/16-digital-technology-for-peace-of-mind.html>
- [4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15:1–15:58, July 2009. [Online]. Available: <http://doi.acm.org/10.1145/1541880.1541882>
- [5] A. C. Tran, S. Marsland, J. Dietrich, H. W. Guesgen, and P. Lyons, "Use cases for abnormal behaviour detection in smart homes," in *Proceedings of the Aging friendly technology for health and independence, and 8th international conference on Smart homes and health telematics*, ser. ICOST'10. Springer-Verlag, 2010, pp. 144–151. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1894439.1894460>
- [6] W. Kang, D. Shin, and D. Shin, "Detecting and predicting of abnormal behavior using hierarchical Markov model in smart home network," in *Industrial Engineering and Engineering Management (IE EM), 2010 IEEE 17th International Conference on*, Oct. 2010, pp. 410–414. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5646583&isnumber=5645891>
- [7] O. Brdiczka, M. Langet, J. Maisonnasse, and J. Crowley, "Detecting Human Behavior Models From Multimodal Observation in a Smart Home," *Automation Science and Engineering, IEEE Transactions on*, vol. 6, no. 4, pp. 588–597, 2009.
- [8] O. A. Baki, joy Zhang, martin Griss, and T. Lin, "A Mobile Application to Detect Abnormal Patterns of Activity," in *Mobile Computing, Applications, and Services, First International ICST Conference, MobiCASE 2009*, vol. 35, Oct. 2010.
- [9] V. Guralnik and K. Z. Haigh, "Learning models of human behaviour with sequential patterns," in *Proceedings of the AAAI-02 workshop "Automation as Caregiver"*, 2002, pp. 24–30, aAAI Technical Report WS-02-02.
- [10] F. Allen, E. Ambikairajah, N.H.Lovell, and B. Celler, "An Adapted Gaussian Mixture Model Approach to Accelerometry-Based Movement Classification Using Time-Domain Features," in *Engineering in Medicine and Biology Society, 2006. EMBS, 2006*, pp. 3600–3603.
- [11] M. Yu, S. Naqvi, and J. Chambers, "A robust fall detection system for the elderly in a Smart Room," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 1666–1669.
- [12] P. Cuddihy, J. Weisenberg, C. Graichen, and M. Ganesh, "Algorithm to automatically detect abnormally long periods of inactivity in a home," in *Proceedings of the 1st ACM SIGMOBILE international workshop on Systems and networking support for healthcare and assisted living environments*, ser. HealthNet '07, 2007, pp. 89–94. [Online]. Available: <http://doi.acm.org/10.1145/1248054.1248081>
- [13] V. Jakkula and D. J. Cook, "Anomaly Detection Using Temporal Data Mining in a Smart Home Environment," *Methods of information in medicine*, vol. 47, no. 1, pp. 70–75, July 2008. [Online]. Available: <http://view.ncbi.nlm.nih.gov/pubmed/18213431>
- [14] Y. Jie, Y. Qiang, and J. Pan, "Sensor-based abnormal human-activity detection," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 8, pp. 1082–1090, 2008.

Modelovanie povrchových štruktúr

Norbert ÁDÁM

Katedra počítačov a informatiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

norbert.adam@tuke.sk

Abstrakt — Geografické informačné systémy v posledných rokoch sa stali súčasťou tvorby informačných systémov v dôsledku nezastupiteľnej úlohy lokalizácie objektov a javov spracovávaných týmito systémami. Vzťahovať dáta a informácie o aktivitách a zdrojoch k polohe v priestore a monitorovať, prípadne predpovedať vývoj v čase je pre modernú spoločnosť nevyhnutné. Kvalitným zdrojom údajov je digitálna mapa. Tento článok pojednáva o možnostiach tvorby vrstevnicových a spádnicových máp. V článku je uvedený algoritmus kreslenia vrstevníc a spádnic za predpokladu existencie analytického vyjadrenia topografickej plochy. V prípade, že nie je známe analytické vyjadrenie topografickej plochy, tak je možné vychádzať z nameraných výškových bodov. Uvedený systém výškových bodov zabezpečuje správne interpretovanie vrstevnicovej a spádnicovej štruktúry, a poskytuje jednoduché rozhranie pre konštrukciu vrstevníc, spádnic a reliéfu terénu.

Kľúčové slová — geografický informačný systém, vrstevnice, spádnice, výškový model

I. ÚVOD

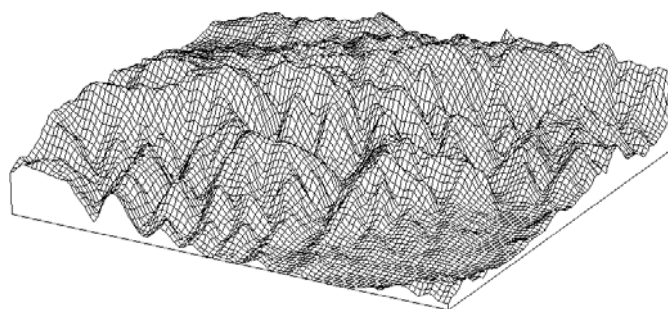
Zber údajov o priestorovom rozmiestnení dôležitých objektov a zdrojov na zemskom povrchu bol vždy výraznou aktivitou ľudských spoločností. Motívy boli rôzne, ale vždy existovala úzka súvislosť s astronómiou, využívali sa poznatky o vzťahoch Zeme a kozmických telies. Ľudské aktivity postupne prinášali stále nové poznatky o priestorových objektoch na Zemi. Výsledkom týchto aktivít boli mapy, napr. astronomické, geologické, tektonické, pôdne, demografické, dopravné, klimatické a pod. Na rozdiel od pôvodnej potreby ako presne určiť polohu objektu, bolo tu potrebné riešiť uloženie rozrastajúceho sa množstva informácií o nich. Riešením sa stali počítačové systémy.

Geografický informačný systém (GIS) ako počítačový nástroj pre mapovanie a analýzu vecí a javov reálneho sveta, spája bežné databázové operácie, ako je zadávanie úloh a štatistické výpočty s jedinečnými možnosťami zobrazenia a priestorovej analýzy, ktoré poskytuje kartografická mapa [1].

Kartografia, zaoberajúca sa zobrazovaním a štúdiom priestorového rozmiestnenia a vzájomných väzieb javov prírody a spoločnosti, používa grafické a modelové prostriedky na tvorbu máp. Časť východiskových informácií pre kartografiu poskytuje geodézia. Výsledky mnohých geodetických meraní tvoria podklady pre spracovanie máp rôznych druhov, ktoré majú rôzne účely [2],[3].

II. MAPY

Pri tvorbe mapy je potrebné zmenšiť obraz skutočnosti [2]. Pri každom zmenšení nastáva zovšeobecnenie (generalizovanie) reality – vylúčenie nepodstatných skutočností a podrobností zo zobrazenia reality na mape (obr. 1). Proces generalizácie nesmie ovplyvniť prehľadnosť, názornosť a estetickosť mapy. Obsah mapy sa vytvára konvenčnými kartografickými výrazovými prostriedkami (kartografickými znakmi), ku ktorým patria čiary (obrysové čiary objektov, izočiar – vrstevnice atď.), tvarové značky (geometrické, názorné, symbolické, písmenové a pod.), čiarové značky, numerické údaje v rôznych farbách, plošné značky, farebné vykrytie plôch, tieňovanie, rastovanie, bodové značky, pohybové značky a ďalšie. Pri rozličných automatizovaných spracovateľských procesoch sa čoraz viac uplatňujú digitálne mapy [6].



Obr. 1 Axonometrické sieťové zobrazenie terénu

Digitálny model terénu je súbor vybraných bodov – uzlových bodov topografickej plochy. Terénna plocha je veľmi mnohotvárná. Digitálny model terénu ju definuje v uzlových bodoch, pričom výšky ostatných bodov terénnej plochy treba určiť na základe vhodnej interpolácie. Terénne body z hľadiska aproximácie topografickej plochy môžu byť regulárne (možno v nich definovať dotykové roviny k topografickej ploche), a singulárne (nemožno v nich vytvárať dotykové roviny, body tvoria „terénne hrany“, napr. body hrán násypov, jám atď.).

Ak všeobecný tvar rovnice spojitaj topografickej plochy nad určitou ohraničenou oblasťou je daný zápisom:

$$z = f(x, y) \quad (1)$$

s neznámym analytickým vyjadrením a neznámymi koeficientmi, potom túto plochu je možné definovať len konečnou množinou n uzlových bodov vhodne rozmiestnených na ploche oblasti, ktorých N -rozmerný vektor súradníc:

$$\vec{z}_{n,l} = \vec{f}_{n,l}(x, y) \quad (2)$$

je známy z meraní. Výšky ďalších bodov (súradnice z) sa určujú interpolačným polynómom m -tého stupňa:

$$P_m(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \dots + a_{m0}x^m + a_{m-1,1}x^{m-1}y + a_{0m}y^2 \quad (3)$$

ktorý bude aproximovať neznáme vyjadrenie topografickej plochy (1). Minimálny počet uzlových bodov na určenie koeficientov (3) je:

$$n_{\min} = \frac{(m+1)(m+2)}{2} \quad (4)$$

pričom n sa zvolí podľa $n > n_{\min} > m$. Topografickú plochu možno aproximovať aj pomocou iných interpolačných mnohočlenov, napríklad pomocou B-spline polynómu alebo pomocou neuniformných racionálnych B-spline polynómov (NURBS – Non Uniform Rational B-spline) a pod. [2].

III. VRSTEVNICE

Základnou požiadavkou na model terénu je možnosť odvodiť výšku terénu pre definované miesto, objekt a pod. Spôsob riešenia tejto požiadavky súvisí hlavne s typom modelu terénu. Popri spôsobe obrazového zobrazenia sa používa aj vrstevnicové zobrazenie povrchov vytvorených podľa Euler-Mongeho [4]. Prednosti sú zjavné hlavne v oblasti technických aplikácií. Takým spôsobom sa vytvárajú napríklad zemepisné a meteorologické mapy, resp. obrázky zobrazujúce rozdelenie potenciálu elektrického poľa.

Vrstevnicové zobrazenie (obr. 2) sa dá rozšíriť aj na funkcie, ktoré nie sú definované na rovinnej oblasti ale na zakrivených plochách, ako je napríklad rozdelenie teploty na povrchu súčastky zložitého tvaru. O hodnotách funkcie svedčia hodnoty pri vrstevniciach, o strmosti plochy zase ich hustota.

Nech funkcia (1) určujúca tvar topografickej plochy je ohraničená na intervale $\Omega = [\alpha, \beta] \times [\gamma, \delta]$. Rovnica vrstevnice s menovitou hodnotou λ bude mať tvar:

$$M(\lambda) = \{(x, y) \in \Omega \mid f(x, y) = \lambda\} \quad (5)$$

Teda uvedená vrstevnica je súhrn tých bodov, v ktorých funkcia nadobúda hodnotu λ .

V prípade že nie sú dané menovité hodnoty jednotlivých vrstevníc, najprv sa vypočítajú extrémny (minimum a maximum) funkcie na intervale Ω :

$$\begin{aligned} f_{\max} &= \max_{\Omega} \{f(x, y)\} \\ f_{\min} &= \min_{\Omega} \{f(x, y)\} \end{aligned} \quad (6)$$

Prakticky sa to dá urobiť pomocou nelineárneho optimalizačného algoritmu, ktorý udáva extrémny danej funkcie. Vypočítané body sa použijú aj pri zobrazení vrstevnicovej štruktúry. Keďže rozlíšenie monitoru je relatívne malé, tento nelineárny optimalizačný algoritmus má veľmi jednoduchý tvar, t.j. vypočítajú sa hodnoty funkcie v každom bode oblasti Ω , ktoré zodpovedajú bodom na obrazovke. Z týchto bodov sa určí minimum a maximum.

Pre nakreslenie c vrstevníc, takých že ich vzájomná vzdialenosť je konštantná, menovité hodnoty sa dajú vypočítať podľa:

$$\lambda_i = f_{\min} + i \cdot \frac{f_{\max} - f_{\min}}{c + 1}; \quad i \in \{0, 1, \dots, c + 1\} \quad (7)$$

Na kreslenie vrstevníc sa použije nasledujúci algoritmus:

VSTUP: zoznam súradníc uzlových bodov terénnej plochy $(x_{0,0}, y_{0,0}, z_{0,0}), (x_{0,1}, y_{0,1}, z_{0,1}), \dots, (x_{n,1}, y_{n,1}, z_{n,1})$

1. Výšky ďalších bodov (súradnice z) sa určujú interpolačným polynómom m -tého stupňa podľa (3).
2. Vypočítajú sa extrémny podľa (6).
3. Nech oblasť Ω je reprezentovaná pomocou obdĺžnika (X_0, Y_0, X_p, Y_q) na obrazovke. Rozdelí sa obdĺžnik tak, aby v každom riadku bolo p a v každom stĺpci q obrazových bodov (pixelov):

$$p = X_p - X_0, \quad q = Y_q - Y_0 \quad (8)$$

4. Vzťah medzi bodmi (x_r, y_s) oblasti Ω a bodmi (X_r, Y_s) na obrazovke sa dá vyjadriť pomocou nasledujúcich rovníc:

$$x_r = \alpha + \frac{r}{p}(\beta - \alpha) \rightarrow X_r = X_0 + r; \quad r \in \{0, 1, \dots, p\} \quad (9)$$

$$y_s = \gamma + \frac{s}{q}(\delta - \gamma) \rightarrow Y_s = Y_0 + s; \quad s \in \{0, 1, \dots, q\}$$

5. V každom bode, kde funkcia

$$g_i(r, s) = f(x_r, y_s) - \lambda_i; \quad i \in \{0, 1, \dots, c\} \quad (10)$$

zmení znamienko, je potrebné zmeniť farbu bodu rastra.

IV. SPÁDNICE

Dôležitou vlastnosťou vrstevníc a spádnic je ich vzájomná kolmost' (obr. 2). Dve spádnice, resp. dve vrstevnice sa nikdy nemôžu pretínať, čo vyplýva aj z ich definície.

Gradientom funkcie (1) je taký vektor, ktorého súradnice sú parciálne derivácie funkcie f , čiže:

$$\nabla f(x, y) = \left(\frac{\partial f}{\partial x}(x, y), \frac{\partial f}{\partial y}(x, y) \right) \quad (11)$$

Vektor gradientu ukazuje vždy v smere najväčšieho rastu. Nech E_x a E_y reprezentujú súradnice vektoru gradientu:

$$E_x = \frac{\partial f}{\partial x}, \quad E_y = \frac{\partial f}{\partial y} \quad (12)$$

Jednotková dĺžka oblúku spádnice sa aproximujeme s úsečkou, ktorá je kolineárna s dotyčnicou a ktorej priemety na osi súradnicovej sústavy Oxy sú dx a dy , t.j.:

$$\frac{dx}{dy} = \frac{E_x}{E_y} \quad (13)$$

Dĺžka spomenutej úsečky je daná vzťahom:

$$ds = \sqrt{dx^2 + dy^2} \quad (14)$$

Tým pádom rovnice jednotlivých priemetov nadobúdajú tvar:

$$dx = \frac{E_x}{E} ds, \quad dy = \frac{E_y}{E} ds \quad (15)$$

kde E je dĺžka gradientu:

$$E = \sqrt{E_x^2 + E_y^2} \quad (16)$$

Na kreslenie spádnic sa použije nasledujúci algoritmus:

1. Zadá sa štartovací bod (x^k, y^k) .
2. Zadá sa jednotková dĺžka oblúka (ds).
3. V tomto bode sa vypočítajú súradnice vektoru gradientu.
4. Použitím vzťahov (15) vypočíta sa dx a dy . Pre súradnice nového bodu spádnice platí:

$$(x^{k+1}, y^{k+1}) = (x^k + dx, y^k + dy) \quad (17)$$

5. Pomocou na to určenej procedúry sa nakreslí úsečka, ktorá je aproximáciou jednotkového oblúka spádnice.
6. Opakuje sa uvedený postup 1 až 5 aj pre novo vypočítané body.

Čím menšia je dĺžka jednotkového oblúka ds , tým väčšiu presnosť je možné dosiahnuť. Koncový bod spádnice je totožný s bodom, určujúcim globálne minimum resp. maximum funkcie $z = f(x, y)$. V týchto bodoch je potrebné definovať podmienku ukončenia algoritmu, lebo v extrémoch parciálne derivácie sa rovnajú nule. Podmienka ukončenia algoritmu má tvar $E < \varepsilon$, kde ε je dostatočne malé kladné číslo. Okrem toho spádnice môžu mať koniec aj na krajoch danej oblasti, preto aj pre tento prípad je potrebné zadať podmienky ukončenia algoritmu.

V. SYSTÉM VÝŠKOVÝCH BODOV

V prípade, že nie je známe analytické vyjadrenie topografickej plochy, tak je možné vychádzať z nameraných, prípadne definovaných výškových bodov. Základom spomenutého modelovania terénu je nasledujúci fyzikálny model:

1. Body plochy ležiace v uzloch siete, ktoré rozdeľujú definičný obor, sa považujú za také hmotné body, ktoré sa môžu pohybovať len vo vertikálnom smere.
2. Body plochy sú spojené s rovinou $z = 0$ v dôsledku pružnej sily:

$$F_1 = -azk \quad (18)$$

kde z je vzdialenosť bodu od roviny xy , a a je koeficient pružnosti.

3. Výškové body (x_k, y_k, z_k) pôsobia príťažlivou silou na body plochy so súradnicami (x, y, z) , ktorej veľkosť je nepriamo úmerná druhej mocnine vzdialenosti medzi nimi a pre túto silu platí:

$$\vec{F}_2 = \sum_{k=1}^n \frac{w_k [(x - x_k)\vec{i} + (y - y_k)\vec{j} + (z - z_k)\vec{k}]}{[(x - x_k)^2 + (y - y_k)^2 + (z - z_k)^2]^{\frac{3}{2}}} \quad (19)$$

kde w_k je váha (hmotnosť) k -tého výškového bodu.

4. Polohu bodu plochy určí nasledujúca silová rovnováha:

$$F_1\vec{k} + F_2\vec{k} = 0 \quad (20)$$

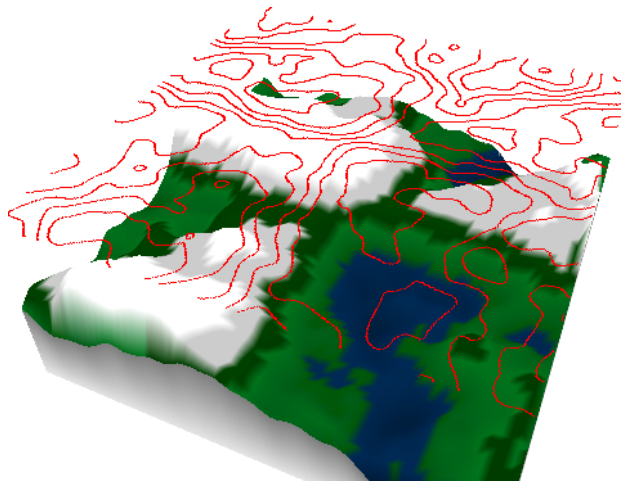
čo pri $a = 1$ vedie k nasledujúcej rovnici:

$$z = \sum_{k=1}^n \frac{w_k (z_k - z)}{[(x - x_k)^2 + (y - y_k)^2 + (z - z_k)^2]^{\frac{3}{2}}} \quad (21)$$

Hore uvedená rovnica je rekurzia, a je možné ju vypočítať iteračným spôsobom (vychádzajúc zo štartovacieho bodu $z = 0$ rovnica (21) sa aplikuje dovtedy, kým rozdiel dvoch

za sebou nasledujúcich hodnôt nebude menšie ako číslo ε , kde číslo ε je dostatočne malé kladné číslo reprezentujúce chybu výpočtu).

Na obrázku 2 je znázornený terén vytvorený na základe uvedeného algoritmu.



Obr. 2 Model terénu so zobrazením vrstevníc s odstupom $c=6$

VI. ZÁVER

Dnešné nové technológie umožňujú získavanie dát s vyššou presnosťou, rýchlosťou a pravidelnosťou, než minulosti. Softvérové prostriedky ako jadro systémových technológií, umožňujú spracovávať surové dáta z rôznych systémov uskutočňujúcich ich zber [5]. V mnohých typoch geografických operácií sa výsledok najlepšie vizualizuje ako mapa alebo graf. Mapy sú veľmi efektívnym prostriedkom uloženia a odovzdávania priestorových informácií.

Článok sa zaoberá tvorbou vrstevnicových a spádnicových máp. V článku je uvedený algoritmus kreslenia vrstevníc a spádnic za predpokladu existencie analytického vyjadrenia topografickej plochy. V prípade, že nie je známe analytické vyjadrenie topografickej plochy, tak je možné vychádzať z definovaných výškových bodov. Systém výškových bodov zabezpečuje fyzikálne správne interpretovanie vrstevnicovej a spádnicovej štruktúry. Poskytuje jednoduché rozhranie pre konštrukciu vrstevníc, spádnic a reliéfu terénu. Nevýhodou použitia tohto systému je to, že rastom počtu výškových bodov rastie čas výpočtu parametrov spomenutých štruktúr.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Tuček, J., *GIS – Geografické informační systémy (Principy a praxe)*, Computer Press, Praha, 1998. ISBN 80-7226-091-X.
- [2] Šutti, J., *Geodézia*, Vydal Alfa, vydavateľstvo technickej a ekonomickej literatúry: Bratislava, ako spoločné vydanie s SNTL – Nakladatelství technické literatúry, Praha, 063 - 560 – 87 GEO.
- [3] Višňovský, P., Čihal, A., *Geodézia a fotogrametria*, Príroda, Bratislava, 1985. 5846 - SÚKK 1823/I-84.
- [4] Füzi, J., *3D grafika a animácia na PC* (in Hungarian). ComputerBooks, Budapest, 1997. 963-618-111-X.
- [5] Vokorokos, L.: *Princípy architektúr počítačov riadených tokom údajov*. Monografia. Copycenter, spol. s.r.o., Košice, 2002. ISBN 80-7099-824-5.
- [6] Czímber K., *Geoinformatika – elektronické poznámky* (in Hungarian), 2001, <http://geo.efe.hu/hun/onlinejegyzet/geoinfo/geoinfo2.htm>.

Zvýšenie návštevnosti webu na základe Long Tail analýzy

Adela TUŠANOVÁ, Ján PARALIČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

adela.tusanova@tuke.sk, jan.paralic@tuke.sk

Abstrakt — Výhody aj dostupnosť internetu a internetového podnikania sa každým rokom neustále zvyšujú. Vznikajú tisíce nových webov s rovnakým zameraním, konkurencia rastie, s ňou stúpa boj o návštevníka. Majitelia webových aplikácií a vývojári si pomaly začínajú uvedomovať dôležitosť kľúčových slov – každý chce byť na prvom mieste vo vyhľadávaní či platiť čo najmenej za kliky pri on-line reklame typu PPC (Pay Per Click). Cieľom je získať čo najviac návštevníkov, ktorí budú konvertovať k určitému cieľu – zakúpiť si produkt, registrujú sa a podobne. V tomto článku chceme poukázať na jednu z metód ako zvýšiť návštevnosť webu a to pomocou tzv. „Long Tail“ analýzy.

Kľúčové slová — Long Tail, kľúčové slová, analýza návštevnosti, segmentácia, optimalizácia, zvyšovanie návštevnosti, Google Analytics

I. ÚVOD

Zvyšovanie návštevnosti webu znamená zvyšovať zdroje návštevnosti ako sú priama návštevnosť, návštevnosť z externých liniek, prirodzená návštevnosť z vyhľadávačov a návštevnosť z reklamných odkazov [3].

Prirodzená návštevnosť z vyhľadávačov - prirodzené výsledky vyhľadávania tvoria najväčší podiel na návštevnosti webových stránok. Vyhľadávače hodnotia stránky indexom, tzv. rankom (Page rank, S-rank,...), do hodnotenia vstupuje až 200 faktorov. Cieľom je dosiahnuť prvé pozície vo výsledkoch vyhľadávania na relevantné kľúčové slová.

Reklamné odkazy sú plateným zdrojom návštevnosti webových stránok. Tie sa zobrazujú najčastejšie vo výsledkoch vyhľadávania alebo na vybraných webových stránkach. Sú prispôbené vyhľadávaciemu dotazu alebo obsahu webu na základe kľúčových slov.

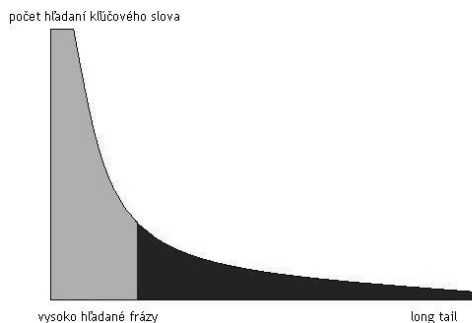
Kľúčové slovo, alebo aj keyword, je zoznam slov, fráza, alebo jedno slovo, ktoré užívateľ zadá do vyhľadávača a očakáva, že vyhľadávač mu na zadané slovo ponúkne relevantné stránky, ktoré dané slovo/frázu obsahujú. Správne použitie kľúčových slov na stránke je jedným z najdôležitejších procesov pri optimalizácii pre vyhľadávače, ako aj pri optimalizácii reklamných odkazov. Nových návštevníkov na danú webovú stránku teda prinášajú kľúčové slová, pre ktoré je webová stránka optimalizovaná [4].

Veľká dilema pre mnoho odborníkov je výber toho, či sa celá optimalizácia bude zameriavať na hlavné kľúčové slová alebo sa pôjde do hĺbky a obsiahne sa síce väčší počet slov, ale s menšou vyhľadávanosťou. Hlavnou výhodou optimalizovania stránky pre hlavné kľúčové slová je väčšinou množstvo ich vyhľadávania za mesiac. Na druhej strane ale takéto frázy sú zväčša všeobecnejšie a preto aj konverzný pomer (počet zrealizovaných udalostí k počtu celkových udalostí) v prípade e-commerce stránok môže byť oveľa nižší ako by tomu bolo v prípade návštev z konkrétnych slov. Taktiež konkurencia na hlavné kľúčové slová je zvyčajne príliš vysoká a preto je ťažké dostať sa na prvé miesta vo vyhľadávačoch. V prípade textových reklám sa za klik na hlavné kľúčové slová platí niekoľkonásobne viac ako na kľúčové slová, ktoré sú menej vyhľadávané. Práve pozície vo vyhľadávačoch a ceny za jednotlivé kliky pri inzerovaní sú motiváciou pre analýzu tzv. long tailových kľúčových slov.

II. LONG TAIL

Termín long tail všeobecne v marketingu popisuje pravú časť grafu predajnosti (viď obr. 1), obsahujúci položky s malým obrátom. Klasický predaj využíva tzv. Pareto princíp, ktorý tvrdí, že 80% obrátu tvorí asi 20% výrobkov a podľa toho obmedzuje svoje skladové zásoby iba na

masovo žiadané produkty. Internet však umožňuje uspokojiť aj klientov žiadajúcich menšinové produkty, čím sa oblasť long tail zväčšuje a rastie jeho ekonomický význam [5]. Long tail teda predstavuje menšinové produkty, ktoré v súhrne môžu priniesť firme rovnaký ak nie väčší zisk ako produkty tvoriace hlavnú časť odbytu.



Obr. 1 Long Tail

Graf znázorňuje počet hľadání na ose Y a kľúčové slová zoradené podľa hľadanosti na ose X. Predpokladáme, že long tail krivka je konvexná a je teda vidieť, že kľúčové slová na dlhý chvost môžu priniesť rovnakú alebo aj vyššiu návštevnosť (čierna plocha pod krivkou) ako tie vysoko hľadané (sivá plocha pod krivkou). Záleží na „dĺžke chvosta“ a na zakrivení long tail krivky. Analýzou tohto grafu potom dokáže majiteľ webu rozhodnúť, na optimalizáciu ktorých kľúčových slov sa má zamerať.

III. ANALÝZA LONG TAILU POMOCOU GOOGLE ANALYTICS

Analýzu long tail-u umožňujú pokročilé štatistické nástroje, z ktorých najpopulárnejší je Google Analytics. Aby sme mohli analyzovať long tailové kľúčové slová, je potrebné nastaviť pokročilú segmentáciu pomocou regulárnych výrazov. Takáto segmentácia nám umožňuje analyzovať historické dáta a aj spätne zistiť najvyhľadávanejšie frázy skladajúce sa z viacerých slov. Regulárne výrazy umožňujú vytvoriť zvlášť segmenty pre frázy skladajúce sa z dvoch, troch, či viacerých slov. Príkladom regulárneho výrazu na segmentáciu fráz skladajúcich sa z dvoch slov je napríklad výraz $^(\S+)\s(\S+)\$,$ kde majú jednotlivé znaky nasledujúci význam:

- ^ začiatok reťazca
- \S akýkoľvek znak okrem „bielych“ znakov
- + minimálne 1krát (maximálne neobmedzené)
- \s „biele“ znaky (medzera, tabulátor, znaky pre zalomenie riadkov)
- \$ koniec reťazca

Tento regulárny výraz teda kontroluje, či sa daný reťazec začína a končí akýmkoľvek znakom resp. znakmi okrem „bielych“ znakov, pričom tieto znaky sú oddelené medzerou, tabulátorom alebo znakom pre zalomenie riadkov.

Konfiguráciu segmentu uvádzame na obrázku 2.

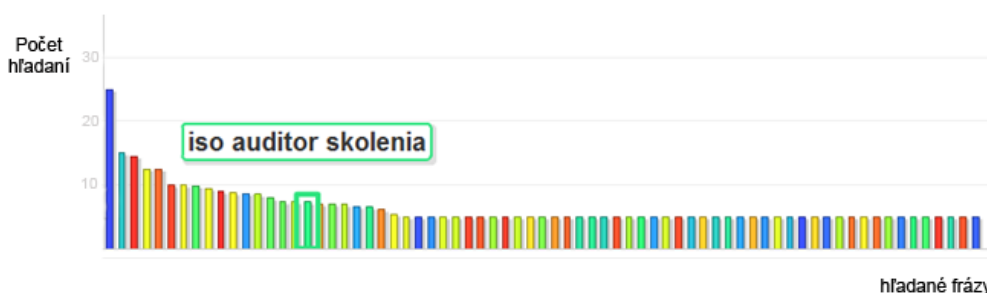
Obr. 2 Konfigurácia segmentácie kľúčových slov v Google Analytics

IV. ANALÝZA LONG TAIL KĹÚČOVÝCH SLOV WEBOVEJ STRÁNKY WWW.QSYSTEM.SK

Webová stránka spoločnosti Q-System s.r.o. bola založená v roku 2008. Jej hlavnou úlohou je poskytovať základné informácie o spoločnosti, ktorá poskytuje služby v oblasti kvality a metrologie. Stránka bola optimalizovaná na hlavné kľúčové slová, ako metrologia, kalibrácie, ISO, apqp, fmea, metrolog, audítor atď. Tieto kľúčové slová spravidla nájdeme medzi prvými 20-timi najvýkonnejšími kľúčovými slovami. V roku 2009 bola stránka propagovaná pomocou reklamných textov v systéme Google AdWords. Reklamné kampane boli opäť optimalizované na základe hlavných kľúčových slov. Návštevnosť sa v roku 2009 oproti roku 2008 vďaka reklame zvýšila o takmer 73%. Pri porovnaní roku 2010 s rokom 2009 však zistujeme, že návštevnosť prudko klesla o takmer 17%. To majiteľov firmy doviedlo k rozhodnutiu zamerať sa na long tail analýzu.

Analýza prebehla v niekoľkých krokoch – v štatistickom rozhraní Google Analytics sme nakonfigurovali segmentáciu kľúčových slov. Následne sa historické údaje za roky 2007-2010 analyzovali zvlášť za jednotlivé roky, čím vzniklo niekoľko rôznych segmentov. Tieto segmenty je možné medzi sebou porovnať a tak zistiť ako sa long tail vyvíjal historicky v čase.

Historický vývoj long tailových slov je možné sledovať aj graficky. Po vytvorení prispôbeného reportu Google Analytics vygeneruje graf, ktorý sa v čase mení. Ako je vidieť na Obr. 3, na osi x sa nachádzajú hľadané frázy usporiadané podľa počtu hľadání (os y) od najviac po najmenej vyhľadávané.



Obr. 3 Grafické zobrazenie Long Tailu v Google Analytics

Výsledkom analýzy je rebríček najkvalitnejších fráz skladajúcich sa z dvoch, troch a viacerých slov. V tabuľke 1 uvádzame prvých 10 najlepších fráz.

TABUĽKA 1
 ZOZNAM 10-TICH NAJLEPŠÍCH LONG TAILOVÝCH KĹÚČOVÝCH SLOV

	2-slovné	3-slovné	Viac slovné
1	iso 9001:2008	iso ts 16949	en iso 9001:2008
2	q system	ts 16 949	diagram príčin a následkov
3	interný audítor	valeo slovakia s.r.o	školenia z metrologie na rok 2009
4	nástroje kvality	en iso 9001:2008	školenie iso 9001:2008
5	školenie metrologia	školenie interný audítor	predám posuvné meradlo mitutoyo
6	školenie ppap	školenie interný audítor	školenie interný audítor ems
7	fmea príklad	valeo slovakia s.r.o.	školenie interný audítor iso 9001:2008
8	školenie fmea	školenie interný audítor	interný audítor ts 16949
9	školenie metrologia	školenie interných audítorov	školenia pre kontrolu kvality
10	ts 16949	predaj meracej techniky	analýza stromu poruchových stavov

Na základe získaných dát môžu majitelia firiem optimalizovať tak stránku, ako aj reklamné kampane na long tailové kľúčové slová, a tak získať výhodu oproti konkurencii.

V. ZÁVER

Cieľom analýzy bolo z množstva štatistických údajov o návštevnosti webovej stránky Q-System s.r.o. vybrať tie údaje, ktoré umožnia efektívne zvýšiť návštevnosť jej webovej stránok. Výstupom je zoznam najlepších long tailových kľúčových slov, ktoré poslúžia optimalizácii tak webovej stránky ako aj reklamných odkazov. Hlavným prínosom je umiestnenie sa na prvých priečkach vo vyhľadávačoch ako aj nižšia cena za klik pri PPC reklamách.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Clifton, B.: Google Analytics. Brno: Computer Press, a.s., 2009. 334 s.: il. ISBN 978-80-251-2231-0
- [2] Tušanová, A.: Vytvorenie webovskej aplikácie a jej vplyv na vývoj malej firmy. Bakalárska práca, KKUI FEI TU v Košiciach, 52s., 2008.
- [3] Zvyšovanie návštevnosti webu [online]. Dostupné na internete:
<<http://www.euroekonom.sk/marketing/internetovy-marketing/zvysovane-navstevnosti-webu/>>
- [4] Kľúčové slová [online]. Dostupné na internete: < <http://www.seochat.sk/klucove-slova/>>
- [5] Long Tail [online]. Dostupné na internete: <http://en.wikipedia.org/wiki/Long_Tail>

Rozpoznávanie slovných povelov s robotom NAO

Zlatko FEDOR, Peter SINČÁK

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

zlatko.fedor@tuke.sk, peter.sincak@tuke.sk

Abstrakt — V článku je stručný úvod do Melovských frekvenčných kepstrálnych koeficientov (MFCC), ktoré sú dominantou v predspracovaní reči, redukujú informačný tok a zároveň do značnej miery kopírujú proces predspracovania reči v ľudskom mozgu.

Po znížení informačného toku a vytvorení dekorelovaných Melovských vektorov jednotlivé slová klasifikujeme pomocou metódy Dynamic Time Warping (DTW). DTW je algoritmus pre meranie podobnosti medzi dvoma sekvenciami (vektormi), ktoré môžu byť rôzne v čase, ale aj rýchlosti.

V závere popisujeme systém, ktorý umožňuje robotovi Nao od firmy Aldebaran Robotics rozpoznávať a učiť sa nové slová pomocou jednoduchého dialógu s človekom.

Kľúčové slová — NAO, MASS, DTW, rozpoznávanie reči

I. ÚVOD

Komunikácia prostredníctvom hovorenej reči je základným a najprirodzenejším spôsobom prenosu informácie medzi ľuďmi. Preto i dnes jednou s najpoužívanejších foriem interakcie človek - robot je práve reč.

V článku predstavíme rozpoznávanie slovných povelov pomocou systému MASS [1], ktorý s využitím metódy DTW rozpoznáva slovné povel. Algoritmus DTW je invariantný voči času a preto dosahuje vysokú presnosť aj pri veľmi malej trénovacej množine nahovorených slov. To umožňuje učenie nových povelov bez zbytočného viacnásobného opakovania povelu.

Jednotlivé textové prepisy slov sú následne preposielané robotovi NAO, ktorý na ne reaguje. Naším cieľom je vytvoriť systém schopný učiť sa nové príkazy v rôznych jazykoch pomocou jednoduchého dialógu s človekom.

Základnou myšlienkou celého systému je referenčná databáza povelov v anglickom jazyku. S jej využitím sme schopní zostaviť dialóg človek - robot a umožniť učenie povelov v iných jazykoch.

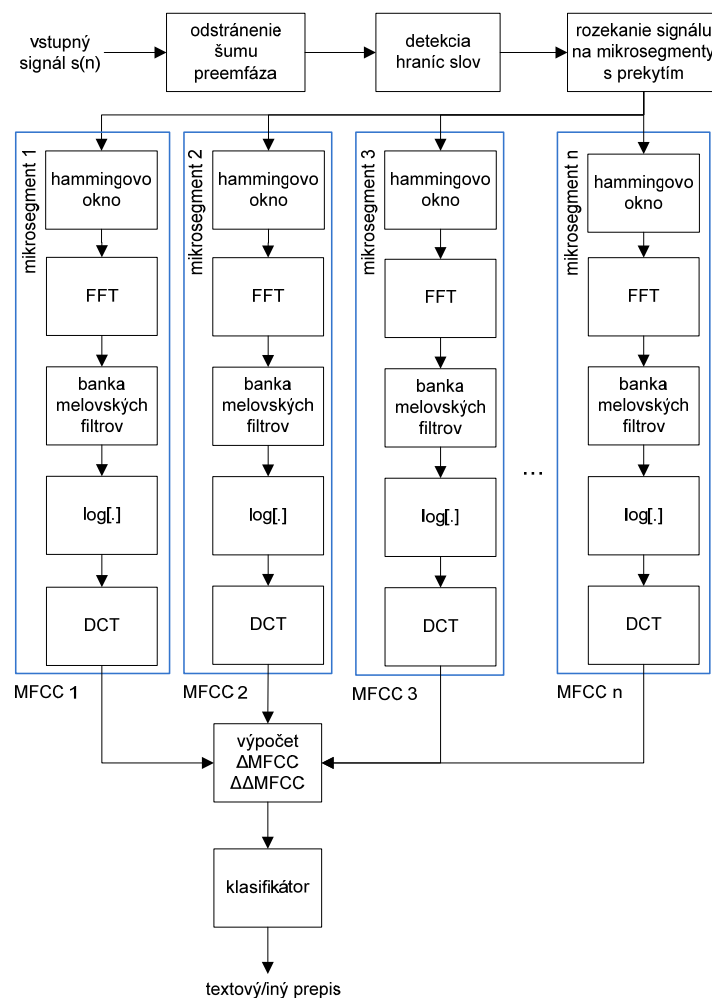
II. MELOVSKE FREKVENČNÉ KEPSTRÁLNE KOEFICIENTY

Algoritmus predspracovania rečového signálu pomocou MFCC je do značnej miery navrhnutý tak, aby čo najviac rešpektoval nelineárne vlastnosti vnímania zvuku ľudským uchom. Jeho hlavnou úlohou je kompenzovať nelineárne vnímanie frekvencií, a to pomocou banky trojuholníkových pásmových filtrov s lineárnym rozložením frekvencií v takzvannej melovskej frekvenčnej škále [2].

Nahráný signál je rozdelený na menšie časové úseky (mikrosegmenty), ktoré sú čiastočne prekryté. Na jednotlivé prekryté mikrosegmenty sú postupne aplikované tieto bloky:

1. Hammingovo okno,
2. rýchla Fourierova transformácia (FFT),
3. banka Melovských filtrov,
4. logaritmus,
5. diskretná kosínusová transformácia.

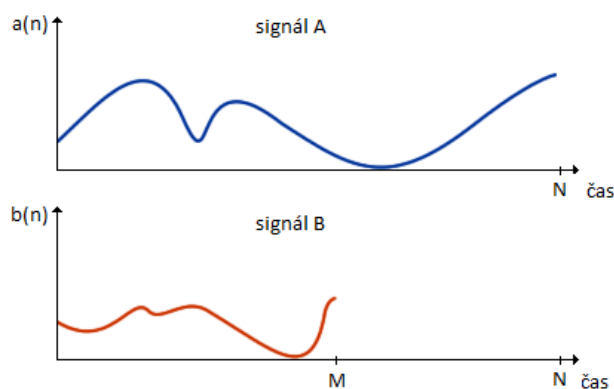
Na nasledujúcom obrázku je znázornený výpočet týchto koeficientov.



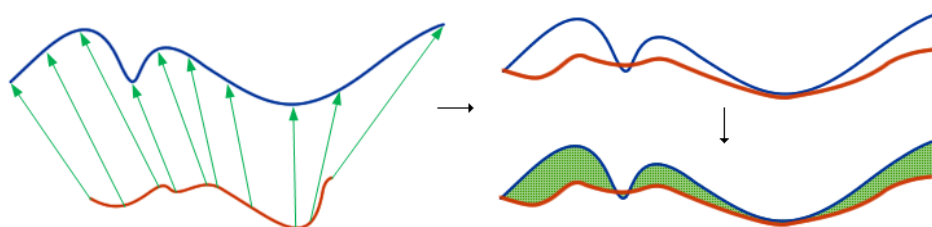
Obr. 1 Základná schéma výpočtu MFCC pri rozpoznávaní slov

III. DYNAMIC TIME WARPING

Algoritmus Dynamic Time Warping (DTW) [3] je v súčasnosti často používaný v rôznych oblastiach, ako napríklad rozpoznávanie reči, počítačové videnie, data mining a mnoho iných. Je založený na metóde dynamického programovania a je nesmierne výkonný pri porovnávaní rôznych signálov s minimalizovaním ich posunu alebo skreslenia v čase. Myšlienkou algoritmu je nájsť minimálnu plochu medzi dvoma signálmi, vzhľadom na všetky prípustné transformačné cesty a rozťahnutím kratšieho signálu na dĺžku dlhšieho. Základný algoritmus DTW vždy porovnáva iba dva vzory, referenčný a testovací (pre porovnávanie viacerých referenčných vzorov naraz je možné použiť Multi Pattern Dynamic Time Warping (MPDTW) [4]). V prípade rozpoznávania reči počíta medzi nimi mieru akustickej odlišnosti. V jednoduchosti si prácu algoritmu môžeme znázorniť graficky. Majme dva signály referenčný $A=(a_1, \dots, a_N)$ a testovací $B=(b_1, \dots, b_M)$.

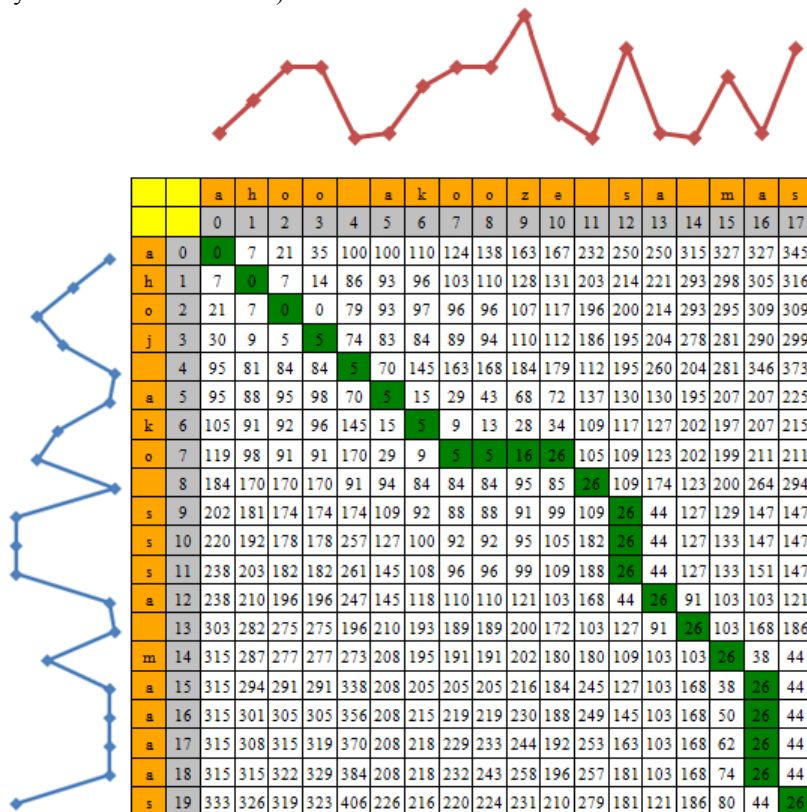


Obr. 2 Referenčný signál A a testovací B



Obr. 3 Nájdenie minimálnej plochy medzi signálmi A a B

Algoritmus dokáže porovnať tieto dva signály v čase $O(NM)$. Pri porovnávaní sa vytvorí takzvaná DTW matica s rozmermi $N \times M$, ktorú je možné vidieť na obrázku 4. V nej sa hľadá taká cesta (Warping funkcia) naprieč celou maticou, aby rozdiel signálov bol minimálny (na obrázku zvýraznená zelenou farbou).



Obr. 4 Ukážka DTW matice pre porovnanie dvoch viet

Jestvuje viacero možností ako vypočítať hodnoty matice. Najčastejší spôsob je:

$$DTW(x, y) = rozdiel(x, y) + \begin{cases} 0, & x = 1, y = 1 \\ \min \begin{cases} DTW(x-1, y) \\ DTW(x, y-1) \\ DTW(x-1, y-1) \end{cases}, & \text{inak,} \end{cases} \quad (1)$$

kde x, y je pozícia v matici pričom $x \in \langle 1, N \rangle$ a $y \in \langle 1, M \rangle$. Hodnoty v matici sa počítajú z bodu $[1;1]$ vždy zľava doprava, riadok po riadku. V prípade, že sa pri výpočte narába s prvým riadkom alebo stĺpcom, narazíme na to, že budeme potrebovať hodnotu, ktorá je mimo matice. Taká hodnota je vždy rovná plus nekonečno. Funkcia rozdiel (x, y) ráta vzdialenosť medzi jednotlivými vzorkami signálov. Ak je vzorka vektor hodnôt je možné pri výpočte použiť rôzne metriky. Najčastejšie používaná je Euklidova a Manhattan metrika.

Pre výsledný výpočet miery podobnosti dvoch signálov postačuje posledná hodnota z DTW matice. A teda výslednú mieru podobnosti signálov A a B vypočítame podľa:

$$D(A, B) = \frac{DTW(N, M)}{N + M}, \quad (2)$$

kde $DTW(N, M)$ je hodnota v DTW matici na pozícii N, M .

IV. ZÁVER

V článku je predstavený inkrementálny učiaci sa systém schopný učiť sa nové slová v rôznych jazykoch. Tento systém je naprogramovaný a plne funkčný. Problém, ktorý musíme vyriešiť je narastajúci čas potrebný pri klasifikácii povelov. S každým novým slovom sa tento čas zvyšuje, pretože pri klasifikácii je nutné prejsť všetky povel, ktoré sa v databáze nachádzajú a vypočítať pre ne DTW maticu.

Navrhli sme niekoľko urýchlení a vylepšení algoritmu, ktoré budeme experimentálne testovať v našom ďalšom výskume.

Taktiež do budúcnosti plánujeme okrem učenia sa nových povelov zakomponovať aj učenie sa nových pohybov s využitím systému kinect. Tieto pohyby budú následne dialógom s robotom previazané so slovnými povelmi.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] FEDOR, Z.: Spracovanie povelov prostriedkami umelej inteligencie, písomná práca k dizertačnej skúške, 2010.
- [2] PSUTKA, J.: Mluvíme s počítačem česky. Academia, Praha 2006. ISBN 80-200-1309-1.
- [3] SENIN, Pavel: Dynamic Time Warping Algorithm Review, Information and Computer Science Department University of Hawaii at Manoa, Honolulu, USA, 2008.
- [4] NAIR, Nishanth.U - SREENIVAS, T.V.: Multi Pattern Dynamic Time Warping for automatic speech recognition, Nair, N.U.; Sreenivas, T.V., TENCON 2008 - 2008, TENCON 2008. IEEE Region 10 Conference 19-21 Nov. 2008 Page(s):1 – 6, Digital Object Identifier 10.1109/TENCON.2008.4766617.

Mobile Robot Navigation Using Fuzzy Cognitive Maps

Martin PALA, Ján VAŠČÁK

Dept. of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

martin.pala@ymail.com, jan.vascak@tuke.sk

Abstract — This paper shows one of the many possible applications of fuzzy cognitive maps in mobile robot navigation. The description and explanation of the method of image recognition, calculation and execution of efficient intervening cause obtained using fuzzy cognitive map are provided in this work. The main purpose is to scheme out a vehicle capable to avoid the obstacles using single camera sensor. The first part of this paper is dedicated to the architecture and hardware configuration of the mobile robot. In the following section the image recognition and fuzzy cognitive map control system is described. The last section is dedicated to the conclusion of the project and the possible improvements considered for the future work.

Keywords — fuzzy cognitive map, image recognition, mobile robot, navigation

I. INTRODUCTION

Navigation is the art of getting from one place to another, safely and efficiently. The means of navigation have changed dramatically over the past decades. From celestial navigation to GPS the methods used for navigation have been upgraded, the share of human factor in navigation has been decreased and it is pushing the human more and more to the role of the observer. This goal can be reached by applying suitable methods of artificial intelligence [2-3].

The robot navigation problem can be split into two main areas of research: Local navigation and global navigation. The local navigation problem deals with navigation on the scale of a few meters, where the main problem is obstacle. And the global navigation problem deals with navigation on a larger scale in which the robot cannot observe the goal state from its initial position [1].

The main topic of this work is to build a vehicle capable of autonomous navigation and obstacle avoidance in an unknown environment using only one camera as a sensor and fuzzy cognitive map (FCM) as a reasoning component for the movement control. The architecture of the vehicle control and hardware configuration is described in the following section. The second section is dedicated to image recognition system and evaluation process of the efficient intervening cause obtained using fuzzy cognitive map. The last two sections describe the experiments performed and conclusion of the proposed navigation system using fuzzy cognitive map approach.

II. ARCHITECTURE OF THE AUTONOMOUS VEHICLE CONTROL

A. Architecture of the autonomous vehicle control and hardware configuration

The architecture of the vehicle control consists of two main parts: Lego NXT vehicle and computer. The vehicle is built by the Lego NXT Mindstorms parts and IP camera A-link IPC2 as the only sensor used in this mobile robot. The computer is not a part of the vehicle but is used to gather the data from the IP camera using WIFI connection. Then the data are evaluated and sent as instructions back to Lego NXT brick via Bluetooth. The following picture shows how the communication and data evaluation works.

Raw images are sent from IP camera to the computer via WIFI where are processed and the position of obstacles are evaluated. Output generated by the agent for image recognition is forwarded as the input to the agent for fuzzy cognitive map evaluation. The output from the fuzzy cognitive map agent represents the action required to avoid the obstacle. This is transformed to the suitable instruction format that is sent to NXT brick.

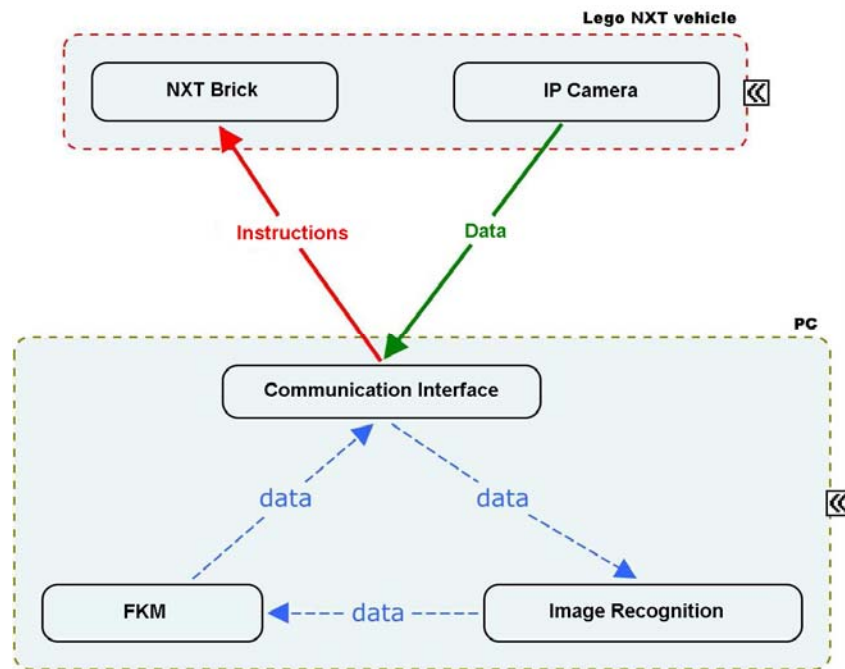


Fig. 1 Architecture of autonomous vehicle control system based on single camera vision for obstacle detection.

III. OBSTACLE DETECTION AND FCM EVALUATION

A. Image Recognition

The images obtained from the IP camera contain a lot of information about the surrounding area of the mobile robot. To extract the information about the obstacles the canny edge detection method is used to lower the amount of unnecessary information and point up the obstacle characteristics. The following picture shows the contrast between the images acquired from IP camera and preprocessed image by image recognition agent implemented for this project.

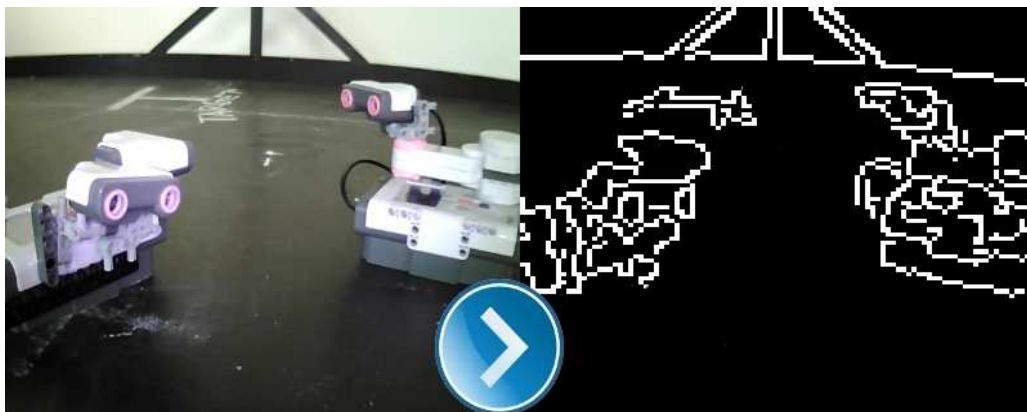


Fig. 2 IP camera image preprocessing using Canny Edge Detection (CED) method to lower the amount of image information.

After image preprocessing the image contains only information about the obstacles and free space where the mobile robot can be moved. To define these areas, image is transformed from bitmap representation to matrix representation so the computational methods are available for further information processing. The edge of the obstacle, white pixel in bitmap representation is transformed to number one in matrix representation and the black pixels defining free space are transformed to number zero in matrix representation. Figure 3 shows the matrix representation of image after the application of CED. In addition the image is divided in three zones and together in seven areas. The green zone represents the areas that are further from the mobile robot. The yellow zone is closer and the red zone is critically close to the mobile robot.



Fig. 3 Binary matrix image representation for obstacle detection system. Note: the red zone is not used in standard image recognition process. This zone is calculated independently and if it is evaluated as an obstacle, the emergency break of the robot is triggered.

Each area represents the position of the obstacle or position of free passable area in front of the vehicle. For example the top left area represents the obstacle or the passable area far left from the robot. To determine which area is considered to be an obstacle and which area is considered to be a passable area further operations are needed. The number of non null rows is counted for each sub-matrix. If the number of non null rows is greater than a threshold value, the area represents an obstacle. The vector of obstacles is generated by applying this algorithm on all the sub-matrices. The full vector representing the obstacles and passable areas can be generated by adding the same vector with the switched values to the end of the vector of the obstacles. Figure 4 shows the example of the full vector of the obstacles and passable areas for the image represented in the figure 3. The threshold value is set to five.

$$P_u = [1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0]$$

Fig. 4 Full vector defining the obstacles and passable areas. Note: First six members of the vector represent the obstacle vector described in the text of this paper. Last six members represent the added vector (vector of passable areas) generated from the obstacle vector by switching the values of the members from number one to zero and from number zero to number one.

The full vector generated from the obstacle vector and vector of passable areas is in the suitable format for the input of the fuzzy cognitive map. Each member of this vector represents the activation of the each input node in fuzzy cognitive map.

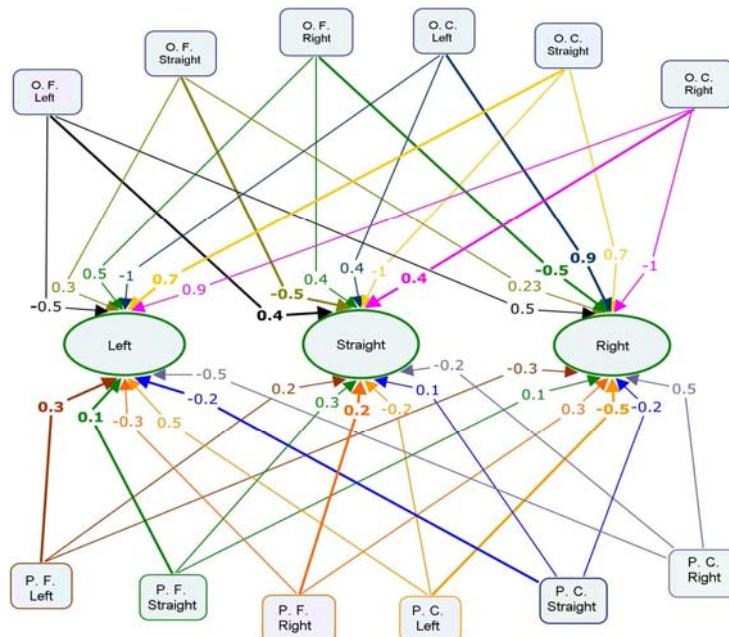


Fig. 5 Fuzzy cognitive map used to evaluate the efficient intervening cause.

A. Fuzzy cognitive map evaluation

The fuzzy cognitive map is designed by a human without an adaptation features. The output from the image recognition system described above is used as an input for fuzzy cognitive map. The following picture shows the FCM used for the mobile robot used in this project. The fuzzy cognitive map consists of fifteen nodes and casual connections between these nodes. The input nodes are on the top and on the bottom of the FCM. The output nodes are in the middle. The activation of the nodes on the top of the map represents the obstacles. O. F. Left represents the obstacle far left and the activation equals to the first member of the full vector of obstacles and passable areas described in section dedicated to image recognition system. O. C. Right represents the obstacle close right and the activation equals to the sixth member of the vector. P. F. Left node represents the passable area far left and the activation equals to the seventh member of the vector. P. C. Right stands for passable area close right and the activation equals to the last member of the vector.

The output from the FCM is the angle that the mobile robot needs to turn to avoid the obstacle. The membership functions of the output nodes are shown in the following picture.

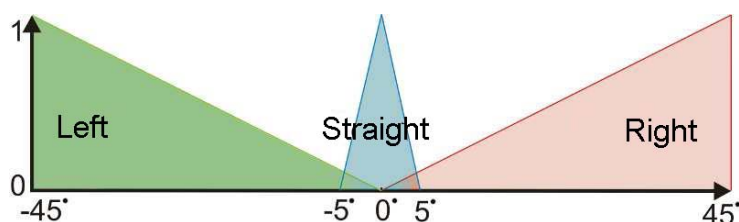


Fig. 6 Membership functions of the output nodes of the fuzzy cognitive map. Note: The Y axis represents the membership degree and the X axis represents the value of the angle.

After the FCM is evaluated the final value has to be defuzzified to get the sharp value of the angle. In this project the center of gravity defuzzification method has been used. In the end the sharp value is transformed to the suitable instruction for the NXT brick and is sent to the vehicle via Bluetooth.

IV. CONCLUSION

The Proposed obstacle avoidance and navigation system has been tested in the environment with the various numbers of obstacles and obstacle alignments. The autonomous vehicle smoothly detected and avoided single obstacle. In the environment with the multiple obstacles the mobile robot also proved that it is capable of multiple obstacle avoidance. In some cases the robot had to use the critical brake system to not to drive in the obstacle, but after a few new tries it smoothly avoided all the obstacles. This error might appear in the areas where the illumination of the testing surface was not adequate. To improve the whole system performance some extensions should be considered. Adaptive fuzzy cognitive map system might improve the smoothness of the movement of the vehicle as well as the improvement of the proposed image recognition system to the dynamic image area distribution.

ACKNOWLEDGMENT

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] M. A. Batalin, G. S. Sukhatme, "Mobile Robot Navigation using a Sensor Network," In IEEE International Conference on Robotics and Automation, pp. 636-642, New Orleans, LA, April 26 - May 1, 2004.
- [2] K. Mls.: Implicit knowledge in Concept Maps and their Revealing by Spatial Analysis of Hand-Drawn Maps, Proc. of the Second International Conference on Concept Mapping - Concept Maps: Theory, Methodology, Technology, Vol. 2, San José, Costa Rica, 2006.
- [3] Vaščák, J.: Fuzzy cognitive maps in path planning / Ján Vaščák - 2008. In: Acta Technica Jaurinensis. Vol. 1, no. 3, pp. 467-479, 2008.

Modifikácie siete MF ARTMAP pre zvýšenie presnosti klasifikácie a zvýšenie generalizačnej schopnosti

Peter SMOLÁR, Peter SINČÁK

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

peter.smolar@tuke.sk, peter.sincak@tuke.sk

Abstrakt — V tejto práci sa venujeme návrhu modifikácie neurónovej siete MF ARTMAP. Tieto návrhy vyšli z analýzy tejto siete počas fázy učenia a testovania hlavne pri klasifikačných úlohách objektov na obraze a týkajú sa minimalizácie počtu vytvorených fuzzy zhlukov a matematického popisu fuzzy relácie, ktorou sa počíta príslušnosť vstupnej vzorky ku fuzzy zhluoku. Pri týchto zmenách očakávame zvýšenie presnosti a generalizačnej schopnosti tejto siete pri akýchkoľvek klasifikačných úlohách a taktiež zrýchlenie úlohy testovania na tejto neurónovej sieti.

Kľúčové slová — neurónová sieť MF ARTMAP, normálne rozdelenie pravdepodobnosti

I. ÚVOD

V súčasnosti si informačné technológie našli svoje miesto asi v každej oblasti ľudskej činnosti. Ich použiteľnosť ďalej rastie aj v závislosti zvyšujúcej sa automatizácie. S prudkým nárastom aplikačnej využiteľnosti informačných technológií takisto stúpajú aj nároky na rýchle spracovanie veľkého množstva dát, tým pádom vzniká tlak aj na používané metódy z dôvodu získavania relevantných informácií, ktoré zákazník požaduje.

Jedným zo spôsobov, ako sa dá pracovať s dátami je použitie metód umelej inteligencie. Tieto riešenia sú špecifické tým, že si „inšpiráciu“ berú z procesov prebiehajúcich v prírode. Dôležitou vlastnosťou niektorých týchto metód je schopnosť adaptovať sa na dynamicky meniace sa problémy, schopnosť učiť sa, schopnosť narábať s neurčitosťou.

Jednou z hlavných metód výpočtovej umelej inteligencie sú neurónové siete. Podmnožina neurónových sietí, ktorej sa budeme venovať ďalej, a ktorou sa riešia hlavne problémy klasifikácie alebo zhlukovania, je trieda sietí ART (Adaptive Resonance Theory – teória adaptívnej rezonancie). Táto teória sa zaoberá problémom tzv. stability-plasticity, čo znamená, že neurónová sieť dokáže spracovať vstupné vzorky, ktoré sú totožné alebo dostatočne „podobné“ už naučeným vstupným vzorkám, ale taktiež sa vie adaptovať na nové vstupné vzorky, ktoré doteraz neboli na vstupe. Členom tejto rodiny neurónových sietí je aj sieť MF ARTMAP (Membership Function ARTMAP), ktorá v sebe spája spolu fuzzy prístup a ART prístup. Inak povedané, že okrem samotnej klasifikácie vstupnej vzorky do tried nám dáva ešte aj informáciu o jej veľkosti príslušnosti k týmto triedam. Danej neurónovej sieti budeme v tomto článku venovať najviac pozornosti s cieľom navrhnúť také modifikácie, aby neurónová sieť dávala na výstupe presnejšie hodnoty príslušnosti k fuzzy triedam a aby pracovala rýchlejšie v normálnom režime.

II. NEURÓNOVÁ SIETĽ MF ARTMAP

Inšpiráciu pre vývoj siete MF-ARTMAP [1] môžeme nájsť v snahe vedcov získať nástroj, ktorý by umožňoval lepšie reprezentovať dáta a vytvárať znalosť v zrozumiteľnejšej podobe. Pri úlohách kontrolovanej klasifikácie sa neraz ťažko rozhoduje o príslušnosti vstupnej vzorky ku niektorej z tried, čo má vplyv na výslednú nepresnosť klasifikácie systému. Keďže vzorky jednej triedy sa môžu nachádzať vo viacerých zhlucoch, situácia je ešte komplikovanejšia. Neurónová sieť MF-ARTMAP však dáva riešenie, ktoré je založené na skombinovaní teórie

fuzzy množín a teórie adaptívnej rezonancie (ART). Vďaka tomu sa dá vyčísliť tzv. „príslušnosť“ konkrétneho prípadu (vstupnej vzorky) k daným triedam. To znamená, že výsledná hodnota klasifikácie nie je „ostré“ rozhodnutie o tom, či vstupná vzorka je členom alebo nie je členom konkrétnej triedy, ale vektor hodnôt, ktorý popisuje „stupeň príslušnosti“ tohto príkladu k všetkým triedam.

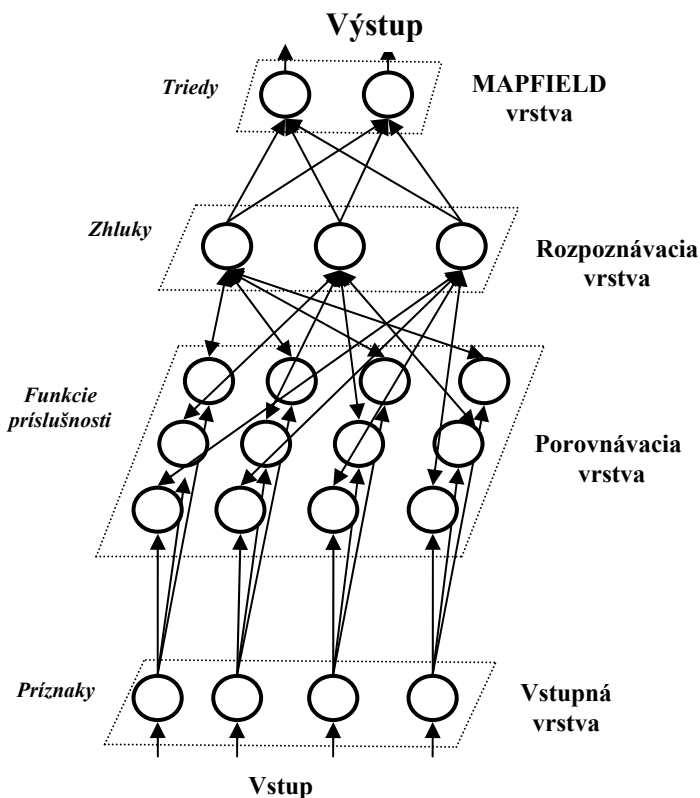
Reprezentácia znalostí

Pri reprezentácii znalostí sa vychádza z hypotézy, že jednotlivé príklady vo vstupnom priestore sa nachádzajú vo „fuzzy zhlukoch“ [2]. Teda, vstupný priestor predstavuje univerzum fuzzy množín, čo znamená, že sa dá akémukoľvek bodu (príkladu) x zo vstupného priestoru vyčísliť hodnota funkcie príslušnosti ku konkrétnemu fuzzy zhluku $\mu_A(x)$ konkrétnej „fuzzy“ triedy.

Každý zhluk predstavuje vzťah medzi vstupnými vzorkami a je definovaný pomocou fuzzy množiny, čo znamená, že je možné ho popísať pomocou fuzzy relácie s parametrami, kde sa práve parametre fuzzy relácie adaptujú v procese učenia, sú teda nositeľmi znalostí [2].

Pre správnu voľbu parametrickej funkcie, ktorou sa fuzzy relácia popisuje, sú kladené určité nároky, ktoré vyplývajú z vlastností celej metódy.

Niektoré z týchto kritérií sú [2]: možnosť meniť tvar fuzzy relácie cez parametre tak, aby bolo možné čo najlepšie opísať nelinearity z popisu vstupných dát, možnosť reprezentovať znalosti ako pravidlá, resp. v inej „zrozumiteľnej“ forme, možnosť explicitného vkladania resp. úpravy znalostí.



Obr. 1 Základná topológia neurónovej siete MF ARTMAP

Vzhľadom na tieto požiadavky je fuzzy relácia v sieťach MF ARTMAP definovaná nasledovne [3]:

$$A(\bar{x}_i) = \int_U \frac{1}{1 + \left| \left(\frac{a_{ij} - x_i}{f_k} \right)^2 \right|} / \bar{x}_i \quad (1)$$

kde jednotlivé indexy znamenajú:

- index m predstavuje počet dimenzií vstupného priestoru,
- $j=1, 2, \dots, m$ predstavuje index dimenzie vstupného priestoru,
- U predstavuje označenie vstupného priestoru,
- index n predstavuje počet všetkých vzoriek v trénovacej množine,

- i predstavuje index aktuálnej vzorky,
- index p predstavuje počet doteraz vytvorených fuzzy zhlukov,
- $k = 1, 2, \dots, p$ predstavuje index aktuálneho fuzzy zhuku,
- označenie $A(x_i)$ predstavuje príslušnosť vzorky x_i k fuzzy zhuku k ,
- vektor x_i predstavuje ľubovoľný bod vo vstupnom priestore,
- vektory x_{sk} , E_k a F_k sú parametre fuzzy zhuku k , kde vektor x_{sk} predstavuje stred fuzzy zhuku k .

III. NÁVRHY MODIFIKÁCII MINIMALIZÁCIE POČTU FUZZY ZHLUKOV

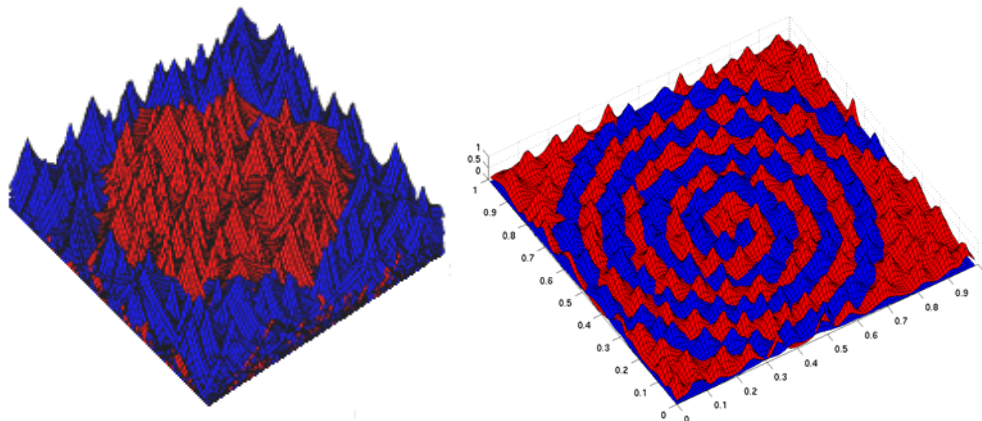
Analýzou správania sa siete MF ARTMAP počas učenia a analýzou výsledkov testovania, boli nájdené určité „slabé miesta“ tejto neurónovej siete, ktoré by sa dali zhrnúť v nasledujúcej vete: Pri klasifikácii vstupné vzorky nadobúdajú rôzne hodnoty príslušnosti ku fuzzy triedam, ktoré však nezodpovedajú žiadaným hodnotám.

Na základe tohto zistenia bola navrhnutá zmena procesu učenia neurónovej siete MF ARTMAP z účelom minimalizovať počet fuzzy zhlukov, čím sa má zvýšiť presnosť klasifikácie vzoriek, ktoré sa inak nachádzali na okrajoch dvoch fuzzy zhlukov prináležiacich k tej istej fuzzy triede.

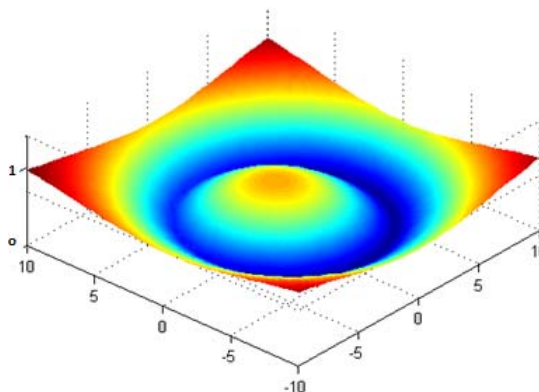
A. Analýza vytváraných fuzzy zhlukov

Ako ilustruje obrázok „kruh vo štvorci“ a „špirála“ (Obr. 2) ak vstupné vzorky do siete MF ARTMAP vstupujú v náhodnom poradí, sieť vytvára veľké množstvo nových fuzzy zhlukov v závislosti od parametra prahu, aj keď tieto fuzzy zhluky sú vedľa seba a patria do rovnakej fuzzy triedy. Ideálne vytvorenie fuzzy zhuku pre obrázok „kruh vo štvorci“ z hľadiska počtu zhlukov ilustruje obrázok (Obr. 3), kde objekt kruhu reprezentuje jeden fuzzy zhuk.

Veľký počet fuzzy zhlukov má za následok, že tie vzorky, ktoré sú na okrajoch dvoch fuzzy zhlukov a patria do tej istej fuzzy triedy, majú nízku hodnotu príslušnosti k fuzzy triede, aj keď patria danej triede rovnako ako, iné vzorky.



Obr. 2 Obrázky "kruh vo štvorci" a "špirála" so znázornenými tvarmi zhlukov a ich funkcií príslušnosti [3]



Obr. 3 Ideálne nastavenie fuzzy relácie siete MF ARTMAP z hľadiska počtu vytvorených fuzzy zhlukov na testovacom obraze "Kruh vo štvorci". Triedu „Kruh“ predstavuje len jeden zhuk [5]

B. Navrhované riešenia pre minimalizáciu počtu fuzzy zhlukov pre jednu fuzzy triedu a zvýšenie presnosti hodnoty príslušnosti ku fuzzy zhuku

Boli navrhnuté tri riešenia na minimalizáciu počtu fuzzy zhlukov pre jednu fuzzy triedu a zvýšenie presnosti hodnoty príslušnosti ku fuzzy zhuku. Tieto riešenia sú:

Prvé riešenie spočíva v zistení, že určitý vplyv na učenie siete má poradie tréningových vzoriek. Riešenie spočíva vo vybratí vzorky z tréningovej množiny do siete podľa stanovenej blízkosti (neprekračujúcej parameter prahu) k fuzzy zhuku, ktoré už boli vytvorené. Tým pádom by sa vytváral nový fuzzy zhuk až vtedy, keby žiadna vstupná vzorka nemohla byť klasifikovaná k už vytvoreným fuzzy zhukom.

Iná varianta modifikácie priebehu učenia spočíva v spustení učiaceho algoritmu neurónovej siete viackrát avšak s iným poradím vstupných vzoriek. Následne sa vykoná zjednotenie fuzzy relácií z jednotlivých behov učiaceho algoritmu a tým sa zvýši aj presnosť hodnoty funkcie príslušnosti vstupnej vzorky ku fuzzy triede. Nevýhoda tohto riešenia je v tom, že počet zhlukov a tým aj počet neurónov v sieti sa znásobí podľa počtu behov.

Tretia možnosť modifikácie spočíva vo vytvorení nadstavby učiaceho algoritmu neurónovej siete MF ARTMAP, ktorý bude zisťovať na konci učenia, či okolité fuzzy zhuky konkrétneho fuzzy zhuku patria do spoločnej fuzzy triedy, alebo nie. Potom na základe okolia fuzzy zhuku, konkrétny zhuk „pohltí“ okolité zhuky, ak tieto fuzzy zhuky patria do rovnakej fuzzy triedy. Nižší počet fuzzy zhlukov bude mať za následok rýchlejší beh algoritmu a presnejšie určenie príslušnosti vzorky k triede.

IV. ZÁVER

Od navrhnutých modifikácií sa hlavne očakáva zvýšenie presnosti určovania stupňa príslušnosti vstupných vzoriek ku fuzzy triedam znížením počtu fuzzy zhlukov jednej fuzzy triedy dosiahnuť

Vo svojej práci by som sa chcel ďalej sústrediť na analýzu a porovnanie mnou navrhnutých modifikácií s klasickou MF ARTMAP sieťou a podobnými sieťami typu ARTMAP.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] P. Maliňák, Inkrementálne učiace algoritmy pre rozpoznávacie systémy, Pisomná práca k dizertačnej skúške, 2008.
- [2] A. Bodnárová. Metódy zhukovania dát s využitím neurónových sietí typu ART. Diplomová práca, Technická univerzita v Košiciach, 2006.
- [3] M. Hric. Inkrementálne učenie na báze ART neurónových sietí. Dizertačná práca, Technická univerzita v Košiciach, 2003.
- [4] P. Sinčák and G. Andrejková. Neurónové siete - Inžiniersky prístup (2. diel). Elfa, Košice, 1996.
- [5] P. Smolár, Rozpoznávanie obrazu neurónovými sieťami typu ART, Pisomná práca k dizertačnej skúške, Technická univerzita v Košiciach, 2011.

Riadenie pohybu robota NAO pri kope do lopty v robotickom futbale

Jaroslav TUHÁRSKY, Ján VAŠČÁK

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jaroslav.tuharsky@tuke.sk, jan.vascak@tuke.sk

Abstrakt — Tento článok pojednáva o spôsobe riadenia pohybu robota pri kope do lopty, ktorý bol použitý najlepšími tímami v súťaži RoboCup 2010. Kop do lopty je po chôdzi druhým najdôležitejším pohybom v robotickom futbale a preto je mu venovaná patričná pozornosť. V predchádzajúcich ročníkoch tejto súťaže tímy použili na návrh pohybových vzorov pre kop do lopty metódu tzv. „key-frames“ čo boli pevne nadefinované trajektórie pohybu expertom. Nevýhodou tohto prístupu je to že dané trajektórie sú určené „offline“ a z tohto dôvodu ich nie je možné adaptovať na zmenenú situáciu. Kop implementovaný takýmto spôsobom mal iba jednu trajektóriu, umožňujúcu kop len priamym smerom. Metóda popísaná v tomto článku adaptuje trajektóriu pohybu „online“ t.j. počas jeho vykonávania a umožňuje kop robota do lopty viacerými smermi.

Kľúčové slová — NAO, RoboCup, robotický futbal, Bézierová krivka

I. ÚVOD

Dôkazom snahy preferovať humanoidnú robotiku je aj súťaž RoboCup, kde roboty medzi sebou súťažia v hraní futbalu. Základnou myšlienkou vzniku tejto súťaže je snaha o usporiadanie futbalového zápasu, medzi majstrami sveta a tímom tvoreným z humanoidných robotických futbalistov, v roku 2050. Ide o veľmi odvážny cieľ s veľkou výzvou pre humanoidnú robotiku, tak z hľadiska konštrukčného vyhotovenia robotov, aby títo boli schopní dostatočne rýchlych a obratných pohybov, ako aj z hľadiska ich softwarového vybavenia, hlavne čo sa týka schopnosti generovať dostatočne rýchly a dynamický pohyb so zaistením dostatočnej stability takéhoto pohybu.

Štandardnou platformou (SPL) v RoboCup-e sa od roku 2008 stal humanoidný robot NAO, keď vystriedal robotického psa AIBO, čím sa začal preferovať vývoj humanoidného robotického futbalistu v tejto súťaži.

II. PROBLEMATIKA KOPU DO LOPTY ROBOTOV NAO VO FUTBALE

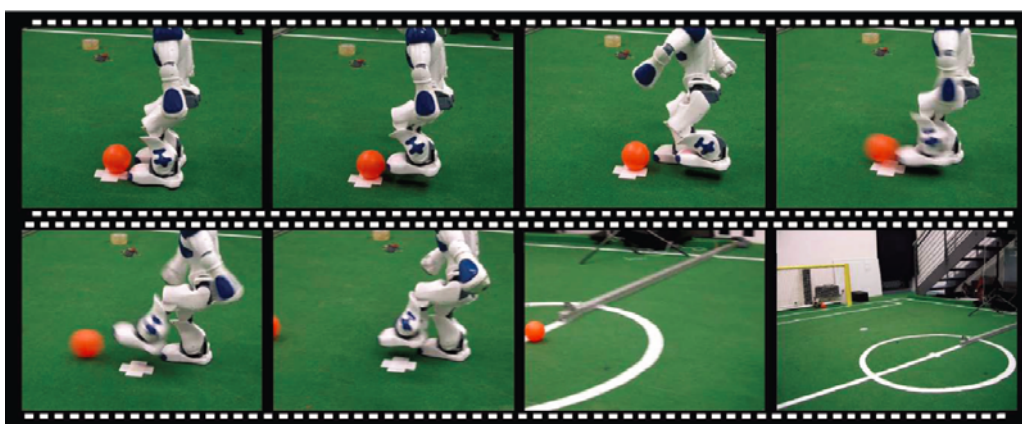
Z dôvodu úspešného zvládnutia hry futbal s humanoidným robotom, je veľká pozornosť venovaná vývoju chôdze a to nielen čo sa týka jej rýchlosti, ale aj jej robustnosti, tak aby bola zaistená dostatočná flexibilita v prípade náhlej zmeny smeru chôdze ako aj schopnosť kompenzovať rušivé vplyvy okolia, spôsobené napr. kolíziami s ďalšími robotmi, alebo zmenou drsnosti povrchu. Podobné požiadavky na robustnosť sú aj v prípade druhého najdôležitejšieho pohybu vo futbale po chôdzi a to kopu do lopty. Počas vykonávania pohybu kopu do lopty sa môže pozícia lopty, t.j. cieľová pozícia pohybu dynamicky meniť a to z rôznych dôvodov, napr. že si robot loptu sám posunul v priebehu vykonávania tohto pohybu. V týchto prípadoch je veľmi dôležitá tzv. „online“ adaptácia trajektórie nohy vykonávajúcej pohyb kopu, tak aby táto presne trafila loptu, čo sa týka bodu a uhla dopadu z dôvodu dosiahnutia požadovaného smeru pohybu lopty. Navyše, aj nepatrné rušivé vplyvy počas vykonávania pohybu môžu spôsobiť stratu presnosti, alebo aj úplne zabrániť úspešnosti kopu. Preto mechanizmus, majúci na starosti vyrovňovanie rovnováhy robota stojaceho na jednej nohe, je kľúčovým faktorom pre úspešne vykonanie tohto pohybu.

III. RIADENIE DYNAMICKÉHO POHYBU PRI KOPE DO LOPTY

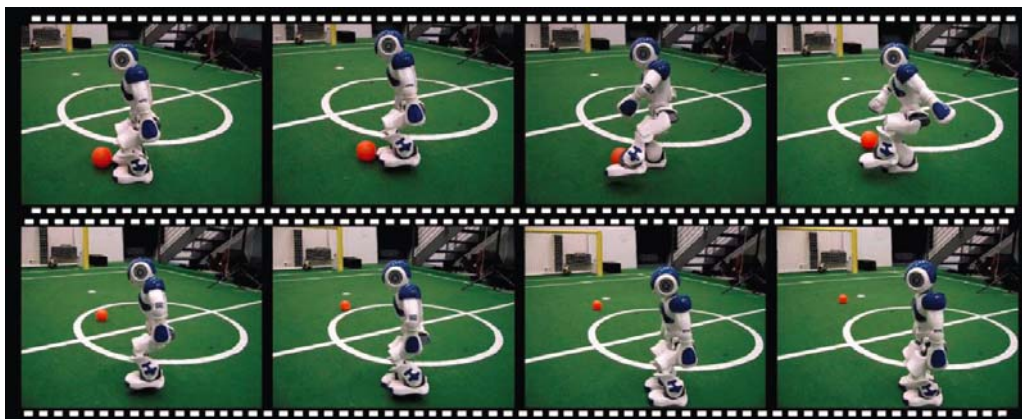
V predchádzajúcich ročníkoch (t.j. pred rokom 2010) súťaže RoboCup v kategórii „liga

štandardnej platformy“ (SPL) t.j. robotov NAO, všetky tímy použili pre kop do lopty priamy pohyb kopajúcej nohy. Z daného vyplýva, že lopta mohla byť kopnutá len jedným smerom a to priamo vpred. Tento zjednodušený návrh pohybu pri kope do lopty, si vyžadoval mnoho času na presné nastavenie sa robota na danú pozíciu oproti lopte, tak aby sa táto po vykonaní kopu pohybovala požadovaným smerom. V ročníku 2010 tímy B-human [1] a rUNSWiift [3] navrhli pohyb kopu do lopty, umožňujúci kopnutie lopty mnohými smermi z daného rozsahu, čo prinieslo značné ušetrenie času, potrebného na odkop lopty požadovaným smerom, pravdepodobne aj toto vylepšenie im pomohlo k dobrému umiestneniu sa v hlavnej časti súťaže RoboCup 2010, keď tím B-human skončil na prvom a tím rUNSWiift na druhom mieste. Tento prístup umožňuje určiť výsledný tvar krivky popisujúcej pohyb kopu „online“, t.j. v závislosti od aktuálnych hodnôt parametrov, akými sú poloha lopty a smer ktorým má byť táto kopnutá.

Hlavnou myšlienkou prístupu [2], použitého víťazným tímom *B-human* [1], je snaha pomocou kombinácie jednoduchých kriviek vytvoriť komplexné krivky, definujúce pohybovú trajektóriu a týmto spôsobom rozdeliť celkový pohyb do jednotlivých pohybových fáz, ako napr. zdvihnutie nohy, alebo kopnutie do lopty.



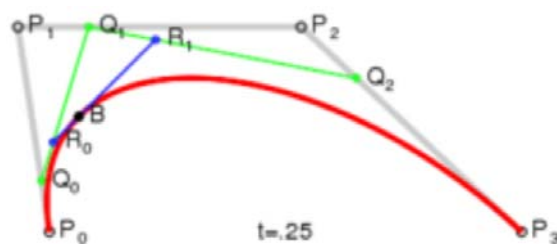
Obr. 1 Príklad kopu do lopty priamo robotom NAO (prevzaté z [2])



Obr. 2 Príklad kopu do lopty bokom robotom NAO (prevzaté z [2])

Na obrázkoch Obr. 1 a Obr. 2 je vidieť kop do lopty, ktorý bol zmenený „online“ počas jeho vykonávania. Robot podľa jeho postavenia k lopte a požadovaného smeru ktorým má byť lopta kopnutá rozhodol o trajektórii pohybu pri kope. na Obr. 1 je vidieť kop priamo a na Obr. 2 kop bokom. Výsledná trajektória kopu je poskladaná z dielčích trajektórií popísanými pomocou kubických Bézierových kriviek, t.j. $n = 3$, pozri rovnicu (1) a Obr. 3. Výsledný tvar Bézierových kriviek je riadený pomocou kontrolných bodov. V prípade kubickej Bézierovej krivky ide o štyri kontrolné body P_0, P_1, P_2, P_3 .

$$b(t) = \sum_{i=0}^3 \binom{3}{i} t^i (1-t)^{3-i} P_i \quad (1)$$



Obr. 3 Príklad kubickej Bézierovej krivky

IV. ZÁVER

Vo všeobecnosti riadenie bipedického pohybu nie je triviálnou záležitosťou a pohyb pri kope do lopty má svoje špecifiká. Tento pohyb by mal mať dostatočnú variabilitu, t.j. aby sa jeho trajektória mohla prispôbiť na zmenu situácie, tak aby robot mohol kopnúť loptu viacerými smermi z danej pozície, v ktorej sa momentálne nachádza a mohol ju tak kopnúť požadovaným smerom. Ako je vidieť z tohto článku dobrým spôsobom na dosiahnutie tohto cieľa je rozdeliť celkový pohyb na viacero fáz, každú fázu riadiť jednou Bézierovou krivkou a výsledný trajektóriu pohybu vytvoriť pospájaním daných Bézierových kriviek riadiacich jednotlivé parciálne trajektórie. Bézierové krivky sú vhodné na riešenie takejto úlohy a to z dôvodu že sa ich tvar dá jednoducho riadiť pomocou kontrolných bodov a je ich možné relatívne ľahko medzi sebou spájať. Týmto sa dosiahne požadovaná variabilita pohybu, riadená robotom „online“, t.j počas vykonávania daného pohybu, čím sa dosiahne schopnosť robota kopnúť loptu viacerými smermi. Táto variabilita pohybu pri kope do lopty riadená „online“ pravdepodobne pomohla tímom, ktoré ju implementovali do svojho systému, umiestniť sa na prvých miestach súťaže RoboCup 2010, keďže tím B-human sa umiestnil na prvom mieste a rUNSWift na druhom.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] T.Rofer, T.Laue, J.Muller, A.Burchardt, E. Damrose, A.Fabisch, F.Feldpausch, K.Gillmann, C.Graf, T.J. de Haas, A.Hartl, D.Honsel, P.Kastner, T.Kastner, B. Markowsky, M.Mester, J.Peter, O.Riemann, M.Ring, W.Sauerland, A. Schreck, I.Sieverdingbeck, F.Wenk, J.H.Worch: B-Human team report and code release 2010 (2010).
- [2] Judith Muller, Tim Laue, and Thomas Rofer. Kicking a Ball - Modeling Complex Dynamic Motions for Humanoid Robots. J. Ruiz-del-Solar, E. Chown, and P.G. Ploeger (Eds.): RoboCup 2010, LNAI 6556, pp. 109–120, 2011. Springer-Verlag Berlin Heidelberg 2011.
- [3] A.Ratter, B.Hengst, B. Hall, B.White, B.Vance, C.Sammut, D.Claridge, H.Nguyen, J.Ashar, M.Pagnucco, S.Robinson, Y.Zhu : rUNSWift Team Report 2010 Robocup Standard Platform League (2010).
- [4] Akin, H.L., Mericli, T., Ozkucur, K.C., Gökce, B.: Cerberus 2010 team description paper (2009).
- [5] Borisov, A., Ferdowsizadeh, A., Mohr, C., Mellmann, H., Martius, M., Krause, T., Hermann, T., Welter, O., Xu, Y.: NAO-Team Humboldt 2009 (2009).
- [6] Foley, J.D., van Dam, A., Feiner, S.K., Hughes, J.F.: Computer Graphics – Principles and Practice, 2nd edn. Addison-Wesley Publishing Company, Reading (1990), xXIII, 1174 S : Ill.
- [7] NAO user guide Ver. 1.10.10 (2011), dostupné na internete : http://users.aldebaran-robotics.com/docs/site_en/index_doc.html

Control of Mobile Robot Joints Using Fuzzy Logic

Ján VAŠČÁK

Dept. of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

jan.vascak@tuke.sk

Abstract — This paper deals with the control of mobile robot joints using fuzzy logic, namely the mobile robot NAO. The controller was designed manually and was tested on the so-called play Bago. The aim of this paper was to verify the possibility and suitability of using fuzzy approach in controlling mobile robots for needs of robotic soccer. The designed method of fuzzy control was tested in the simulator Webots.

Keywords — mobile robot, fuzzy logic controller, robotic soccer

I. INTRODUCTION

This paper deals with the control of mobile robots using fuzzy logic. Mobile robots are included in the automation industry and they are creating new opportunities how to replace human in hazardous activities. Mobile robots are used at home to assist in various types of work or in the hospitals as a staff. Mobile robots seen in the demanding are as such as competition for example Robocup – robotic soccer.

This chapter is based on Saffionti [1] and Certo [2]. Despite the impressive advances in control of mobile robots in recent years were done, a number of problems remain. Most of the difficulties originate in the nature of real-world, unstructured environments, and in the large uncertainties that are inherent to these environments. The first, the prior knowledge about the environment is, in general, incomplete, uncertain and approximate. The second, perceptually acquired information is usually unreliable. The limited range, combined with effect of environmental features and of adverse observation conditions, leads to noisy and imprecise data. The third, real-world environments typically have complex and unpredictable dynamics. Finally, the effect of control actions is not completely reliable.

It often claimed that the qualitative nature of fuzzy logic makes it a useful tool for dealing with problems characterized by the pervasive presence of uncertainty. This paper follow are devoted to test the validity of this claim in control of mobile robots. Before we begin our algorithm design, however, we need better to clarify the intended meaning of key terms “fuzzy logic” and “fuzzy controller”.

Most concepts used in everyday language, such as “high temperature” is not clearly defined. In 1965 ZADEH proposed a mathematical definition of those classes that lack precisely defined criteria of membership. Zadeh called them fuzzy sets. Membership in a fuzzy set may be indicated by any number from 0 to 1, representing a range from “definitely not in the set” through “partially in the set” to “completely in the set”. Fuzzy sets are a generalization of ordinary sets, and they may be combined by operations similar to set union, intersection, and complement. However, some properties of ordinary set operations are no longer valid for fuzzy sets.

In technical applications, fuzzy control refers to programs or algorithms using fuzzy logic to allow machines to make decisions based on the practical knowledge of a human operator. The fundamental problem of automatic control is determining the appropriate response of the system, or production plant, for any given set of conditions. Fuzzy control does not require an exact theoretical model but only the empirical knowledge of an experienced operator. This knowledge is then expressed as a set of linguistic rules of the form “if [present conditions], then [action to be taken].” The control strategy can be encoded as an algorithm or computer program. During the operation of the machine, sensors measure the current values of the input

variables (angles with robots, in this case), and a computer then determines the appropriate values of the action variables (e.g., kicking ball robot).

For simulation of humanoid robot NAO and whole soccer field, simulation environment Webots - Robotstadium is used. The input variables creation of humanoid robot – NAO for fuzzy controller is illustrated in Fig. 1.

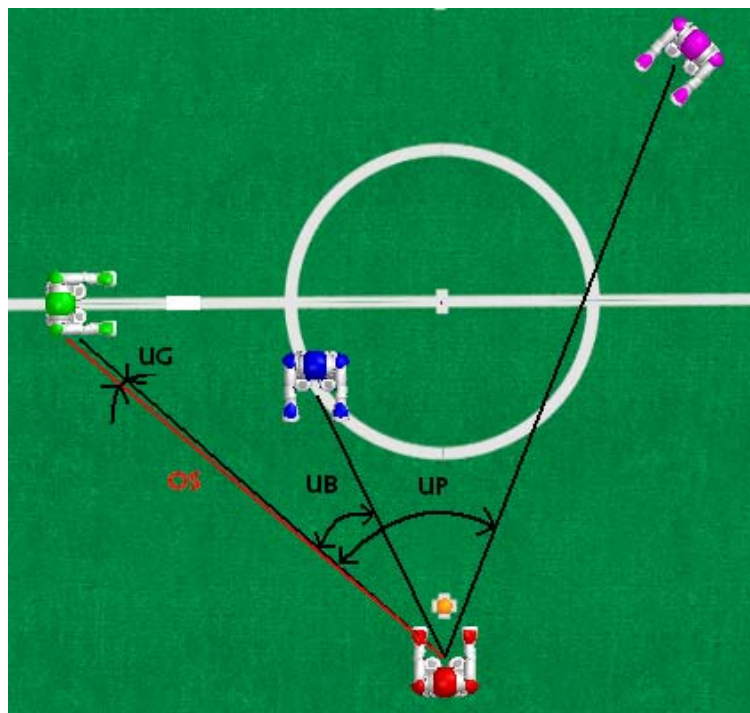


Fig. 1 Inputs variable for fuzzy controller. OS is axis of the robot. UG is angle to the green robot from the center of the robot. UB is angle to the blue robot from the center of the robot. UP is angle to the purple robot from the center of the robot.

II. PROPOSED SYSTEM

The overall structure of the system is shown on Fig. 2 and consists of individual parts which are interconnected and which are described respectively.

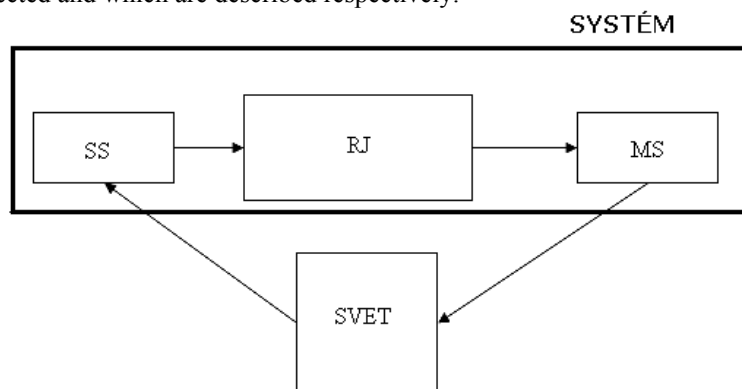


Fig. 2 The structure of the system

SVET – it is word under which we understand the environment in which the robot is located and which will solve your problem or your task. In this case the simulation environment Webots - Robotstadium is used.

SS – it is sensor-based system. This system is used to gain inputs into the robot. In this case it is camera.

MS – it is motor system. This system is responsible for the modification and movement in environment.

RJ – it is controller. As controller, we used Mamdani type fuzzy controller, whose structure and operation of individual blocks will be discussed in the next chapter III. Mamdani fuzzy controller is suitable for effective use of the complexity of controlling the mobile robot. The advantage of fuzzy controller is its robustness and ability to work with blunt values.

III. MAMDANI FUZZY CONTROLLER

This chapter is based on Vaščák [3], Pivoňka [4]. Mamdani fuzzy controller is used to control knowledge expert in the form of production rules. It is knowledge-based system working in real time. Schematic diagram of Mamdani fuzzy controller is realized in Fig. 3.

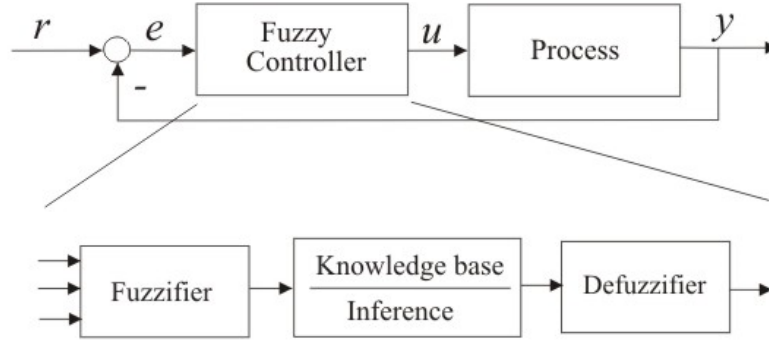


Fig. 3 Mamdani fuzzy controller

The following chapters describe the individual blocks designed of Mamdani fuzzy controller.

A. Fuzzifier

The fuzzifier determines the membership degrees of the controller input values in the antecedent fuzzy sets. Input values for fuzzy controller are angles with robots show on Fig. 1.

B. Knowledge base

The general form is given by the base rules:

IF antecedent **THEN** consequent **ELSE** the next rule

Antecedent is usually composed logical condition in which they are bound by rules of logical connections which is logical, then drained the second condition as consequent rule:

IF x_1 is A_{1j} **AND** x_2 is A_{2j} **AND...AND** x_N is A_{Nj} **THEN** u is B_j **ELSE** next rule
 where $j = 1, 2 \dots M$ is K -th rule antecedent composed of N partial terms. The input variables x_i are angles with robots. A_{ij} , B_j are fuzzy sets and u is the output variable.

Base rules for the proposed method in which the whole basis of the rules generated using the program. The count of rules in rules-based fuzzy controller is set depending on the function of the number of entries as follows:

$$\text{Count of rules} = x^y$$

where x is count of membership function and y is count input values of fuzzy controller.

C. Inference

The inference mechanism combines fuzzifier variables with the knowledge stored in the rules and determines what the output of the rule-based system should be. In general, this output is again a fuzzy set.

The most common method is the implementation of inference Min – Max.

Method Min-Max of fuzzy inference needs operation:

- Conjunction $\forall x \in X, \mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$
- Disjunction $\forall x \in X, \mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$

Method Min-Max of fuzzy inference is defining:

$$\max_j \{ \min_i [\mu_{A_{ij}}(x), \mu_{B_j}(x)] \}$$

D. Defuzzifier

The *defuzzifier* calculates the value of this crisp signal from the fuzzy controller outputs. There are many defuzzifier methods. We will use the defuzzifier method of middle maximum. This method is calculated from the first maximum and last maximum following:

1. calculated method first maximum

$$u_{\min}^* = \left\{ u_i \mid i \neq j \quad u_i < u_j \quad \forall i, j \quad \mu_g(u_i) = \mu_g(u_j) = \max(\mu_g(u)) \right\} \quad (1)$$

2. calculated method last maximum

$$u_{\max}^* = \left\{ u_i \mid i \neq j \quad u_i > u_j \quad \forall i, j \quad \mu_g(u_i) = \mu_g(u_j) = \max(\mu_g(u)) \right\} \quad (2)$$

3. calculated method middle maximum through the use of (1) and (2) :

$$u^* = u_{\min}^* + \frac{u_{\max}^* - u_{\min}^*}{2} \quad (3)$$

E. Membership function

Due to the simplicity of the computing process triangular membership functions were designed for input and output show as Fig. 4. The count of linguistic variables ranging $<3.9>$, because the maximum number of linguistic variables allows us to optimize and finer control and the minimum number is a necessity to define the object.

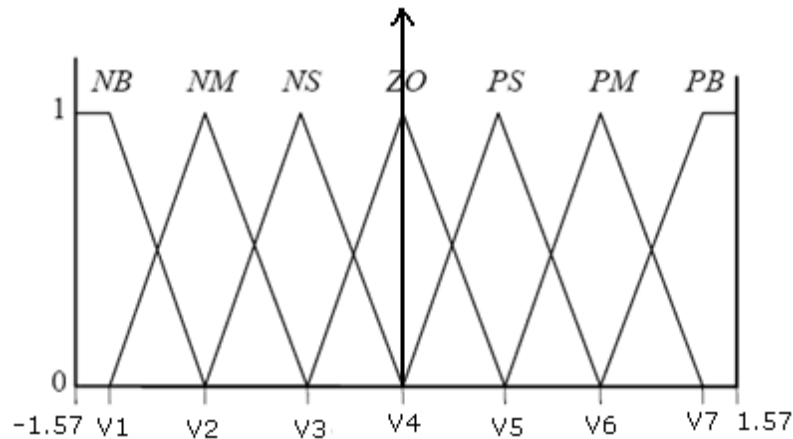


Fig. 4 The shape and distribution membership function of the universe for the standardized input and output

The size of the universe for input and output was standardized to the interval $<-1.57; 1.57>$, which is limiting deflection heads left and right.

IV. EXPERIMENTS

The experiments will find such a set of fuzzy controller, where the error angle passes the robot, which would be the lowest. Differences between experiments represent a change in the simulations:

- count of membership functions for inputs, which reduces the count of rule base
- combination of gaming, which represent the different layout of players of both teams in a simulation environment

A. Change in count of membership functions

The following experiments will adjust the count of the membership function, which reduces the count of rules in the rule base.

1) EXPERIMENT 1

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	3
Membership function output	max. 9
Rules	27
Games combination	15

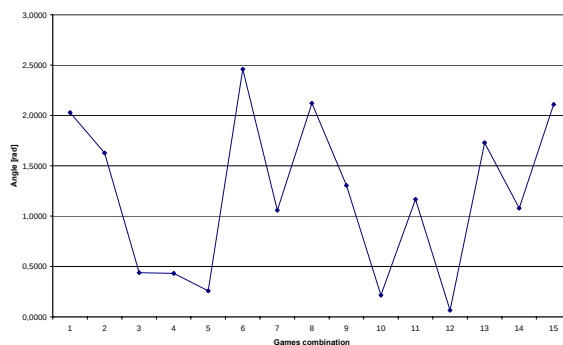


Fig. 5 Error angle passes in experiment 1

2) EXPERIMENT 2

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	5
Membership function output	max. 9
Rules	125
Games combination	15

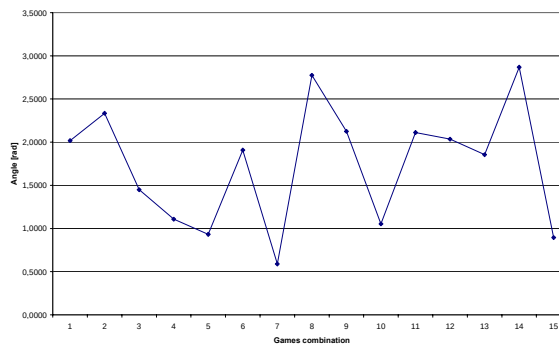


Fig. 6 Error angle passes in experiment 2

3) EXPERIMENT 3

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	7
Membership function output	max. 9
Rules	343
Games combination	15

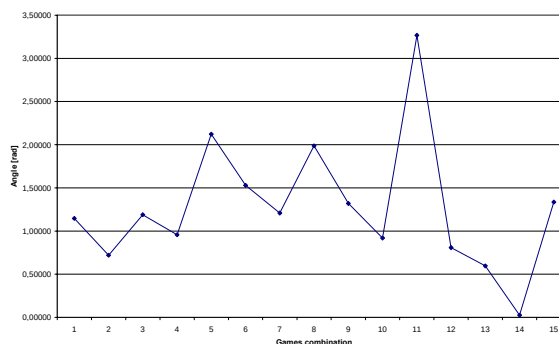


Fig. 7 Error angle passes in experiment 3

4) EXPERIMENT 4

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	9
Membership function output	max. 9
Rules	729
Games combination	15

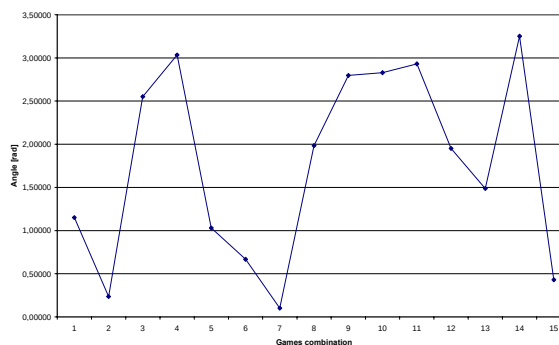


Fig. 8 Error angle passes in experiment 4

5) EXPERIMENT 5

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	9
Membership function output	9
Rules	6561
Games combination	15

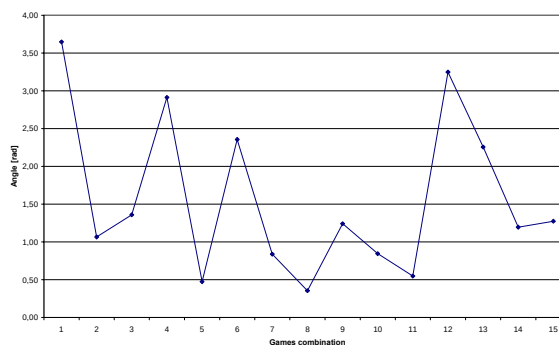


Fig. 9 Error angle passes in experiment 5

B. Change in combination of gaming

TABLE I
CHANGE IN THE SIMULATION

Count	Number
Membership function inputs	9
Membership function output	9
Rules	6561
Games combination	100

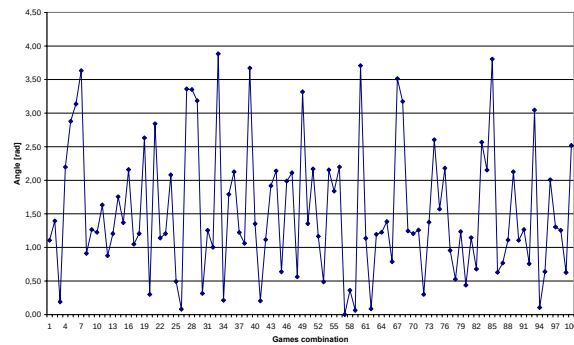


Fig. 10 Error angle passes in experiment 6

V. CONCLUSION

After performing the experiments, we have found out following facts:

- Experiment's change in count of membership functions shows that the count of the membership functions and hence the count of fuzzy rules, the regulator has no effect on the recording angle of the robot teammate. Depending on the count membership function of the fuzzy controller is independent with regard to finding a quality solution that will meet a number of game combinations.
- Experiment's change in combinations did not affect the game to find quality solutions. It is only changing of the combination of gaming regulation and visible results for different combination of these options.

ACKNOWLEDGMENT

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] SAFFIONTTI, Alessandro. The uses of fuzzy logic in autonomous robot navigation. In: Soft Computing. 1st ed. vol. 4 (1997), pp. 180-197. On the Internet:
 <<http://www.fmi.uni-sofia.bg/courses/biomed/robots/ITSolutions/hfc98.pdf>>
- [2] CERTO, Giuseppe: Nao model and simulation for webots. In: Bachelor semester project, EPFL – School of Computer & Communication Sciences, 2009. On the Internet:
 <<http://birg.epfl.ch/webdav/site/birg/users/170218/private/report.pdf>>
- [3] VAŠČÁK, Ján: Fuzzy logika v regulácii, Košice: TU-FEI-KKUI, 2008. On the Internet:
 <<http://www.ai-cit.sk/source/fuzzy/FLvR.pdf>>
- [4] PIVOŇKA, Petr: Vyšší formy řízení, Brno: VUTBR FSI, 2003 On the Internet:
 <<http://www.kves.uniza.sk/kvesnew/dokumenty/%C5%A0EP/PID%20Control/VFR03.pdf>>

Nastavovanie parametrov fuzzy regulátora pre navigáciu robotov pomocou evolučných prístupov

Ján VAŠČÁK

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jan.vascak@tuke.sk

Abstrakt — Tento príspevok sa zaoberá využitím evolučného prístupu pre návrh bázy znalostí fuzzy regulátora pre potreby navigácie mobilného robota vo futbalovej hre „bago“. Hra sa využíva pre tvorbu stratégie prihrávok a chytania hráčskych robotov. Vzhľadom na to bol navrhnutý systém, ktorého základom je fuzzy regulátor, pričom požadované parametre tohto regulátora sa nastavovali genetickým algoritmom. Pomocou genetického algoritmu sme upravovali bázu pravidiel regulátora. Systém bol navrhnutý v simulačnom prostredí, kde bol aj aplikovaný na humanoidného robota NAO pri hre bago.

Kľúčové slová — genetický algoritmus, fuzzy regulátor, mobilný robot, robotický futbal

I. ÚVOD

Evolúcia ľudstva je tisíce až milióny rokov vývoja, kedy sa z najjednoduchších jednobunkových organizmov stávali zložité mnohobunkové živočíchy. Náš telesný vývoj už dosiahol hranice, ale naša inteligencia je ešte len na začiatku. V mojom programe vyvíjam mobilných robotov, ktorý sa snažia chápať, myslieť a vnímať to čo sa okolo nich deje. Reagovanie robotov síce nie je veľmi rýchle, ale tak ako sa vyvíjalo ľudstvo, vyvíja sa aj technika konštrukcie mobilných robotov. Pomocou evolučného prístupu som sa snažil naprogramovať robotov tak, aby sa aspoň trochu podobali skutočným futbalistom. Vybral som také metódy, spôsoby a postupy, aby robot vedel rozpoznávať svoju pozíciu, farbu a uhol svojho spoluhráča, uhol svojho protivníka a vzdialenosť a uhol lopty. Program by mal fungovať tak, aby robot vedel rozmýšľať, kde sa nachádza lopta a komu má loptu kopnúť a komu zasa nie. Evolúcia, čiže vývoj robotov ešte len začína a momentálne je na dobrej ceste za zdokonaľovaním.

II. NAVIGÁCIA MOBILNÉHO ROBOTA

Navigácia mobilných robotov v prostredí vyžaduje schopnosť robota dané prostredie určitým spôsobom vnímať. Na to slúži senzorový systém robota. Ten poskytuje aktuálne informácie o okolí prostredí, ktoré spracúva riadiaci systém robota. Výstupom riadiaceho systému je informácia (napr. vzdialenosť prekážky, uhol od prekážky atď.), ktorá slúži na vykonanie potrebného kroku pri ceste z štartovacieho bodu do bodu cieľového. Veľká časť výskumných inžinierov sa špecializuje na vytvorenie takého algoritmu, pomocou ktorého by sa robot vedel orientovať v akomkoľvek neznámom prostredí. Podľa [332] navigácia je:

Je proces určovania a zachovania kurzu či trajektórie z jedného miesta na druhé. Procesy, ktoré odhaľujú pozíciu jednotlivca vzhľadom na poznaný svet, sú v navigácii kľúčové. Poznaný svet sa skladá z plôch, ktorých jednotlivé umiestnenia sú zobrazené na mape.

Navigácia sa z pohľadu snímania prostredia rozdeľuje do dvoch skupín na Globálnu a Lokálnu.

- *Globálna navigácia* je to navigácia robota pri ktorej je potrebné vedieť globálnu informáciu o prostredí. Taká navigácia sa vykonáva na základe poznania mapy priestoru, ktorú nám poskytujú kognitívne mapy. Kognitívna mapa sa vytvára napríklad z umiestnenia kamery nad prostredím.

- *Lokálna navigácia* je taká navigácia, pri ktorej sa určuje nasledovný krok len zo senzorických vlastností robota. Robot nemá informáciu o každom bode prostredia, len o určitej časti, v ktorej sa daný robot nachádza a pokiaľ mu to umožňujú senzorické vlastnosti.

III. INTELIGENTNÉ METÓDY NAVIGÁCIE MOBILNÉHO ROBOTA

Aplikácia metód a algoritmov umelej inteligencie pri riešení zložitých problémov rozhodovania a riadenia tak v technických, ako aj v netechnických vedách zaznamenala nebývalý rozvoj hlavne vďaka vzniku metód výpočtovej inteligencie (fuzzy množín, umelých neurónových sietí a evolučných algoritmov). Inteligentné systémy vyznačujúce sa aplikovanými prvkami umelej inteligencie sa sú charakterizované nasledujúcimi vlastnosťami:

- Majú schopnosť sa učiť a získavať znalosti
- Majú schopnosť dané znalosti uchovávať
- Vedia nadobudnuté znalosti použiť na riešenie problému

Podľa existujú úrovně riadenia inteligentných systémov, ktoré rozdeľujeme:

- Nultá úroveň - Robustný regulátor
- Prvá úroveň – Adaptívny regulátor
- Druhá úroveň – Optimálny regulátor
- Tretia úroveň - Plánovací regulátor

Robustný regulátor: Je to typ regulátora, ktorý je robustný voči zmenám respektíve zachováva sa jeho kvalitatívne vlastnosti (stabilita, kvalita regulácie). Slovom *Robustnosť* môžeme definovať systém, ktorý sa správa voči zmenám a poruchám algoritmu neohľaduplne respektíve jeho ďalšie pôsobenie sa zaobíde len s malou ujmovou na spôsobe regulácie.

Adaptívny regulátor: Používa sa v prípadoch kde sa v priebehu času menia dynamické alebo statické vlastnosti regulovanej sústavy (napr. zmena koeficientu prestupu tepla usadeninami, zmena prietoku zanášaním atď.), ktoré môžu mať za následok zhoršenie kvality regulácie, nakoniec i jej zlyhanie. Adaptívny regulátor je schopný tieto zmeny identifikovať a prestaviť svoje konštanty, tak aby bola zaistená pôvodná kvalita regulácie.

Optimálny regulátor: Je regulátor, ktorý má z určitého hľadiska najlepšie možné vlastnosti. Môžeme napr. požadovať odstránenie regulačnej odchýlky v minimálnom čase, alebo s najmenším prerégulovaním (dynamická optimalizácia), alebo s vynaložením minimálnej energie alebo s maximálnou účinnosťou.

Plánovací regulátor: Je regulátor, ktorého obsah pozostáva z plánovacích funkcií, ktorých náplňou je:

- Plánovanie rizika
- Plánovanie porúch
- Plánovanie chýb
- a pod....

IV. GENETICKÝ ALGORITMUS

Prírodná evolúcia je úspešná a robustná adaptácia v biologických systémov.

Genetický algoritmus sa objavil začiatkom 60. rokov minulého storočia a nachádzajú sa na pomedzí deterministických a nedeterministických algoritmov. Princíp práce genetického algoritmu je postupná tvorba generácií rôznych riešení daného problému. Pri riešení sa uchováva tzv. populácia ktorej každý jedinec predstavuje jedno riešenie daného problému. Ako populácia prebieha evolúciou riešenie sa zlepšuje.

Jedince sa označujú názvom fenotyp a pre jeho reprezentáciu sa používa názov genotyp, alebo chromozóm. Chromozóm sa delí na jednotlivé gény, ktoré sú lineárne usporiadané. Gén nadobúda rôzne hodnoty. Jedinci môžu byť zakódované rôznym spôsobom. Spôsob ako je jedinec zakódovaný má veľký vplyv na úspech alebo neúspech riešenia. Najjednoduchším spôsobom ako zakódovať jedinca je binárny jedinec nejakej dĺžky. (2)

$$(1,0,1,0,1,1,0,0)$$

$$(1)$$

1. Inicializácia – Vytvorenie prvej populácie
2. Začiatok – Na základe určitej metódy vyberieme z populácie niekoľko jedincov s vysokou zdatnosťou
3. Pomocou genetických operátorov vytvoríme novú populáciu
 - Križenie – Výmena časti jedincov medzi sebou
 - Mutácia – Náhodná zmena časti jedinca
 - Reprodukcia – Kopírovanie jedinca bez zmeny
4. Vyhodnotenie jedincov novej populácie
5. Ukončenie algoritmu na základe ukončovacej podmienky
6. Koniec algoritmu – Výstupom je jedinec s najlepšimi vlastnosťami

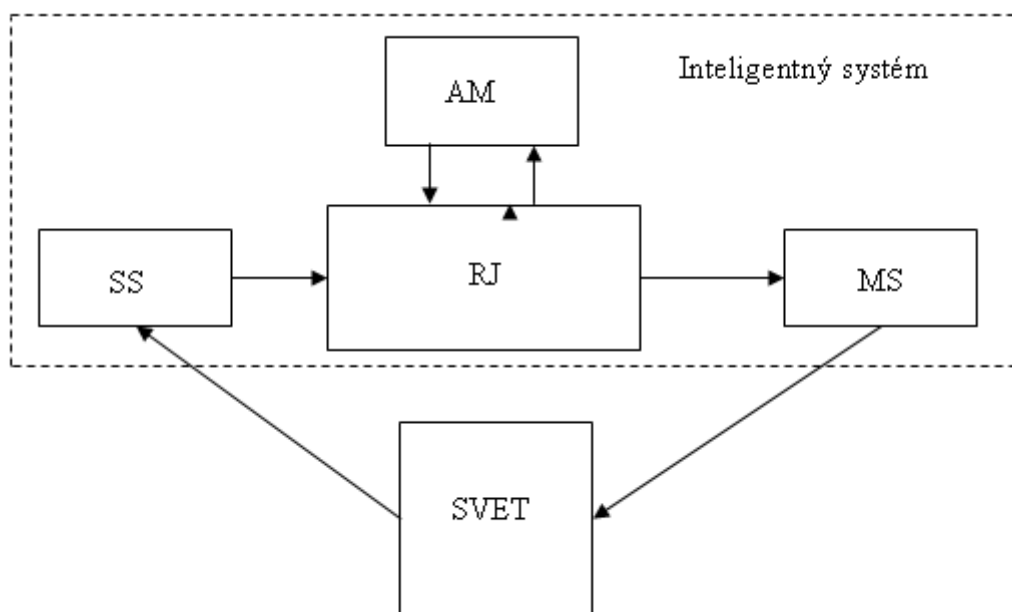
Genetické algoritmy sú väčšinou používané v praxi na optimalizáciu úloh, ale miernou úpravou toho algoritmu ho môžeme použiť na rozmanité úlohy.

V. NÁVRH SYSTÉMU

Štruktúra systému pozostáva zo štyroch častí, ktoré sú navzájom prepojené:

1. Senzorický systém (SS)
2. Riadiaca jednotka (RJ)
3. Motorický systém (MS)
4. Adaptačný mechanizmus (AM)

Štruktúra systému je znázornená na obrázku (Obr. 1)



Obr. 1 Štruktúra inteligentného systému

A. Senzorický systém

Na zber dát z prostredia bola použitá kamera, ktorá plnila účel, ako sú u ľudí oči a na základe klasických goniometrických funkcií sa dané vzdialenosti a uhly objektov dokázali prepočítavať

B. Riadiaca jednotka

Riadiaca jednotka tvorila jadro navigácie robota a pozostávala s fuzzy regulátora typu TSK (Takagi – Sugeno - Kang) je to typ regulátora, ktorý nemá schopnosť adaptácie a preto som tam vytvoril hybridný prístup, kde na učenie fuzzy regulátora bol použitý genetický algoritmus.

Fuzzy regulátor pozostáva s piatich častí: (Obr. 2)

1. Fuzzyfikácia – Mení ostré hodnoty na stupne príslušnosti jednotlivých lingvistických premenných.
2. Báza pravidiel – Obsahuje jednotlivé pravidla typu IF – THEN

3. Bába dāt – Definuje jednotlivé funkcie príslušnosti
4. Vyhodnotenie pravidiel - je určenie výsledného stupňa príslušnosti celej predpokladovej časti pre každé pravidlo.
5. Defuzzyfikácia - je prevod výslednej funkcie príslušnosti výstupov na reálnu hodnotu (ostrú).

Fuzzy regulátor typu TSK je o niečo jednoduchší. Pri tomto type regulátora odpadá potreba defuzzyfikácie, čo ma za následok to, že obsahuje pravidla v tvare: (2)

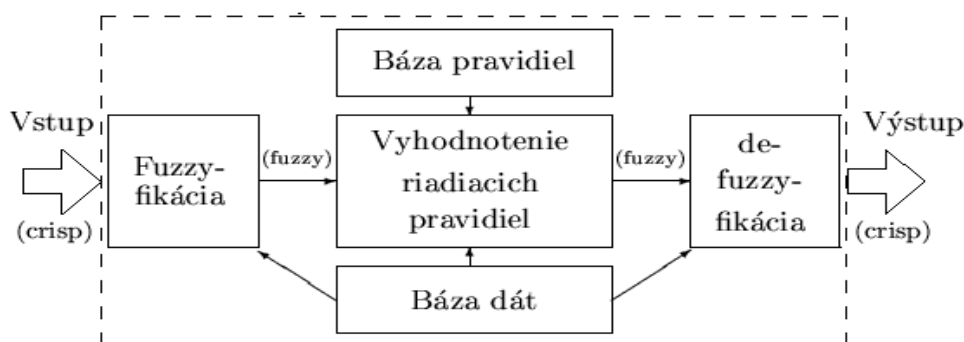
$$\text{Ak } x \text{ je } A \text{ \& } y \text{ je } B \text{ potom } z = f(x,y) \quad (2)$$

kde:

x, y – sú vstupy privádzané do fuzzy regulátora.

A, B – sú funkcie príslušnosti priradené danému vstupu.

z – je funkcia vstupu x a y , už ostrá hodnota jedného pravidla.



Obr. 2 Štruktúra Mamdaniho fuzzy regulátora

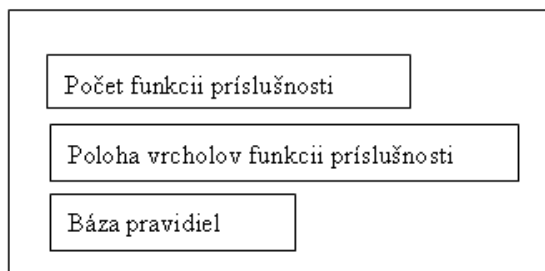
C. Adaptačný mechanizmus

Samotný TSK fuzzy regulátor je neadaptívny typ regulátora a tak máme snahu tento fakt zmeniť a vytvárame hybridný prístup, ktorý rozširuje daný TSK regulátor o schopnosť vyvíjať sa. Ako adaptačný mechanizmus sme použili Genetický algoritmus, ktorý má na základe výskumu najlepšie vlastnosti v oblasti riadenia. Genetický algoritmus patrí do skupiny evolučných algoritmov. Samotná adaptácia fuzzy regulátora sa týka úlohy zmeny bázy pravidiel, kde sa populácia skladá z jedincov, kedy jeden jedinec predstavuje jedno riešenie celého problému.

D. Štruktúra jedinca

Samotná štruktúra jedinca pozostáva z parametrov fuzzy regulátora: (Obr. 3)

1. Počet funkcií príslušnosti
2. Poloha vrcholov funkcií príslušnosti
3. Samotná báza pravidiel



Obr. 3 Štruktúra jedinca

Prvá časť chromozómu pozostáva z trojprvkového chromozómu odpovedajúcemu počtu vstupov do regulátora. Pre každý vstup je vygenerovaný počet funkcií príslušnosti z intervalu $\langle 3, 7 \rangle$. Táto časť chromozómu sa počas adaptácie po vygenerovaní nemení.

Druhá časť chromozómu obsahuje hodnoty vrcholov funkcií príslušnosti. Keďže na základe vrcholov dokážeme určiť polohu jedného a polohu druhého nosiča funkcie príslušnosti

potrebného na výpočet stupňa príslušnosti k danému vstupu. Dĺžka tohto chromozómu je rôzna a závisí od počtu nadefinovaných funkcií príslušnosti. Pri tejto časti chromozómu je dôležité zachovať fuzzy partíciu. Táto čas chromozómu sa počas adaptácie po vygenerovaní nemení.

Posledná časť chromozómu je venovaná tvorbe pravidiel, jeho počet je počas adaptácie premenlivý, kde jedno pravidlo predstavuje päť prvkové pole (Obr. 4). Počet vygenerovaných pravidiel je zo spodnej hranice ohraničený na zvládnutie danej úlohy a hornú hranicu ma stanovenú na základe počtu funkcií príslušnosti vygenerovaných pre dané vstupy do regulátora. Maximálny počet pravidiel pre jedného jedinca je vypočítavaný na základe $N_{I1}xN_{I2}xN_{I3}$.

Kde N_{I1} - predstavuje počet funkcií príslušnosti pre prvý vstup.

1	3	7	0.15	0.5
---	---	---	------	-----

Obr. 4 Štruktúra pravidla

- Prvé tri pozície pravidla prislúchajú vstupom a hovoria ktorá funkcia príslušnosti je priradená danému vstupu.
- Štvrtá pozícia je priradená hodnota výstupu z intervalu $<-1.57;1.57>$.
- Piata pozícia v pravidle prislúcha váhe pravidla nadobúda hodnoty z intervalu $<0;1>$.

VI. GENETICKÝ ALGORITMUS PRE NÁVRH BÁZY ZNALOSTÍ

Populáciu tvorí určitý počet jedincov, kde jedinec predstavuje jeden návrh fuzzy regulátora. Štruktúra jedinca pozostáva z niekoľko typov chromozómov. Úlohou genetického algoritmu je navrhnúť takú bazu znalosti, ktorá by mala najlepšie vlastnosti na riešenie daného problému.

A. Vytvorenie populácie

Vytvorenie populácie je proces pri ktorom sa jedinci nagenerujú náhodne a to podľa vopred definovaných podmienok.

B. Ohodnotenie jedincov

Každý jedinec po vytvorení nesie so sebou svoje vlastnosti či už dobre alebo zle. Aby sme vedeli či vlastnosti daného jedinca sú vhodné pri vytváraní správania potrebného na zvládnutie úlohy je potrebné ich nejakým spôsobom merať. Pri meraní schopnosti jedinca použijeme fitness funkciu, ktorá bola definovaná na základe presnosti nahrávky daného hráča svojmu spoluhráčovi. Grafické znázornenie prepočtu funkcie príslušnosti je na obrázku (Obr. 5).

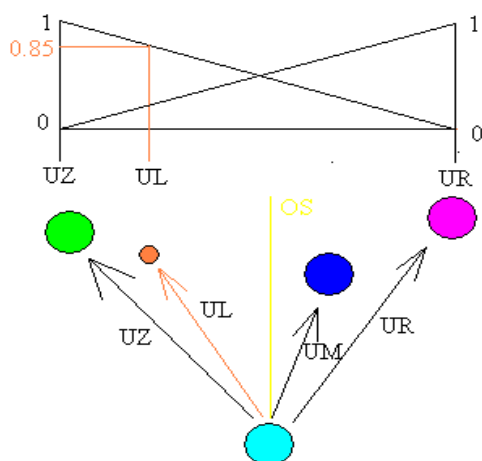
Na určovanie hodnoty funkcie príslušnosti som použil jednoduchý trojuholník, kde som prepočítaval hodnoty na základe analytickej geometrie podľa vzorca (3)

$$fitness = \frac{U_{holH2} - U_{holH1}}{U_{holH2} - U_{holLopty}} \quad (3)$$

Takýto spôsob sa používal na ohodnotenie jednej hracej situácie. Pri zadefinovaní väčšieho počtu hracích kombinácií sa fitness prepočítala ako aritmetický priemer cez všetky hracie kombinácie čo nám umožnilo, že daný jedinec nezabudne naučenú kombináciu pri učení inej, ale sa kontroluje správanie jedinca pri viacerých možných hracích kombinácií a vypočítaná hodnota fitness popisovala vlastnosti daného jedinca.

C. Selekcia

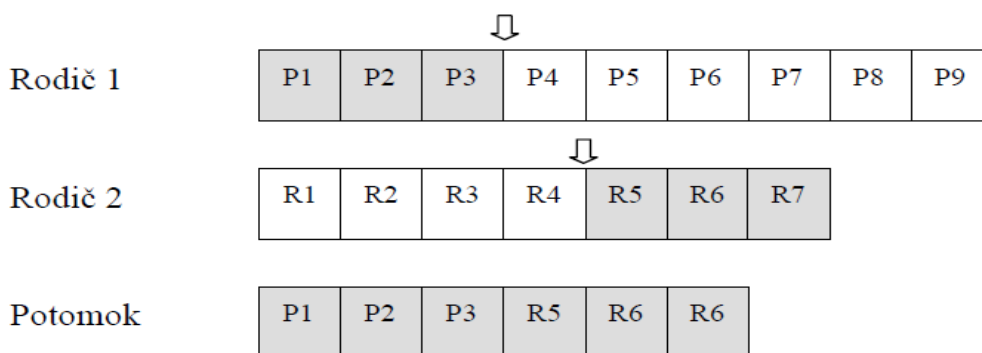
Tento blok genetického algoritmu je veľmi dôležitý pre tvorbu jedincov do novej populácie. Podstata selekcie je výber dvoch rodičov, ktorý sa budú podieľať na reprodukčnom procese, kde z dvoch vybraných rodičov sa vytvorí jeden potomok. Pri selekcii rodičov sa museli zadefinovať ohraničenie kvôli možnému kríženiu rodičov. Ohraničenie spočívalo v tom, že do reprodukčného procesu sa mohli dostať len takí rodičia, ktorý majú rovnakú časť štruktúry chromozómu, v podstate sa jedná o to, že musia mať rovnakú prvú časť chromozómu, čiže rovnaký počet nadefinovaných funkcií príslušnosti pre dané vstupy. Rodičia boli vybraný náhodným spôsobom či už obsahovali dobré alebo zle vlastnosti, čo nám prinieslo väčšie množstvo genetického materiálu s ktorým mohol daný genetický algoritmus pracovať.



Obr. 5 Výpočet hodnoty fitness funkcie

D. Križenie

Pri tomto procese som použil metódu jednobodového križenia. Podstata spočíva v tom, že rodičia môžu mať rôzny počet nadefinovaných pravidiel a pre každého z rodičov sa náhodne vygeneroval rôzny bod križenia. (Obr. 6) Po úspešnom vytvorení potomka sa musela uplatniť jedna modifikácia jedinca, ktorá odstrániť potenciálne rovnaké pravidla. Takýto jedinec bol vložený do populácie potomkov. Celý proces sa opakoval pokiaľ nebola vytvorených toľko potomkov koľko bolo rodičov.



Obr. 6 Jednobodové križenie rodičov

E. Vytvorenie novej populácie

Nová generácia bola vytvorená výberom medzi populáciou rodičov a populáciou potomkov, na základe dosiahnutých vlastností pri riešení daného problému. Dosiahnuté vlastnosti sa ohodnocovali fitness funkciou. Pri vytvorení novej populácie bol použitý mierny elitizmus, kde jedinci s najlepšimi vlastnosťami s predošlej generácie boli hneď presunutý do novej generácie, pričom týmto prechodom nestratili možnosť reprodukcie.

VII. EXPERIMENT

Pri vývoji môjho genetického algoritmu som sa rozhodol nastavovať parametre, ktoré priamo súvisia s algoritmom a ktoré súvisia len okrajovo respektíve súvisia, ale s modifikáciou genetického algoritmu. Nastavované parametre sú:

1. Počet jedincov
2. Počet generácií
3. Počet herných kombinácií
4. Medzná hodnota podobných pravidiel
5. Počet elityckých jedincov

Zmenou týchto pravidiel by som chcel zlepšiť evolučný proces a tým dosiahnuť uspokojivé výsledky pri navigácii mobilného robota v futbalovej hre „bago“.

Počet jedincov má veľký vplyv na riešenie úlohu pričom na istý počet herných kombinácií je prislúcha úmerný počet jedincov. Väčšie množstvo herných kombinácií si vyžaduje veľké množstvo znalostí, ktoré sú uložené v genetickom materiáli jedincov. Zväčšenie genetického materiálu dosiahnem zväčšením počtu jedincov čiže veľkosti populácie. Toto správanie si môžeme všimnúť na experimentoch 1 - 4. Podľa toho vyplýva že na 5 herných kombinácií potrebujeme aspoň 15 krát väčšie množstvo jedincov.

Pri zmene prahovej hodnoty podobných pravidiel nastáva problematická otázka, kedy sú dane pravidla rovnaké a kedy umelo ovplyvňuje evolučný proces tým že zatracujeme genetický materiál, ktorý by mohol byť pre nás dosť zaujímavý. Tento návrh som robil metódou pokus omyl. Tato metóda nebola až tak veľmi nepriateľna lebo ak sme nastavili prahovú hodnotu vyššiu tak genetický algoritmus nám neprinášal tak dobré výsledky a naopak ak hodnota prahu bola príliš nízka začali sa nám hromadiť pravidla čo malo za následok, že pomaly sa nám genetický materiál začal dosť na seba podobáť a jedinec ostal v lokálnom minime. Pri ideálnom nastavení sa prahovej hodnoty sa výrazne zmenila najlepšia vhodnosť jedinca.

VIII. ZÁVER

Použitie fuzzy regulátora s adaptáciou na riadenie smeru nahrávky pri futbalovej hre (bago) sa javí ako veľmi dobrá voľba. Fuzzy regulátor sa mal adaptovať na 30 herných kombinácií čo si vyžaduje veľký počet uchovávaných znalostí. Veľkou výhodou použitia fuzzy regulátora s možnosťou adaptácie je, že ak na vstup príde podobná situácia ako je situácia na ktorú bol naučený dokáže veľmi pekne reagovať a výsledky sú pomerne dobré. Pri vyvinutí takých základných herných kombinácií, ktoré by pokrývali celý herný priestor, dokázal by robot s použitím fuzzy regulátora s možnosťou adaptácie reagovať na akúkoľvek situáciu a priniesť kvalitný výsledok, či už na ňu naučený je alebo nie.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] JURIŠICA, L.: LAT&T Journal 5/2003 – Robotika.
- [2] NASCIMENTO, J.: Artificial NN in Control and Optimization. Doctor thesis, Manchester UK: Faculty of Technology, University of Manchester 1994.
- [3] SINČAK P. a ANDREJKOVÁ G.: Neurónové siete Inžiniersky prístup 1. Elfa: Kosice, 1996.
Vaščák J.: Fuzzy logika v regulácii. Podrobný úvod do problematiky fuzzy riadenia. 2008. Dostupné na internete: <<http://www.ai-cit.sk/source/fuzzy/FLvR.pdf>>
- [4] I. Sekaj: Evolučné výpočty, IRIS Bratislava, 2005.

Využitie neurónových sietí pri extrakcii pravidiel pre fuzzy regulátor

Ján VAŠČÁK

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jan.vascak@tuke.sk

Abstrakt — Jedným z hlavných nedostatkov fuzzy regulátorov je ich neschopnosť sa učiť. V tomto príspevku využívame na získavanie bázy znalostí fuzzy regulátora neurónové siete. Konkrétne sa zaoberáme sieťou typu FALCON, ktorá svojou štruktúrou umožňuje extrahovať znalosti bez ďalšej nutnosti ich pretransformovania do vhodnej podoby. V príspevku sú načrtnuté základy prístupy učenia tohto typu siete, pričom sa v ďalšom zaoberáme tzv. hybridným učením. Výkonnosť učenia a kvalita takto získaných znalostí bola experimentálne overená na príklade dodržiavania bezpečnej vzdialenosti vozidiel.

Kľúčové slová — neurónová sieť FALCON, fuzzy regulátor, hybridné učenie

I. ÚVOD

Úspešnosť riadenia závisí na popise sústavy, ktorú chceme riadiť. Ak je sústava silno nelineárna, matematicky ťažko popísateľná, veľmi komplikovaná alebo citlivá na prudké zmeny akčného zásahu je vhodné použiť fuzzy regulátor. Fuzzy regulátor obsahuje bázu znalostí, ktorá je tvorená rozhodovacími pravidlami typu AK-POTOM. Pomocou týchto pravidiel je možné popísať sústavu, ktorú chceme riadiť. Výsledok fuzzy riadenia závisí od skúseností človeka(expert), ktorý tieto pravidlá tvorí. V mnohých prípadoch je nedostatok týchto expertov alebo je ťažké popísať systém rozhodovacími pravidlami. Z tohto dôvodu sa začalo pracovať na systémoch, ktoré bez prítomnosti človeka dokážu vytvoriť bázu znalostí a pravidiel. K týmto systémom patrí aj neurónová sieť FALCON.

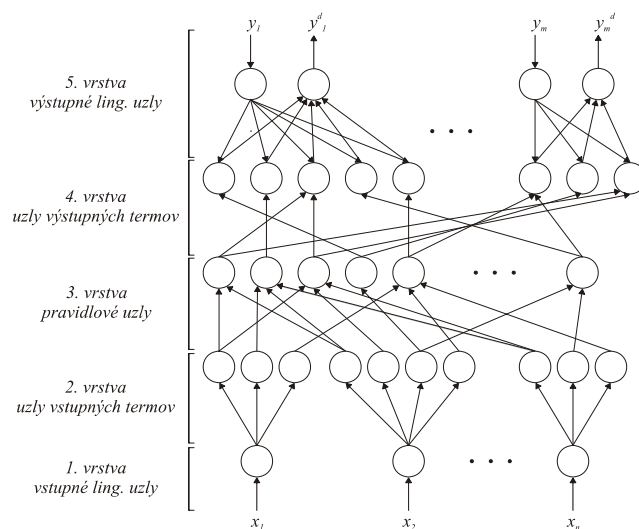
II. OPIS SIETE FALCON

FALCON je dopredná, viacvrstvová sieť. Jej názov je odvodený zo začiatočných písmen anglických slov fuzzy adaptive learning control network. Je zaujímavá tým, že v jej štruktúre sú zahrnuté vlastnosti fuzzy regulátora, ktoré majú schopnosť učiť sa.

A. Štruktúra siete

FALCON je päťvrstvová sieť pričom má tri skryté vrstvy. Uzly v prvej vrstve sú vstupné uzly, ktorými sú reprezentované lingvistické premenné a preto sa tieto uzly nazývajú vstupné lingvistické uzly. Podobne aj uzly v piatej vrstve, ktoré sa nazývajú výstupné lingvistické uzly. Ak je jeden výstup zo siete, tak výstupné lingvistické uzly sú dva. Jeden uzol do siete vstupuje a druhý z nej vystupuje. Uzol, ktorý do siete vstupuje obsahuje požadovaný výstup(trénovacie dáta) a uzol, ktorý z nej vystupuje obsahuje aktuálny výstup(výstup zo siete). Ak je výstupov zo siete viac, tak počet výstupných lingvistických uzlov je stále dvojnásobný. Uzly v druhej a štvrtej vrstve sa nazývajú uzly termov. V druhej vrstve sú to uzly vstupných termov a vo štvrtej vrstve uzly výstupných termov. Sú v nich uložené funkcie príslušností lingvistických premenných. Uzly v tretej vrstve sa nazývajú pravidlové uzly. Reprezentujú sa v nich fuzzy logické pravidlá. Spojenia v tretej vrstve definujú predpoklady pravidlových uzlov a spojenia vo štvrtej vrstve definujú závislosť týchto pravidlových uzlov. Z tohto dôvodu sa v každom pravidlovom uzle nachádza maximálne jedno spojenie s uzlom zo štvrtej vrstvy. Môže nastať prípad, keď sa medzi týmito uzlami nenachádza ani jedno spojenie. To znamená, že daný pravidlový uzol má veľmi malý respektíve nemá žiadny vplyv na danú výstupnú lingvistickú

premennú. Spojenia v druhej a v piatej vrstve sú plne prepojené. Šípky na spojeniach znázorňujú smer toku dát v sieti.



Obr. 1 Štruktúra neurónovej siete FALCON

B. Funkcie jednotlivých vrstiev

V tejto časti podľa [1] popíšem jednotlivé vrstvy siete a vysvetlím čo sa na každej vrstve počíta. Každá vrstva má svoju vstupnú funkciu f a aktivačnú funkciu a . Hodnoty váh sú označené písmenom w .

Vstupné lingvistické uzly

Neuróny v tejto vrstve iba posielajú vstupné hodnoty do ďalšej vrstvy. Hodnoty váh sú konštantné. Dolný index i určuje lingvistickú premennú, v ktorej sa daný vstup nachádza a horný index (1) označuje číslo vrstvy siete.

$$\begin{aligned} f &= u_i^{(1)}, \\ a &= f, \\ w_i^{(1)} &= 1. \end{aligned} \quad (1)$$

Uzly vstupných termov

V týchto neurónoch sa vytvárajú funkcie príslušností. Každý neurón transformuje vstupné hodnoty na hodnoty konkrétnej funkcie príslušnosti. Tieto hodnoty sa tiež nazývajú stupeň príslušnosti. Sú v rozmedzí od 0 do 1. Hodnota približujúca sa k nule vyjadruje malú príslušnosť k danej funkcii príslušnosti. Hodnota približujúca sa k jednotke hovorí o veľkej miere príslušnosti k danej funkcii príslušnosti.

Premenná m_{ij} je stred Gaussovej krivky j -tej funkcie príslušnosti a i -tej lingvistickej premennej a σ_{ij} je variancia Gaussovej krivky.

$$\begin{aligned} f &= -\frac{(u_i^{(2)} - m_{ij})^2}{\sigma_{ij}^2}, \\ a &= \exp(f), \\ w_{ij}^{(2)} &= m_{ij}. \end{aligned} \quad (2)$$

Pravidlové uzly

Spojenia v tejto vrstve definujú predpoklady pravidlových uzlov. Preto tieto pravidlové uzly vykonávajú fuzzy logický súčin. Tento fuzzy logický súčin je realizovaný pomocou operácie minima.

$$\begin{aligned} f &= \min(u_1^{(3)}, u_2^{(3)}, \dots, u_n^{(3)}), \\ a &= f, \\ w_i^{(3)} &= 1. \end{aligned} \quad (3)$$

Uzly výstupných termov

Neuróny v tejto vrstve pracujú v dvoch režimoch. V prvom režime zdola-nahor spojenia tejto vrstvy akumulujú pravidlá s rovnakým záverom. Môže nastať prípad, keď naakumulovaná hodnota prekročí hodnotu 1. Vtedy sa táto hodnota oreže na 1.

$$\begin{aligned} f &= \sum_i u_i^{(4)}, \\ a &= \min(1, f), \\ w_i^{(4)} &= 1. \end{aligned} \quad (4)$$

V druhom režime zhora-nadol vykonávajú neuróny rovnakú funkciu ako neuróny v druhej vrstve (2).

Výstupné lingvistické uzly

Aj v tejto vrstve sa pracuje v dvoch režimoch. V režime zhora-nadol vstupujú tréningové dáta do siete. Neuróny majú v podstate rovnakú funkciu ako neuróny v prvej vrstve.

$$\begin{aligned} f &= y_i, \\ a &= f, \\ w_i^{(6)} &= 1. \end{aligned} \quad (5)$$

V režime zdola-nahor neuróny pracujú ako defuzzifikátor, čiže prevádzajú fuzzy hodnoty na ostré čísla. Nasledujúce vzťahy simulujú ťažiskovú metódu, ktorá je veľmi často používanou metódou defuzzifikácie.

$$\begin{aligned} f &= \sum_j (m_{ij} \sigma_{ij}) u_{ij}^{(5)}, \\ a &= \frac{f}{\sum_j \sigma_{ij} u_{ij}^{(5)}}, \\ w_{ij}^{(6)} &= m_{ij} \sigma_{ij}. \end{aligned} \quad (6)$$

III. HYBRIDNÉ UČENIE

A. Štruktúrne učenie

Štruktúrne učenie je prvou fázou hybridného učenia a je použité na nájdenie počiatočných parametrov funkcií príslušností a na vytvorenie fuzzy logických pravidiel. V tejto fáze sieť pracuje v režime zhora-nadol, to znamená, že tréningové dáta vstupujú do siete z oboch smerov.

Na približné určenie stredov funkcií príslušností je podľa [1] použité kohonenovo pravidlo

$$\begin{aligned} \|x(t) - m_{\text{najbližší}}(t)\| &= \min\{\|x(t) - m_i(t)\|\}, \\ m_{\text{najbližší}}(t+1) &= m_{\text{najbližší}}(t) + \alpha(t)[x(t) - m_{\text{najbližší}}(t)]. \end{aligned} \quad (7)$$

Premenná $x(t)$ je vstup alebo výstup lingvistickej premennej, $\alpha(t)$ je parameter učenia, $m_{\text{najbližší}}(t)$ je počiatočná hodnota stredu funkcie príslušnosti.

Keď poznáme všetky stredu funkcií príslušností, je potrebné určiť variancie funkcií príslušností. Na výpočet variancií je podľa [1] použité pravidlo prvého najbližšieho suseda

$$\sigma_i = \frac{|m_i - m_{\text{najbližší}}|}{r}. \quad (8)$$

Premenná σ_i je variancia i -tej funkcie príslušnosti, r je parameter prekrytia, ktorý sa zvolí experimentálne.

Po zistení parametrov funkcií príslušností sú signály, ktoré boli privádzané z oboch strán siete, na druhej a na štvrtej vrstve. Výstupy z druhej vrstvy prechádzajú treťou vrstvou cez počiatočné spojenia tretej vrstvy, čím získame silu pravidla $o_i^{(3)}$ každého pravidlového uzla. Na základe síl pravidiel $o_i^{(3)}$ a výstupov zo štvrtej vrstvy $o_j^{(4)}$ potrebujeme určiť výsledné spojenia medzi pravidlovými uzlami a uzlami výstupných termov na štvrtej vrstve. Tieto spojenia sú na začiatku plne prepojené. Na výpočet spojení je podľa [1] použité kompetitívne pravidlo učenia

$$w_{ji}(t+1) = o_j^{(4)}(-w_{ji}(t) + o_i^{(3)}). \quad (9)$$

Premenná $w_{ji}(t)$ je váha spojenia medzi i -tým pravidlovým uzlom a j -tým výstupným term uzlom. Po vypočítaní váh spojení, jednotlivé váhy reprezentujú silu existencie výsledného pravidlového uzla. Zo spojení spájajúcich pravidlový uzol a výstupné term uzly sa vyberie nanajvýš jedno spojenie, ktorého váha ma maximálnu hodnotu. Z toho vyplýva, že každý pravidlový uzol môže byť spojený iba s jedným term uzlom.

Ak sú všetky váhy spájajúce pravidlový uzol a výstupné term uzly veľmi malé, tak sa tieto spojenia môžu vymazať. Znamená to, že daný pravidlový uzol ma malý alebo žiadny vplyv k výstupnej lingvistickej premennej a preto sa môže tento pravidlový uzol odstrániť.

B. Parametrické učenie

Parametrické učenie je druhou fázou hybridného učenia a je použité na optimálne nastavenie parametrov funkcií príslušností. V tejto fáze sieť pracuje v režime zdola-nahor. V parametrickom učení sa používa algoritmus spätného šírenia chyby, pričom hlavným cieľom je minimalizácia chybovej funkcie E , kde $y^d(t)$ je požadovaný a $y(t)$ je aktuálny výstup zo siete

$$E = \frac{1}{2} (y^d(t) - y(t))^2. \quad (10)$$

Následne sa podľa [1] počíta chyba $\partial E / \partial w$ od štvrtej vrstvy po druhú vrstvu, za predpokladu, že w je nastavovateľný parameter v uzle

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial a} \frac{\partial a}{\partial w}. \quad (11)$$

Zároveň sa upravujú stredy a variancie funkcií príslušností podľa požadovaného výstupu a v závislosti na chybe. V nasledujúcej časti popíšem podľa [1] výpočty stredov, variancií a chýb na jednotlivých vrstvách.

5. vrstva

Na tejto vrstve sa nastavujú stredy a variancie iba pre výstupné funkcie príslušností. Použitím rovníc (11) a (6) sa podľa [1] odvodil vzťah pre stred m_i

$$m_i(t+1) = m_i(t) + \eta [y^d(t) - y(t)] \frac{\sigma_i u_i^{(5)}}{\sum_i \sigma_i u_i^{(5)}}, \quad (12)$$

kde η je parameter učenia. Podobne pre varianciu σ_i

$$\sigma_i(t+1) = \sigma_i(t) + \eta [y^d(t) - y(t)] \frac{m_i u_i^{(5)} \left(\sum_i \sigma_i u_i^{(5)} \right) - \left(\sum_i m_i \sigma_i u_i^{(5)} \right) u_i^{(5)}}{\left(\sum_i \sigma_i u_i^{(5)} \right)^2}, \quad (13)$$

Pre chybu platí vzťah

$$\delta^{(5)} = y^d(t) - y(t), \quad (14)$$

4. vrstva

Na tejto vrstve sa nenastavuje žiaden parameter funkcie príslušností. Počíta sa tu len chyba. Z rovníc (6) a (14) sa podľa [1] odvodil vzťah pre chybu

$$\delta_i^{(4)} = [y^d(t) - y(t)] \frac{m_i \sigma_i \left(\sum_i \sigma_i u_i^{(5)} \right) - \left(\sum_i m_i \sigma_i u_i^{(5)} \right) \sigma_i}{\left(\sum_i \sigma_i u_i^{(5)} \right)^2}, \quad (15)$$

3. vrstva

Tak ako vo 4. vrstve, aj v tejto vrstve sa počíta iba chyba. V prípade ak je jeden výstup zo siete, pre chybu podľa [1] platí

$$\delta_i^{(3)} = -\frac{\partial E}{\partial a_i^{(3)}} = \delta_i^{(4)}, \quad (16)$$

Ak je výstupov zo siete viac, pre chybu podľa [1] platí

$$\delta_i^{(3)} = \sum_k \delta_k^{(4)}, \quad (17)$$

pričom chyba pravidlového uzla sa sčítava iba z tých chýb uzlov na 4. vrstve, ktoré sú jeho dôsledkami.

2. vrstva

Na tejto vrstve sa nastavujú stredy a variancie iba pre vstupné funkcie príslušností. Najprv sa počíta chyba. Z rovnice (3) podľa [1] platí

$$\frac{\partial a_i^{(3)}}{\partial u_i^{(3)}} = \begin{cases} 1 & \text{ak } u_i^{(3)} = \min(\text{vstupy pravidlového uzla } i) \\ 0 & \text{ináč} \end{cases}, \quad (18)$$

Z rovníc (16) a (18) podľa [1] platí

$$-\delta_i^{(2)} = \sum_k q_k, \quad (19)$$

pričom suma sa vykonáva z chýb pravidlových uzlov

$$q_k = \begin{cases} -q_k^{(3)} & \text{ak } a_i^{(2)} \text{ je minimum v } k\text{-tom pravidlovom uzle} \\ 0 & \text{ináč} \end{cases}, \quad (20)$$

Použitím rovníc (11) a (2) sa podľa [1] odvodil vzorec pre aktualizáciu stredov m_{ij} vstupných funkcií príslušností

$$m_{ij}(t+1) = m_{ij}(t) + \eta \delta_i^{(2)} \exp\left(\frac{-(u_i^{(2)} - m_{ij})^2}{\sigma_{ij}^2}\right) \frac{2(u_i^{(2)} - m_{ij})}{\sigma_{ij}^2}, \quad (21)$$

a pre variancie σ_{ij}

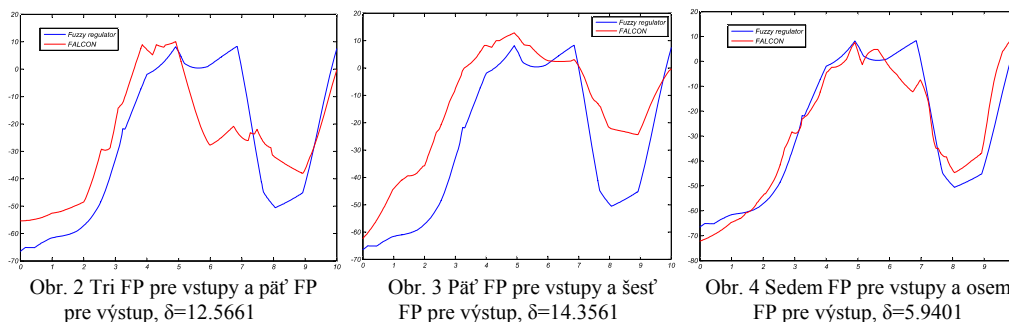
$$\sigma_{ij}(t+1) = \sigma_{ij}(t) + \eta \delta_i^{(2)} \exp\left(\frac{-(u_i^{(2)} - m_{ij})^2}{\sigma_{ij}^2}\right) \frac{2(u_i^{(2)} - m_{ij})}{\sigma_{ij}^3}, \quad (22)$$

IV. VÝSLEDKY EXPERIMENTOV

Cieľom experimentov je porovnať výsledky učenia neurónovej siete FALCON pre rôzne počty funkcií príslušností (FP) a overiť ich na rozličných tréningových dátach.

V experimentoch sme porovnávali výstupy fuzzy regulátora vytvoreného človekom s regulátorom, ktorý bol vytvorený z parametrov vyextrahovaných z FALCON-a. Na posúdenie kvality učenia sme použili priemernú chybu δ (23), t.j. rozdiel medzi výstupom z pôvodného fuzzy regulátora a výstupom z fuzzy regulátora vytvoreného FALCON-om predelený počtom vzoriek.

$$\delta = \sum_{i=1}^n \frac{\|y_i^d - y_i\|}{n}, \quad (23)$$



V. ZÁVER

V tomto článku sme prezentovali neurónovú sieť FALCON, ktorá bola použitá pri tvorbe fuzzy regulátora. Z experimentov sme zistili, že FALCON dosahoval lepšie výsledky pri väčšom počte FP. Na kvalitu učenia má vplyv viacero faktorov. Sú to napríklad počet FP pre jednotlivé premenné, tréningové dáta, parameter prekrytia, parameter učenia. Najlepšie výsledky dosahoval FALCON pri šiestich a siedmich vstupných FP. Priemerná chyba sa tu veľmi často pohybovala v rozmedzí 5 až 9, čo sú veľmi dobré výsledky.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Lin, Ch. T. – Lee, G.: Neural Fuzzy Systems – A Neuro-Fuzzy Synergism to Intelligent Systems; Prentice Hall PTR, New York, 1996, s. 533-544, ISBN 0-13-235169-2
- [2] Vaščák, J.: Fuzzy logika v regulácii – Podrobný úvod do problematiky fuzzy riadenia; 2008
- [3] Fuzzy Logic Toolbox – User's Guide; dostupné na internete: http://www.mathworks.com/access/helpdesk/help/pdf_doc/fuzzy/fuzzy.pdf
- [4] Virtual Reality Toolbox – User's Guide; dostupné na internete: http://www.mathworks.com/access/helpdesk_r13/help/pdf_doc/vr/vr.pdf

From Imitation Learning to Innovation in Designing Behaviors for Humanoid Robots

Mária VIRČÍKOVÁ, Peter SINČÁK

Dept. of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and
Informatics, Technical University of Košice, Slovak Republic

{maria.vircikova, peter.sincak}@tuke.sk

Abstract — The proposed research deals with a usage of artificial intelligent techniques in humanoid robotics. The objective is to design an interactive robotic behavior. The focus is on social robotics and how to improve the interaction between human and robot. In the first part user interactively cooperates with the System and this is the way how the robot acquires knowledge. Our approach is based on real manipulation where the robot freely interacts with humans and objects, and learns during continuous interaction with the environment.

In the second part of the system we use the Interactive Evolutionary Computation – a technique which optimizes systems based on subjective human evaluation. So far, the algorithm is applied to a system of evolution of pleasant motions or dances. It adapts to a user's preferences and we think adaptability can be considered a requisite feature of intelligence. A user can build a robot behavior without any programming knowledge and then the evolutionary algorithm helps user to discover new possibilities of motions.

Keywords — cognitive architecture, evolutionary computation, human—robot interaction, interactive evolution, learning from demonstration, robot NAO

I. INTRODUCTION

*“Art begins in imitation
and ends in innovation.”
Mason Cooley (1927—2002)*

The focus of robotic research continues to shift from industrial environments, in which robots perform a repetitive task in a very controlled environment, to mobile service robots operating in a wide variety of environments, often in human—habited ones. New applications for robots make grow many problems which must be solved in order to increase their autonomy. These problems are as said in [3], but are not limited to, navigation, localization, behavior generation and human—machine interaction.

According to Brooks [1], the implicit dream of Artificial Intelligence (AI) has been to build human level intelligence. Building a humanoid is the challenge par excellence for AI and robotic workers. AI tries to make programs or computers do things that, when done by people, are described as having indicated intelligence. The goal of AI has been characterized as both the construction of useful intelligent systems and the understanding of human intelligence. Since AI's earliest days there have been thoughts of building truly intelligent autonomous robots. In academic research circles, work in robotics has influenced work in AI and vice versa [2]. Recent generations of humanoid robots (HRs) increasingly resemble human in shape and capacities. HRs are robots that are at least loosely based on the appearance of the human body. We work with humanoid robot Nao developed by Aldebaran Robotics. The robot (fig. 1) has a human—like appearance and various sensors for interacting with humans.

In community of AI, by humanoid we mean not only a robot that has the physical appearance of a human, but also has a fairly advanced artificial intelligence, allowing it to reasonably approximate human behavior and interactions. New trends in research consider the robots ability to interact with people safely and in natural way. The current leading topics in HRs are so—called service and social robotics — where the target is getting robots closer to people and

their social needs. There are many attempts to create robots capable of expressing emotions or communicating in a natural language.

This paper starts by reviewing the basic features of the learning by imitation paradigm and considers a new way how to map the teacher skills to a robotic agent during the interaction – teaching process. We suggest an idea of learning behaviors one robot from another. The paper continues with a discussion of the emergence of creativity – innovative behaviors – by employing the Interactive Evolution in the development of the behaviors (which are in our case motor skills of the robot).

II. LEARNING FROM DEMONSTRATION

A. Robots learning by imitating human

An enduring challenge in social robotics has been to improve the interaction between human users and robots – it is still unnatural, slow and unsatisfactory for the human interlocutor. Robots have different abilities, knowledge, and expectations than humans. They need to react appropriately to human expectations and behavior. With this respect, scientific advances have been made to date for applications in entertainment and service robotics that largely depend on intuitive interaction.

Dautenhahn and Billard [8] called the type of systems which we are trying to develop *social robots*: Embodied agents that are part of a heterogeneous group: a society of robots or humans. They are able to recognize each other and engage in social interactions, they possess histories – perceive and interpret the world in terms of their own experience – and they explicitly communicate with and learn from each other.

It is a huge research field including gesture and natural language communication, perceiving emotions and expressing their own artificial emotions, establishing social relationships, exhibition of different personalities and characters, recognition of interaction partner and many others. In this paper we present our work in **learning motion tasks from observation**.

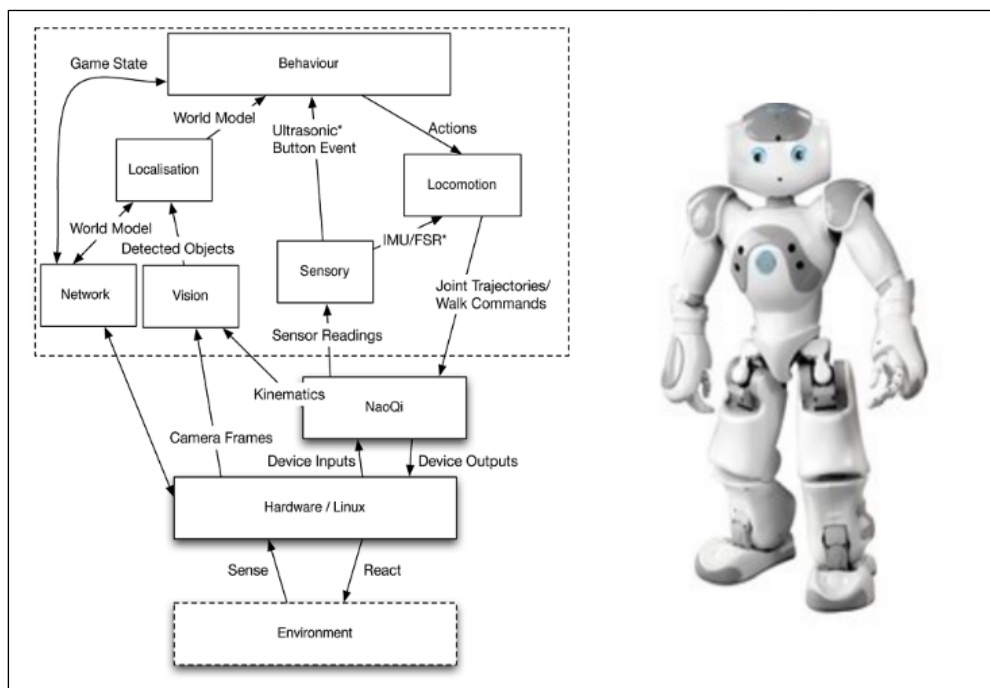


Fig. 1 Humanoid robot Nao comes with 25 degrees of freedom for great mobility. The inertial sensor and closed loop control provide great stability while moving and enables positioning within space. Its state-of-the-art onboard actuators give Nao extreme precision in its movements. It supports multiple programming environments and programming languages like URBI Script, C++, Python, .Net Available on Linux. It can be programmed and controlled using Linux, Windows or Mac OS. The most used simulators for Nao are Microsoft Robotics Studio, Gostai Urbi Studio, Pyro, Choregraphe and Webots. The position in angles of the degrees of freedom – servos or motors of the robot will represent the phenotype of the genetic algorithm. The robot is running on a Linux platform and can be programmed using a proprietary SDK called NaoQi. It is based on a client-server architecture, where NaoQi itself acts as a server. The modules plug into NaoQi either as a library or as a broker, with the latter communicating over IP with NaoQi. Running a module as a broker thus has some IP-related overhead, but also allows the module to be run on a remote machine, which is very useful for debugging. NaoQi itself comes with modules which we use.

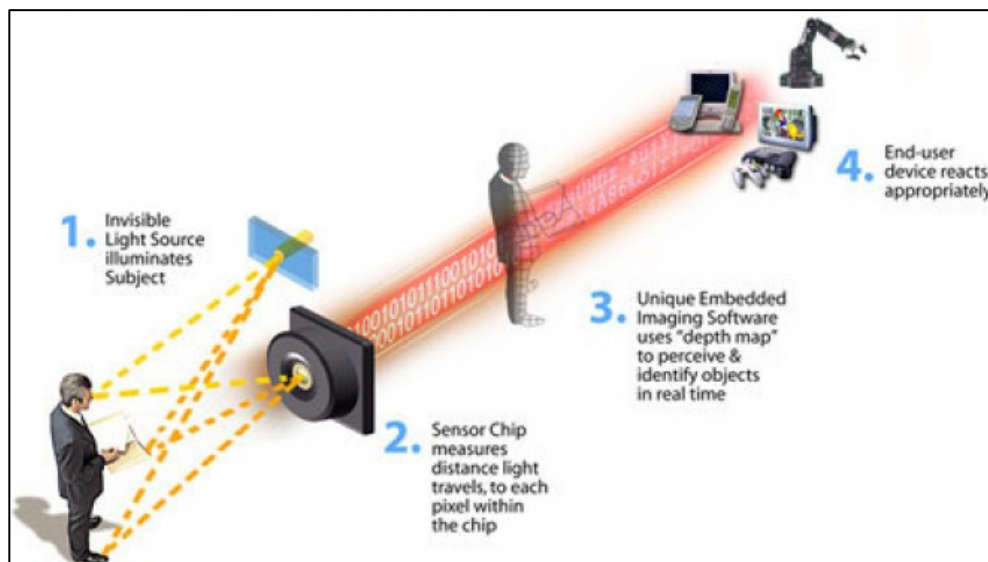


Fig. 2 Article [21] describes how Kinect works: It generates a three-dimensional (moving) image of the objects in its field of view, and recognizes (moving) human beings among those objects.

Authors [9] formulated a statement: *Imitation takes place when an agent learns a behavior from observing the execution of that behavior by a teacher.* They also enumerated the advantages a learning mechanism should give: adaptation, efficient communication, compatibility with other learning mechanisms, efficient learning and thus, learning by imitation seems to be one of the major components of general intelligent behavior. Imitation encompasses a set of different competences such as recognizing other people's actions, recognizing the goal of a particular action and the objects and/or subjects involved. This project attempts to implement a similar set of abilities in a humanoid robot following a biologically motivated perspective.

Learning tasks from demonstrations offers a convenient mechanism to customize and train a robot by transferring task related knowledge from a user to a robot. This avoids the time-consuming and complex process of manual programming. The way in which the user interacts with a robot during a demonstration plays a vital role in terms of how effectively and accurately the user is able to provide a demonstration. Teaching through demonstrations is a social activity, one that requires bidirectional communication between a teacher and a student.

Approach "Learning these tasks from demonstrations according to [10] is a *convenient mechanism to customize and train a robot by transferring task related knowledge from a user to a robot.* This avoids the time-consuming and complex process of manual programming. The way in which the user interacts with a robot during a demonstration plays a vital role in terms of how effectively and accurately the user is able to provide a demonstration. Teaching through demonstrations is a social activity, one that requires bidirectional communication between a teacher and a student. The key question was how to **map a teacher action to an agent action**. These agent actions in our experiments are the motor capabilities of the robots.

We mimic some user movements using Microsoft Kinect sensor device, a horizontal bar connected to a small base with a motorized pivot and is designed to be positioned lengthwise above or below the video display. The device features an "RGB camera, depth sensor and multi-array microphone running proprietary software" [11], which provide full-body 3D motion capture, facial recognition and voice recognition capabilities. Figure 3 illustrates how Kinect sensor tracks a user. As described in [21], "Kinect relies on depth mapping to determine an object's location in a field. Much like sonar, the unit measures the time it takes for light to bounce off the object and return to the source. It creates a 3D visualization of the object this way. This method offers greater reliability than older techniques, which relied on color and texture differences to determine the location and nature of an object. In that model, the results can be thrown off by low contrast situations, like wearing a shirt the same color as the wall behind you. Depth mapping also doesn't rely on external light sources to measure anything. Kinect will ostensibly work in a pitch black room, because it is generating its own "light pings." Once the device has data to process, the processor crunches the data to make it work. It differentiates between faces, limbs, and other environmental aspects of the field, as well as recognizes people and remembers who they are when they step back in the field."

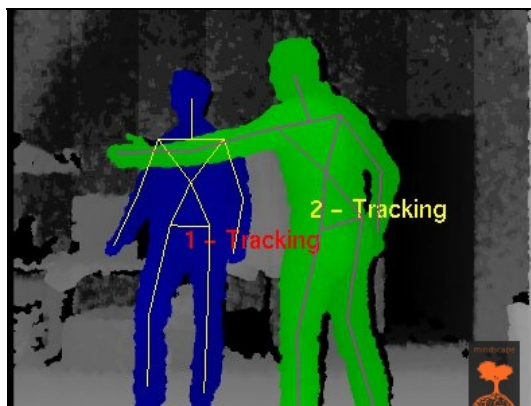


Fig. 3 Skeleton tracking of users using Kinect sensor.

B. Robots learning from each other

At the present time we are trying to model different robot's personalities via its interaction with several human users and this way we personalize robot's behaviors according to the interaction of the specific human user. This way various models for multiple robot approach are developed. Robots are interacting between each other and are learning new motion tasks in the community of robots.

When robots will start to learn from each other they will be able to adapt much faster to new environments and to perform even much complicated tasks. This approach means also that all robots within the group could acquire tremendous quantities of useful knowledge and they would not have to reinvent their abilities and skills from beginning again.

III. DISCOVERING NEW BEHAVIORS ACCORDING TO USER'S PREFERENCE VIA INTERACTIVE EVOLUTION

A. Genetic Algorithm: The wisdom is in the population

Jorge Luis Borges in his *The Library of Babel* writes about a library which consists of indeterminate and infinite number of books. Most of them contain nonsense and rational relation is almost miraculous exception, but somewhere there are books that contain all the knowledge of the universe.

A traveler in such a Library wanting to find something rational has a very difficult task, as the most of the books are just a cluster of incomprehensible letters. How we could help this traveler to find a "masterpiece" in the endless space of options". We need a tool that would fly in this space effectively and thanks to it we would find something inspiring. One of such a tool could be an evolutionary algorithm.

An evolutionary algorithm is an optimization algorithm which uses some mechanisms inspired by biological evolution coming out from Darwinian Theory. Evolutionary computation is a biologically inspired general computational concept and includes Genetic Algorithms (GA), Genetic Programming and Evolutionary Strategies. GA provides a very efficient search method working on population, and has been applied to many problems of optimization and classification.

General GA process is as follows:

- 1) Initialize the population of chromosomes.
- 2) Calculate the fitness for each individual in the population using **fitness function**.
- 3) Reproduce individuals to form a new population according to each individual's fitness.
- 4) Perform crossover and mutation on the population.

Crossover selects genes from parent chromosomes and creates a new offspring.

After a crossover is performed, mutation takes place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring.

- 5) Go to step (2) until some condition is satisfied.

The EC is a population-based searching algorithm and outputs multiple candidates, each called an individual, as system outputs. The space of all feasible solutions (it means objects among those the desired solution is) is called search space (also state space). Each point in the

search space represents one feasible solution. Each feasible solution can be "marked" by its value or fitness for the problem.

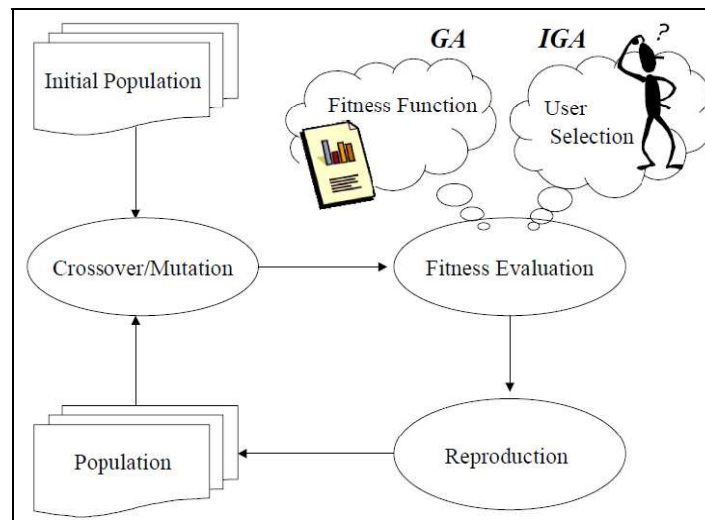


Fig. 4. The process of Genetic Algorithm and Interactive Genetic Algorithm where the fitness function is replaced by the subjective user evaluation [5].

A **fitness function** is a particular type of objective function that prescribes the optimality of a solution (that is, a chromosome) in a genetic algorithm so that that particular chromosome may be ranked against all the other chromosomes. Optimal chromosomes, or at least chromosomes which are *more* optimal, are allowed to breed and mix their datasets by any of several techniques, producing a new generation that will (hopefully) be even better. We could say that the fitness function is at the heart of an EC application because it determines which solutions within a population are better at solving the concrete problem.

In reality, there are two types of target systems for system optimization: systems whose optimization performances are numerically — or at least quantitatively — defined as evaluation functions and systems whose optimization indexes are difficult to specify. How would you mathematically express your feelings looking at some object of art? There are many applications when the fitness function cannot be explicitly defined – not only in the art domain. The second question is whether the evolutionary algorithms could be verily designated as innovative.

B. Interactive Evolution: a Source of Inspiration

Evolution's use of competition, variation, reproduction and selection provides an inspiration for a method creating artificial system [15]. Fogel [12] has written that design variations acquired by such systems are passed on to their offspring. We examine an interactive genetic algorithm (IGA) that serves as a creativity tool for users.

Interactive Evolutionary Computation, introduced by Takagi [5], is an optimization method that adopts EC among system optimization based on subjective human evaluation. It is simply an EC technique whose **fitness function is replaced by a human user** or better said; it uses human's response as a fitness value. The user acts as critic of the robotic behaviors. We can say that the IEC is a technology that embeds **human preference, intuition, emotion, psychological aspects, or a more general term, KANSEI**, in the target system.

The goal of our design of pleasant movements is to create some 'nice motion performance'. However, there is no general standard of 'beauty of motions', and it is almost impossible to organize fitness function. IGA might be a solution for this: it can reflect personal preference because it perceives fitness directly from user instead of computing using some function. Users evaluate robotic motions and give them marks according to their preferences. The process starts from motion performances learned before (Section II: Learning from Demonstration). After user evaluates the first generation of motions of the robots, the algorithm produces the next generation. We use the tournament selection method: the winner of each tournament is the one with the highest fitness value and this one is selected for crossover.

Goldberg [22] formulated the "innovation intuition" within these two relations:

- 1) **Selection + Mutation = Continual Improvement**
- 2) **Selection + Recombination = Innovation**

In GA, the crossover & the mutation are sources of variation. Much research has been done where IGAs are used to **assist users in the development of creative ideas**. (Evaluating creativity is of course still mostly subjective.) Under creativity we mean producing something new or unexpected and valuable. As reported by [20] a well known inhibitor of creative thinking is design fixation, where the suggestion of a particular solution makes it difficult to imagine other good solutions. Unlike many other computational search algorithms, EC methods work with populations of “fairly good” solutions. Therefore, there is less danger that creativity will be harmed by design fixation on one “best” solution.

IV. DESIGN OF ROBOTIC PLEASANT MOTIONS

Our goal was to discover new frontiers of humanoid behaviors for aesthetic pleasure, entertainment or communication. In the field of developing robotic dances, numerous works are presented in [1]. Researchers at the University of Tokyo have developed the learning—from—observation training method, which enables a robot to acquire knowledge of what to do and how to do it from observing human demonstrations. At the Kyoto University applied a method called intermodality mapping to generate robot motion from various sounds and also to generate sounds from motions using the back—propagation through—time algorithm. Other approach from Tokyo University is using Chaos to trade synchronization and autonomy in a dancing robot. Dancing Robot Partner built at the Tohoku University is well—known in the domain of robotics. Their robot acts as a female partner and realizes ballroom dances in coordination with a male human dancer by estimating his intention.

Although some of the mentioned systems are examples where interaction between human and robot is the key factor of their success, they are exceptions in the world of robotic dance. The majority of dance systems of robots are only pre—programmed motions and after a while user find them boring.

The proposed system in this work interacts with human and the motion is evolving in accordance with his evaluation of the seen behaviors learned in the first part of the project. The system is a good aid for inexperienced observers to create their own “robotic dance”. The result is that every person can adapt, in our case, the dance choreography, in accordance to his own expectations and preferences.

V. CONCLUSION

In the first part of the project we engage people in face—to—face interaction with a humanoid robot Nao. A naive human teacher can develop new motion tasks via interaction process. In the second part we presented a multi—robot system capable of evolution of motions using an interactive evolutionary computation. We believe that the interactive evolutionary computation will make interactive robots more personal and will help user to creatively discover robot’s behaviors during the interaction process that he does not imagine before.

Thus, there is a need to extend the search space in our model with an explicit expansion of the robot behaviors solution space by adding more variables.

Our future dream is for Nao to acquire its own mental model of people. Currently, he does not reason about the emotional state of others. Thus, we want to extract the information about the own preferences of human during his evaluation of the behavior of the robot in the interactive evolutionary algorithm and make this process autonomous.

According to Brooks and Breazeal [2]:

“The ability to recognize, understand, and reason about another’s emotional state is an important ability for having a theory of mind about other people, which is considered by many to be a requisite of adult—level social intelligence.”

ACKNOWLEDGMENT

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] Aucouturier, J. J.—Ikeuchi, K. et. al.: *Cheek To Chip: Dancing Robots and AI’s Future*. In Intelligent Systems, Ieee, 23(2): Pp. 74-84, 2009.
- [2] Brooks, R.A.—Breazeal, C. et. al.: *The Cog Project: Building A Humanoid Robot*. In: Computation For Metaphors, Analogy and Agents. C. Nehaniv (Ed), Lecture Notes In Artificial Intelligence 1562. New York, Springer, Pp. 52-87, 1999.
- [3] Garcia, E.—Jimenez, M., A. et. al.: *The Evolution Of Robotics Research, From Industrial Robotics To Field And Service Robotics*. In: Ieee Robotics And Automation Magazine, March 2007.

- [4] Gouaillier, D.—Hugel, V. et. al.: *The Nao Humanoid: A Combination Of Performance And Affordability*. In: Ieee Transactions On Robotics, 2008.
- [5] Takagi, H.: *Interactive Evolutionary Computation. Cooperation Of Computational Intelligence And Human Kansei*. In: Proceedings Of The Ieee, Vol. 89, Issue 9, Pp. 1275 - 1296, Issn: 0018-9219, September 2001.
- [6] Vircikova, M.: *Artificial Intelligence In Humanoid Systems*, Diploma Thesis, 2010.
- [7] Fong, T.—Nourbakhsh, I.—Dautenhahn, K.: *A Survey of Socially Interactive Robots: Concepts, Design and Applications*, technical report No. Cmu-Ri-Tr-02-29, Carnegie Mellon University 2002.
- [8] Dautenhahn, K.—Billard, A.: *Bringing up Robots or The Psychology of Socially Intelligent Robots: from Theory to Implementation*. In: Agents '99, Proceedings of the third annual conference on Autonomous Agents, ISBN 1-58113-066-X, ACM New York, 1999.
- [9] Berthouze, L.—Bakker, P.—Kuniyoshi, Y.: *Learning of Oculo-Motor Control: A prelude to Robotic Imitation*. In: Proceedings of the 1996 Lee/RSJ International Conference on Intelligent Robots and Systems, 1996.
- [10] Koenig, N.—Takayama, L.—Mataric, M.: 2010 Special Issue: *Communication and Knowledge Sharing in Human-Robot Interaction and Learning from Demonstration*. In: Journal Neural Networks, Vol. 23, Issue 9, Elsevier Science ltd. Oxford, 2010.
- [11] Totilo: Natal Recognizes 31 Body Parts, Uses Tenth Of Xbox 360, Computing Resources, *Kotaku*, Gawker Media Retrieved November 25, 2010.
- [12] Fogel, D.: *Evolutionary computation: Toward a new philosophy of machine intelligence, 2nd edition*. IEEE Press, 2000.
- [13] Calinon, S.—Billard, A.: *Incremental learning of gestures by imitation in a humanoid robot*, in: Proceedings of the 2nd ACM/IEEE International Conference on Human_Robot Interactions, HRI'07, 2007.
- [14] Ogino, M.—Toichi, H.—Yoshikawa, Y.—Asada, M.: *Interaction rule learning with a human partner based on an imitation faculty with a simple visuomotor mapping*, in: The Social Mechanisms of Robot Programming by Demonstration, Robotics and Autonomous Systems 54 (5), 2006.
- [15] Goldin, D. — Keil, D.: *Interaction, Evolution and Intelligence*, IEEE Press, 2001.
- [16] Ishiguro, T.—Ono, M. Imai—T. Kanda.: *Development of an interactive humanoid robot "Robovie"— An interdisciplinary approach*, in *Robotics Research*, R. A. Jarvis and A. Zelinsky, Eds. New York: Springer-Verlag, 2003.
- [17] Banerjee, A.—Quiroz, J.C.—Louis, S. J: *A Model of Creative Design Using Collaborative Interactive Genetic Algorithms*, pages 397–416. Springer, 2008.
- [18] DiPaola, S.—Gabora, L.: *Incorporating characteristics of human creativity into an evolutionary art algorithm*. In (D. Thierens, Ed.). Proceedings of the Genetic and Evolutionary Computing Conference (GECCO). July 7 — 11, 2007, University College London, England, pp. 2442-2449, 2007.
- [19] Boden, M.A.: *What is creativity?*, in Boden (ed.), Dimensions of creativity, The MIT Press, Cambridge, Massachusetts, USA, 1994.
- [20] Buelow, P.: *Advantages Of Evolutionary Computation Used For Exploration In The Creative Design Proces*. In: Journal of Integrated Design and Process Science, IOS Press, ISSN 1875-8959, Vol. 11, No 3, 2007, online date 2009.
- [21] Carbody, T.: Wired, Gadget Lab: *How Motion Detection works in Xbox Kinect*. November 3, 2010. <http://www.wired.com/gadgetlab/2010/11/tonights-release-xbox-kinect-how-does-it-work/all/>
- [22] Goldberg, D.: *The Design of Innovation: Lessons from Genetic algorithms for the Real World*. In Technological Forecasting And Social Change, Vol. 64 (2000), Pp. 7-12.

Analýza historických dát z kolaboratívneho prostredia KP-LAB

Matúš FEDOR, František BABIČ, Stanislav PETRULA

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

matus.fedor@student.tuke.sk, stanislav.petrula@student.tuke.sk,
frantisek.babic@tuke.sk

Abstrakt — Tento článok popisuje možné prístupy pre analýzu dát reprezentujúce akcie a aktivity používateľov v kolaboratívnom prostredí KP-LAB. V tomto prípade budú prezentované dva prístupy a to automatické dolovanie zaujímavých sekvencií prostredníctvom asociačných pravidiel alebo manuálna identifikácia na základe používateľských preferencií a požiadaviek. V prvom prípade bol na analýzu bol použitý softvérový nástroj spoločnosti IMB, SPSS Clementine. Získané dáta z kolaboratívneho prostredia vo forme logov boli predspracované z pohľadu konzistencie, čistoty dát a ich úplnosti. V druhom prípade bola navrhnutá a implementovaná vlastná aplikácia v jazyku C#, ktorá umožňuje vytvorenie vlastných vzorov zaujímavých sekvencií.

Príručné slová — kolaboratívne prostredie, logy, asociačné pravidlá, sekvencie akcií

I. ÚVOD

Tento článok popisuje spôsob spracovania a následnej analýzy záznamov z kolaboratívneho systému KP-LAB. Extrahované a zobrazené informácie reprezentujú prehľad správania používateľov počas kolaboratívnych procesov. Samotný záznam (log) obsahuje 14 položiek, ktoré predstavujú dostatočné historické údaje pre analytické účely. Tieto dáta boli zozbierané počas finálnej fázy vývoja systému, preto ich kvalita je z časového pohľadu rôznorodá: integrácia monitorovacích služieb, úprava používateľského prostredia s cieľom efektívneho logovania, úprava monitorovacích služieb z pohľadu výkonnosti a úplnosti logovaných dát, zber dát z reálnych pilotných kurzov. Tieto aspekty boli dôležitou motiváciou pre predspracovanie s cieľom získať konzistentnú množinu dát pre aplikáciu vhodných analytických prístupov.

II. HISTORICKÉ DÁTA

A. Zdroj údajov a ich formát

Historické dáta pochádzajúce z prostredia KP-Lab boli k dispozícii vo formáte sql v objeme viac ako 118 tisíc záznamov. Každý log je popísaný 14 atribútmi [3]:

[ID, Type, Actor, Actor Type, Actor Name, Entity, Entity Type, Entity Title, Belongs to, Time, Custom properties, Custom data]

- ID – unikátny identifikátor logu
- TYPE – typ vzniknutej akcie, napr. vytvorenie, vymazanie, editácia
- ACTOR – unikátny identifikátor autora, ktorý realizoval danú akciu
- ACTOR TYPE – rola autora (používateľa), ktorá je delegovaná na základe relevantnej časti užívateľského rozhrania, napr. študent, administrátor, učiteľ
- ACTOR NAME - užívateľské meno získané z modulu správy používateľov
- ENTITY – unikátny identifikátor zdieľaného objektu, ktorý je predmetom danej akcie
- ENTITY TYPE – typ zdieľaného objektu, napr. úloha, dokument, linka, komentár
- ENTITY TITLE - konkrétny názov daného zdieľaného objektu
- BELONGS TO – unikátny identifikátor príslušnej časti užívateľského prostredia, kde bola daná akcia realizovaná
- TIME – čas, kedy bola udalosť uložená do databázy (vo formáte HH:MM:SS)
- CUSTOM DATA AND PROPERTIES – tento parameter sa používa v situácií, keď si konkrétna používateľská aplikácia uloží nejaké premenné alebo údaje typické pre ňu.

B. Popis dát

Surové dáta sú uložené v MySQL databáze, viď. obr. 1:

Browse activities - all

Showing max 100 activities starting from 115820. Number of all activities in database is 115919

Navigate: [previous](#) [next](#) | [first](#) [1153](#) [1154](#) [1155](#) [1156](#) [1157](#) [1158](#) [1159](#) [1160](#) [1161](#) [1162](#) [last](#)

ID/Type	Actor/Type/Name	Entity/Type/Title	Belongs To/Refers To/Used Tool
115823 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315-1311-55d47be9-2fb9-4844-8cbd-6970cbb5865f user	http://www.kp-lab.org/ontologies/nn#Note_100413-2336-b1a8cc9f-79a5-41c9-a69c-526db03d6de0 note	http://www.kp-lab.org/system-model/TLO#CollectiveSpace:2010-03-23T12:41:07.541b126-4bb5-a799-8253c739d380 --
115824 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315-1311-55d47be9-2fb9-4844-8cbd-6970cbb5865f user	http://www.kp-lab.org/system-model/TLO#UploadableContentItem_100414-0057-df6cbb9f-a324-45e5-a9dc-f2b86edbf file	http://www.kp-lab.org/system-model/TLO#CollectiveSpace:2010-03-23T12:41:07.541b126-4bb5-a799-8253c739d380 --
115825 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315-1311-55d47be9-2fb9-4844-8cbd-6970cbb5865f user	http://www.kp-lab.org/system-model/TLO#UploadableContentItem_100420-0154-9bc4e2ed-f675-4490-a30a-4c50d546be76 file	http://www.kp-lab.org/system-model/TLO#CollectiveSpace:2010-03-23T12:41:07.541b126-4bb5-a799-8253c739d380 --

Obr. 1 Prehľad uložených logov z prostredia KP-Lab cez webové rozhranie

Ako vidno z obrázku, nie všetky parametre logu sú vždy vyplnené, majú hodnotu null. Táto situácia sa zlepšila postupným vylepšovaním logovacích služieb a ich komunikácie s prezentačnou vrstvou. Aktuálne uchovávané logy sú podstatne detailnejšie. Na obr. 2 je zobrazený popis dát získaný nástrojom SPSS Clementine. K jednotlivým atribútom logu na ľavej strane patria všetky typy hodnôt, ktoré nadobudli v dátach (udalosti). Získané údaje bolo nutné predspracovať pomocou vhodných techník predspracovania (ako je napr. integrácia, normalizácia). Záznamy je možné analyzovať pomocou ďalších funkcií v SPSS Clementine, ako je napríklad funkcia zobrazenia dát, histogram, matice, tabuľky.

Activity	Event			
act_id	ID			
act_type	activity	create	Login User	Register New User
	assigning to phase	creation	modification	unspecified group
	assigning to session	Creation	modify	
	assigning to theme	delete	none	
	assignment	deletion	opening	
actor_id	Subjekt ID			
actor_type	user	teacher		
	"bez hodnoty"	student		
	NULL	individual		
	Map-It User	User		
actor_name	v 50% nezadaná hodnota = NULL, inak meno			
entity_id	Object ID			
entity_type	shared space	note	comment	

Obr. 2 Popis vybraných parametrov logu prostredníctvom nástroja SPSS Clementine

III. PREDSPRACOVANIE DÁT

Na to, aby sa údaje vo forme logov mohli analyzovať, je nutné ich vyčistiť, odstrániť zašumené a nekonzistentné dáta. Až potom je možné relevantné údaje z rôznych zdrojov integrovať.

44009	44282 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...	user	Kari-Olli Ko...	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...
44010	44283 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...	user	Kari-Olli Ko...	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...
44011	44284 opening	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...	user	Kari-Olli Ko...	http://www.kp-lab.org/ontologies/mapit#MapItAgent_080315...
44012	\$null\$	44285		modificat...	http://www...

Obr. 3 Príklad nekonzistentných dát

Na obr. 3 sú zobrazené problémy, s ktorými sme sa stretli pri práci s log záznamami. Jedným z nich sú nulové hodnoty, kde je potrebné rozlíšiť či daná chýbajúca hodnota neexistuje v reálnom svete (v našom systéme) alebo len nebola zadaná a je ju možné dopočítať, prípadne doplniť expertom [3]. Na obrázku je tiež prípad, kedy je záznam nekonzistentný, zle zapísaný a pre ďalšie analýzy nepoužiteľný. Tieto záznamy boli následne zmazané. Záznamy, ktorých atribúty obsahovali nulové hodnoty, boli taktiež zmazané vzhľadom na nemožnosť ich doplnenia. Toto nie najlepšie riešenie však bolo časom nahradené kvalitnejšími dátami, v rámci ktorých sa počet nulových hodnôt podstatne znížil.

Ďalším krokom bolo odstránenie rôznych foriem popisu tých istých hodnôt atribútov, viď.

obr. 4, čo bolo spôsobené postupným vývojom a integráciou logovacích služieb rôznymi používateľskými nástrojmi na prezentačnej vrstve.

Value	Proportion	%	Count
opening	52,58	48884	
modification	18,62	17308	
creation	13,25	12317	
unspecified group	13,06	12145	
deletion	2,2	2041	
assigning to theme	0,09	83	
assigning to phase	0,05	44	
assignment	0,04	38	
Login User	0,03	24	
create	0,02	21	
assigning to session	0,02	20	
Creation	0,02	16	
modify	0,01	13	
activity	0,0	3	
none	0,0	3	
delete	0,0	2	
	0,0	1	
Register New User	0,0	1	

Obr. 4 Príklad všetkých možných hodnôt atribútu, ktorý popisuje typ realizovanej aktivity

V tomto prípade ide napr. o zjednotenie hodnôt „creation“, „create“ and „Creation“ do jednej. Takýto prístup bol aplikovaný na všetky nutné možnosti s cieľom získať zoznam základných typov akcií, ktoré mohli používatelia realizovať.

IV. EXPERIMENTY

A. Dolovanie asociačných pravidiel

SPSS Clementine¹ je nástroj, ktorý slúži na dolovanie dát, umožňuje rýchly vývoj prediktívnych modelov použitím podnikových expertíz a ich rozvoj v organizáciách pre podporu rozhodovania. Jeho návrh s využitím priemyselného štandardu CRISP-DM [2], umožňuje vykonávať celý proces dolovania v dátach, od identifikácie cieľov až po nasadenie do praxe.

Využitím algoritmu Apriori, ktorý je implementovaný v Clementine, boli realizované experimenty na získanie asociačných pravidiel. Ako výsledný parameter bol postupne vybraný typ akcie, typ objektu a identifikátor subjektu; ako podmienkové parametre boli vždy nastavené ostatné parametre oproti výslednému. Tento prístup priniesol pravidlá ako IF Entity = 2049 & TYPE = modification THEN ACTOR = 31 (ak objekt 2049 vykonával modifikáciu, potom na 98% modifikoval objekt 31, toto pravidlo má vysokú presnosť 98,31% avšak nízku podporu 1,18%, preto sa vyskytuje len veľmi zriedka v celej množine dát) alebo IF Actor type = user & Entity type = shared space THEN Type = opening (ak používateľ (user) vykonával akciu na type objektu shared space, potom na 71% vykonával akciu otvárania. Toto pravidlo má pomerne vysokú presnosť 62,78% aj podporu 71,55% a preto popisuje veľkú časť množiny dát) – toto zistenie predstavuje typickú situáciu pre kolaboratívne virtuálne prostredia, kde väčšinu akcií reprezentujú akcie prehľadávania prostredia samotného alebo jednotlivých zdieľaných objektov. Cieľom je získať prehľad o možnostiach, ktoré nám dané prostredie a jeho elementy ponúka pre dosahovanie stanovených cieľov. Napriek tomu, získané asociačné pravidlá nemajú vysokú hodnotu pre napr. učiteľa, ktorý chce zhodnotiť priebeh výukového kurzu, zistiť ako pracujú jeho študenti a aké prístupy používajú na riešenie konkrétnych úloh. Preto sme sa rozhodli navrhnúť aplikáciu pre manuálnu definíciu potenciálne zaujímavých vzorov, ktoré budú následne vyhľadávané v databáze historických údajov.

B. Identifikácia zaujímavých sekvencií

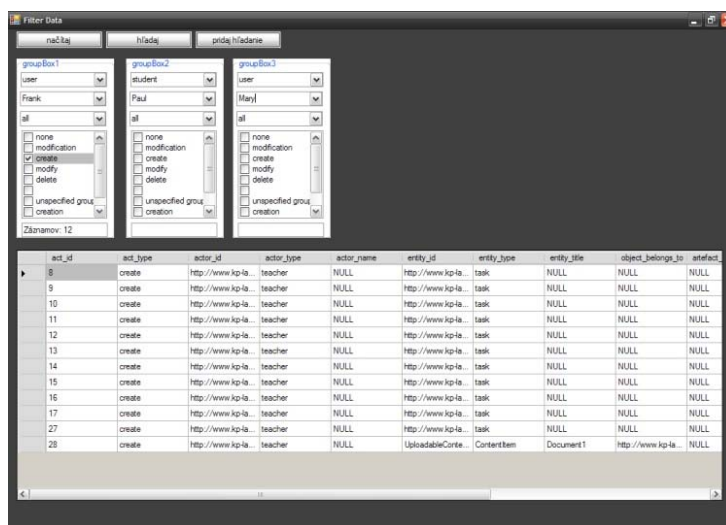
Manuálna identifikácia zaujímavých sekvencií znamená možnosť vytvoriť vzor na základe používateľských preferencií alebo požiadaviek, ktorý môže obsahovať vybrané atribúty logu. To znamená, že nemusí ísť len o jednoduchú kombináciu akcií, ale používateľ si môže overiť napr. postupnosť akcií, ktoré boli realizované na danom objekte vybranými prispievateľmi. Napr. chcem vedieť, či pri vytvorení daného objektu, ktorým bola prednáška vo forme pdf, nasledovali nejaké ďalšie akcie zo strany študentov, ktoré by signalizovali ich záujem o prednesenú látku a prínos danej prednášky pre ich ďalšiu prácu.

Na tieto účely bola navrhnutá a implementovaná aplikácia v jazyku C#, ktorá má uľahčiť hľadanie vzorov správania z dostupných dát systému KP-Lab. Obrazovka pozostáva z vyhľadávacieho okna, priestoru pre zobrazenie výsledkov a funkčných tlačidiel pre načítanie, vyhľadávanie a pridanie ďalšieho vyhľadávacieho okna.

¹ <http://www.spss.com/software/modeler/>

Načítanie dát je možné vo formáte Excelovskej tabuľky, vykoná sa po stlačení tlačidla *načítať*, v zobrazovacom okne sa zobrazia všetky dostupné dáta a vo vyhľadávacom okne sa aktivujú doteraz prázdne vyhľadávacie polia. Aktivácia znamená, že aplikácia si načíta používateľov podľa mena, k nim relevantné úlohy, ktoré vykonávali a samozrejme aj operácie s ktorými pracovali.

Vo vyhľadávacom okne, je možné zvoliť kritéria, podľa ktorých sa majú údaje prehľadávať. Príklad je uvedený na obr. 5, kde sa vyhľadávajú všetky aktivity užívateľa, na základe používateľom zvolených kritérií. Aplikácia nám ponúkne zoznam všetkých užívateľov, ktorých aktivita bola zaznamenaná v log záznamoch systému KP-LAB. V názornom príklade je vybratý užívateľ Frank, k nemu nám aplikácia ponúkne zoznam všetkých úloh, s ktorými Frank pracoval, necháme však zvolenú možnosť *all* (všetky úlohy), na záver z checkBoxListu vyberieme aktivity, ktoré chceme skúmať, v našom prípade bolo zvolené *create*. Takto sme vytvorili prvú časť vzoru, ku ktorej môžeme podobným postupom doplniť ďalšie časti. Jednotlivé časti vzoru sú spájané prostredníctvom logických spojok ako AND alebo OR. Tento variant ponúka variabilitu pri definovaní vzoru, keďže umožňuje hľadať aj podobné výskyty definovaného vzoru.



Obr. 5 Príklad definície vzoru, ktorý bude následne vyhľadaný v databáze historických údajov, a jeho výskyty budú zobrazené v zozname dole

V. ZÁVER

Analýza správania používateľov v kolaboratívnom prostredí na základe logov ich aktivít je dôležitým prvkom pri vyhodnocovaní, pri identifikácii zaujímavých prístupov pre riešenie úloh alebo návrhu nových funkcií pre dané kolaboratívne prostredie. Cieľom tohto článku bolo ukázať potenciálne zaujímavý prístup v tejto oblasti, ktorý umožní používateľom definovať ich vlastné vzory očakávaného správania, ktoré budú následne overené na databáze historických údajov vo formáte logov. Momentálne máme k dispozícii prvý prototyp aplikácie, ktorý bude ďalej vylepšovaný s cieľom poskytnúť prostredníctvom jednoduchého používateľského prostredia možnosť definovať komplexnejšie vzory.

POĎAKOVANIE

Táto práca bola vytvorená s podporou projektu VEGA č. 1/0042/10 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] Paralič, Ján: *Objavovanie znalostí v databázach*. Košice: TU-FEI, 2003. 88 s.
- [2] FAYYAD, Usama – PIATETSKY-SHAPIO, Gregory - SMYTH, Padhraic: *From Data Mining to Knowledge Discovery in Databases: American Association for Artificial Intelligence*. 1996.
- [3] PARALIČ, Ján - BABIČ, František: *New A virtual environment for support and evaluation of various knowledge practices based on principles of trialogical learning*. Košice. 2009.

Vytvorenie B2B aplikácie postavenej na platforme .NET

Marek FURIMSKÝ, František BABIČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

marek.furimsky@student.tuke.sk, frantisek.babic@tuke.sk

Abstrakt — Tento článok sa zaoberá vytvorením B2B aplikácie postavenej na platforme .NET 4.0. Využíva biznis procesy, ktoré spĺňajú štandard UBL 2.0, aj keď pokrýva iba niektoré z procesov rozsiahleho štandardu UBL 2.0. Tieto biznis procesy sú následne transformované na biznis procesy technológie Windows Workflow Foundation 4 a implementované prostredníctvom webových služieb. Ako klient webových služieb je použitá Silverlight aplikácia, ktorá tvorí zároveň užívateľské rozhranie. Takto navrhnuté riešenie umožňuje jednoduché ovládanie a zároveň podporuje tzv. long-running biznis procesy. Cieľom aplikácie je vytvoriť základ pre jednoduchý elektronický trh.

Kľúčové slová — B2B, .NET, webové služby, elektronický trh

I. ÚVOD

Súčasný globalizačný trend tlačí firmy v konkurenčnom prostredí k lepšiemu využívaniu svojho potenciálu, znižovaniu nákladov a hľadaniu nových pridaných hodnôt, ktoré jej pomôžu získať konkurenčnú výhodu. Jedným z faktorov, ktorý firmy považovali za rozhodujúcu výhodu pri svojom podnikaní, sa stalo využívanie informačných a komunikačných technológií na podporu obchodných aktivít. Tento proces vyústil do rozvoja elektronického obchodovania.

II. ELEKTRONICKÉ OBCHODOVANIE

A. Elektronické obchodovanie

Elektronické obchodovanie (e-commerce [1]) je nakupovanie a predávanie výrobkov alebo služieb cez elektronické systémy ako napr. internet alebo podobné počítačové siete. Týmto systémami sa uskutočňuje veľké množstvo obchodov, pri ktorých sa používajú elektronické formy prevodov finančných prostriedkov, internetový marketing, on-line spracovanie transakcií, elektronická výmena dát (EDI), systémy manažmentu zásob a automatizovaný zber dát. Súčasne elektronické obchodovanie zvyčajne využíva World Wide Web aspoň v jednej z fáz životného cyklu transakcie.

Pojem transakčný mechanizmus označuje akýkoľvek spôsob, ktorým je možné obchodovať. Je to dôležitý faktor úspechu využívania informačných a komunikačných technológií. Vhodne zvolený mechanizmus môže firme napomôcť pri maximalizácii zisku, či minimalizácii vstupov.

B. B2B

B2B¹ je všeobecné označenie pre internetové transakcie a služby, ktoré umožňujú firmám nakupovať a predávať tovar a služby on-line iným firmám. Tento typ elektronického obchodu využíva internet pre zjednodušenie komunikácie v obchodnom reťazci. Pri B2B sa spravidla jedná o dlhodobejší obchodný vzťah, ktorý je často zmluvne podchytený. Nejedná sa teda o klasické nakupovanie, ale o uzatváranie kontraktov medzi firmami.

C. Elektronický trh

V súčasnosti sa hlavne v B2B obchodovaní vytvára nový model, ktorý je možné prirovnať k obchodovaniu s akciami na burzách. Tento model sa označuje ako elektronický trh (e-marketplace, e-exchange). Vzniká prepájaním veľkého množstva nákupcov a predajcov a

¹ <http://en.wikipedia.org/wiki/Business-to-business>

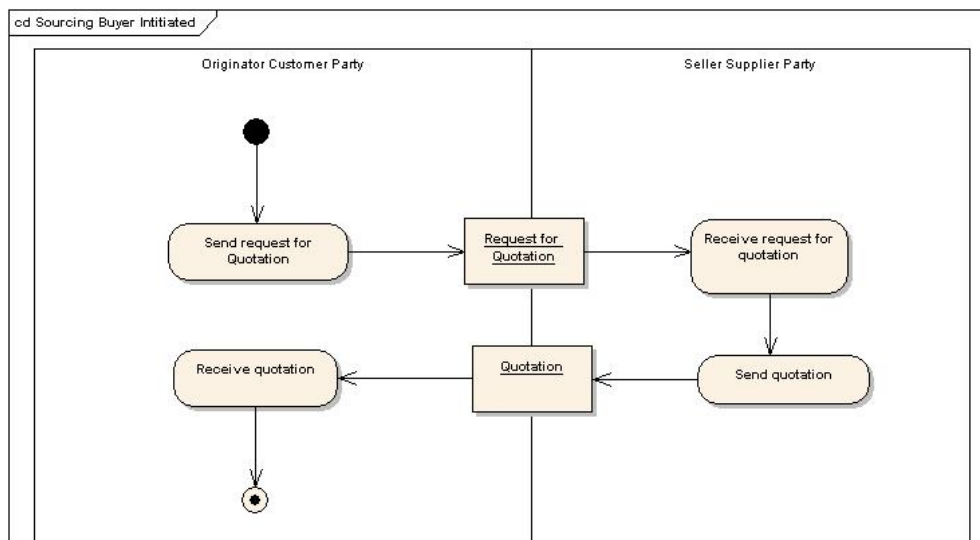
automatizovaním transakcií medzi nimi. Elektronické trhy rozširujú možnosti nákupcov, predajcom uľahčujú prístup k novým klientom a znižujú transakčné náklady pre všetkých zúčastnených. Zároveň zvyšujú transparentnosť a kvalitu informácií. Vlastník elektronického trhu môže spoplatňovať každú transakciu, prípadne môže zaviesť poplatky formou predplatného na určité obdobie. Navyše sú elektronické trhy softvérového charakteru a teda je pomerne jednoduché implementovať vylepšenia bez veľkých dodatočných investícií.

III. UBL

UBL (Universal Business Language) je knižnica štandardných elektronických biznis dokumentov, ako napr. objednávky, faktúry a pod. Tento jazyk bol založený konzorciom OASIS (Organization for the Advancement of Structured Information Standards²) za spolupráce s mnohými ďalšími organizáciami na štandardizáciu dát. Aktuálna verzia nesie označenie UBL 2.0³ a bola oficiálne spustená v októbri 2006. Vlastníkom UBL ostalo konzorcium OASIS, ale tento jazyk je k dispozícii pre každého bez licenčných poplatkov. Všetky UBL 2.0 procesy sú popísané formou ako na obr. 1. UBL 2.0 v súčasnosti popisuje tieto biznis procesy:

A. Obstarávanie

Proces obstarávania je zvyčajne prvým procesom v rámci kúpy, resp. predaja tovaru. UBL popisuje rôzne formy obstarávania – poskytnutie elektronického katalógu, obstarávanie iniciované zákazníkom a tzv. punchout obstarávanie.



Obr. 1 Proces zákazníkom iniciovaného obstarávania v UBL 2.0

B. Objednávanie

Nasleduje spravidla po úspešnom procese obstarávania. Zabezpečuje najmä elektronickú výmenu objednávok medzi kupujúcim a predávajúcim. V tomto procese je ešte možná zmena ako i zrušenie celej objednávky.

C. Expedícia

Tento biznis proces popisuje transfer tovaru alebo služieb od dodávateľa k príjemcovi. Zabezpečuje výmenu elektronických expedičných správ. Tieto správy obsahujú informácie o expedícii, ktorá môže byť úplná alebo čiastočná. O úplnej expedícii hovoríme vtedy, ak transfer obsahuje všetky položky objednávky. Čiastočná expedícia zabezpečuje transfer niektorých položiek objednávky.

D. Fakturácia

Proces fakturácie je jedným zo základných procesov, pri ktorých dochádza k výmene elektronických dokumentov, v tomto prípade faktúr. Faktúra je vystavená predajcom tovaru, resp. poskytovateľom služieb po dodaní tovaru, resp. poskytnutí služby.

² <http://www.oasis-open.or>

³ <http://docs.oasis-open.org/ubl/cs-UBL-2.0/UBL-2.0.html>

E. Platba

Pri tomto procese dochádza k uhradeniu faktúry, najčastejšie elektronickým presunom finančných prostriedkov medzi účtami kupujúceho a predávajúceho. Tento proces je najčastejšie realizuje medzi účtovníckym oddelením kupujúceho a predávajúceho.

F. Doprava

Tento proces definuje objednávanie logistických služieb pri medzinárodných obchodoch. Predávajúci v ňom informuje prepravcu, že tovar, ktorý má byť doručený kupujúcemu, je k dispozícii a môže byť transportovaný. Prepravca zasa informuje kupujúceho o termíne dodávky.

G. Kontrola originality tovaru

Tento proces zabezpečuje elektronickú výmenu certifikátu originality, ktorý deklaruje, že daný tovar má skutočne daný pôvod. Tento certifikát vystavuje exportér a po potvrdení príslušnými úradmi ho dostáva importér tovaru.

H. Informácie o stave zásielky

Tento proces zabezpečuje elektronickú výmenu reportov, ktoré vystavuje prepravca, aby informoval kupujúceho o stave zásielky. Report obsahuje zvyčajne informácie o aktuálnej pozícii zásielky, prípadne o termíne dodania.

V aplikácii nie sú implementované všetky tieto biznis procesy. Nakoľko cieľom aplikácie bolo vytvoriť základ pre malý elektronický trh, vybrané boli iba základné procesy a to proces obstarávania, objednávanie a fakturácie.

IV. POUŽITÉ TECHNOLOGIE

A. .NET

.NET⁴ je robustná vývojová platforma, ktorá umožňuje vytváranie rôznych druhov aplikácií pre operačné systémy Windows – desktopové, webové, serverové, mobilné, či cloud aplikácie. Aktuálna verzia .NET 4.0 prináša so sebou množstvo novinek medzi inými i Windows Workflow Foundation 4⁵, Windows Communication Foundation 4⁶, Silverlight 4⁷ a pod., ktoré tvoria základ celej aplikácie.

B. C#

Programovací jazyk C#⁸ bol vytvorený spoločnosťou Microsoft, aby skĺbil robustnosť C++ s jednoduchosťou Javy. Výsledkom bol objektovo orientovaný programovací jazyk, ktorý patrí medzi základné programovacie jazyky pre platformu .NET. Aktuálna verzia C# 4.0 bola vydaná 12. Apríla 2010.

C. Windows Workflow Foundation

Základný framework pre tvorbu biznis procesov na platforme .NET. Obsahuje množstvo preddefinovaných aktivít, ktoré môžu byť rozšírené o nové aktivity vytvorené vývojárom. Do aktuálnej verzie Windows Workflow Foundation 4 bol pridaný úplne nový prvok - Workflow Service, ktorý umožňuje integrovať webové služby do biznis procesov. Táto integrácia zabezpečuje správne poradie volania tej ktorej webovej služby, no zároveň umožňuje medzi jednotlivé volania pridať aktivity, ktoré sú vykonávané na strane servera bez toho, aby ich musel spúšťať klient, čo v celkovom meradle znižuje objem sieťovej komunikácie. Windows Workflow Foundation 4 prinieslo zároveň i ďalšiu novinku a to je Persistence. Persistence je schopnosť uložiť aktuálny stav procesu do databázy po stanovenom limite nečinnosti. V prípade, že sa vyžaduje pokračovanie procesu, tento proces je z databázy obnovený do pôvodného stavu a môže bez problémov pokračovať. Týmto spôsobom je zabezpečené, že aj pri tzv. long-running procesoch nie sú zbytočne zaťažované výpočtové kapacity systémov.

⁴ <http://www.microsoft.com/net/>

⁵ <http://msdn.microsoft.com/en-us/netframework/aa663328>

⁶ <http://msdn.microsoft.com/en-us/netframework/aa663324>

⁷ <http://www.silverlight.net/getstarted/silverlight-4/>

⁸ <http://msdn.microsoft.com/en-us/vcsharp/aa336809>

D. Windows Communication Foundation

Základný framework pre tvorbu službovo - orientovaných aplikácií. Umožňuje vytvárať aplikácie, ktoré medzi sebou komunikujú prostredníctvom webových služieb. WCF podporuje SOAP, REST, WS-* a i. Aktuálna verzia má označenie Windows Communication Foundation 4 a okrem iných vlastností je lepšie integrovaná s Windows Workflow Foundation 4.

E. ASP.NET MVC

Jedna⁹ z novších technológií od firmy Microsoft určená pre vývoj webových aplikácií. Už podľa názvu je zrejme že jej základom je návrhový vzor MVC, ktorý od seba oddeľuje zobrazovanie, užívateľskú interakciu a biznis model. Tento vzor má svoje nesporné výhody, ktoré sú veľmi cenné najmä v prípade rozširovania aplikácie a jej testovania. Aktuálna verzia je ASP.NET MVC 3 a priniesla nový view engine – RAZOR, ako i lepšiu kooperáciu s JQUERY¹⁰ a integrovanú podporu pre JSON¹¹ formát.

F. Silverlight

Vývojová platforma pre tvorbu interaktívnych webových aplikácií. Distribuuje sa pomocou pluginu, ktorý sa nainštaluje do web prehliadača podobne ako konkurenčný Flash. Umožňuje tvorbu tzv. RIA aplikácií (Rich Internet Applications¹²) – webové aplikácie, ktoré majú v mnohých smeroch rovnaké vlastnosti ako klasické desktopové, ale majú obmedzený prístup k lokálnym súborom a hardvéru počítača. Aktuálna verzia má označenie Silverlight 4. Istý čas po jej vydaní sa objavil PivotViewer komponent, ktorý prináša nový pohľad na prácu s obrovským množstvom dát vo vizuálnej podobe. V aplikácii bol použitý ako tzv. visual search, ale môže sa uplatniť napr. i v business intelligence, či rôznych výskumoch. Nakoľko však tento komponent nemá ešte ani rok, má stále svoje slabé miesta. Tie by sa však mali výrazne zredukovať s príchodom Silverlight 5, ktorý bol ohlásený na rok 2011.

V. ŠTRUKTÚRA RIEŠENIA

Finálne riešenie B2B aplikácie sa skladá z niekoľkých projektov, pričom každý projekt sa zameriava na určitú časť implementácie. Medzi základné projekty patria:

A. E-MARKET.CLIENT

Tento projekt je vlastne Silverlight aplikácia, ktorá tvorí základ pre používateľa systému. Predstavuje užívateľské rozhranie a zároveň klienta webových služieb a biznis procesov. Základom je PivotViewer¹³ komponent, ktorý slúži ako tzv. visual search pre vyhľadanie požadovaného výrobku. Vľavo je niekoľko filtrov, ktoré po zakliknutí zredukujú počet vyhovujúcich záznamov, a tým aj počet obrázkov vpravo. Zároveň majú obrázky vždy taký rozmer, aby vyplnili čo najviac prideleného priestoru. Visual search je zobrazený na obr. 2. Po nájdení požadovaného výrobku sa zobrazia aktuálne ponuky na tento výrobok. Následne začína proces objednávaní, kde je možné ešte dohodnúť podmienky ako cena, množstvo a pod.

B. E-MARKET.CLIENT.WEB

Je to ASP.NET MVC web aplikácia, ktorej základnou úlohou je hosťovať Silverlight aplikáciu, vytvorenú v projekte E-Market.Client. Ďalšími úlohami, ktoré tento projekt zabezpečuje sú funkcie prihlasovania a registrácie do systému.

C. E-MARKET.CONTRACTS

Projekt obsahuje všetky triedy používané v systéme. Každá trieda je anotovaná príslušným príznakom, aby bolo možné prenášať objekty danej triedy cez webové služby, prípadne správne validovať danú vlastnosť objektu.

D. E-MARKET.DATA

Tento projekt sa zameriava na vytvorenie mapovania tabuliek databázy na triedy v jazyku C#. Tento postup sa nazýva ORM (Object Relational Mapping) a v tomto prípade bola na mapovanie použitá technológia ADO.NET Entity Framework 4.0.

⁹ <http://www.asp.net/mvc>

¹⁰ <http://jquery.com/>

¹¹ <http://www.json.org/>

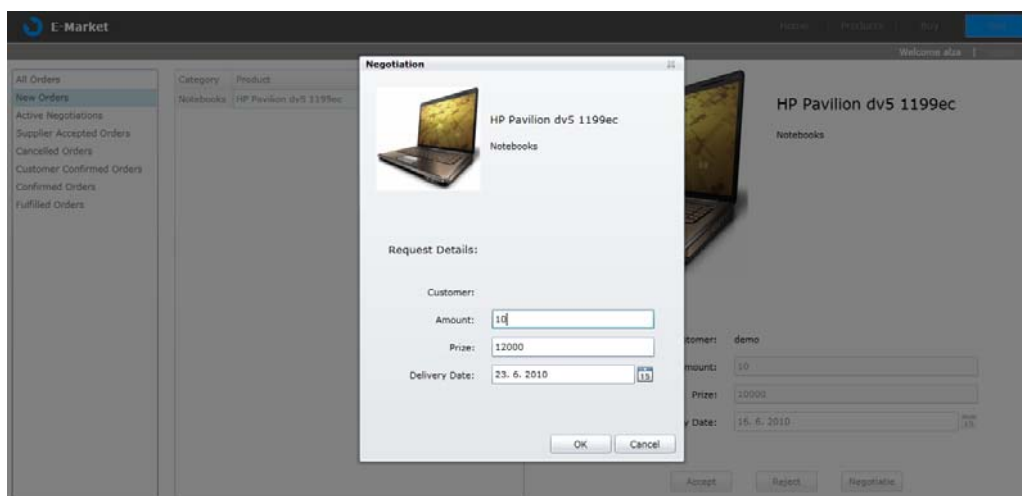
¹² http://en.wikipedia.org/wiki/Rich_Internet_application

¹³ <http://www.microsoft.com/silverlight/pivotviewer/>



Obr. 2 PivotViewer v Silverlight aplikácii

Táto aplikácia reprezentuje malý elektronický trh, ktorý umožní výmenu tovarov medzi firmami, keďže ide o model B2B. Hlavné užívateľské rozhranie je rozdelené na dve časti, kde prvá obsahuje galériu tovarov, ktoré sú momentálne k dispozícii a môžu byť predmetom obchodu. V našom prípade je naplnená testovacími vzorkami. V ľavej časti je možné aplikovať rôzne filtre, ktoré umožnia používateľovi jednoduchú orientáciu a výber na základe jeho požiadaviek alebo očakávaní. Po výbere tovaru nasleduje proces objednávania s možnosťou negociácie, ktorý je zobrazený na obr. 3.



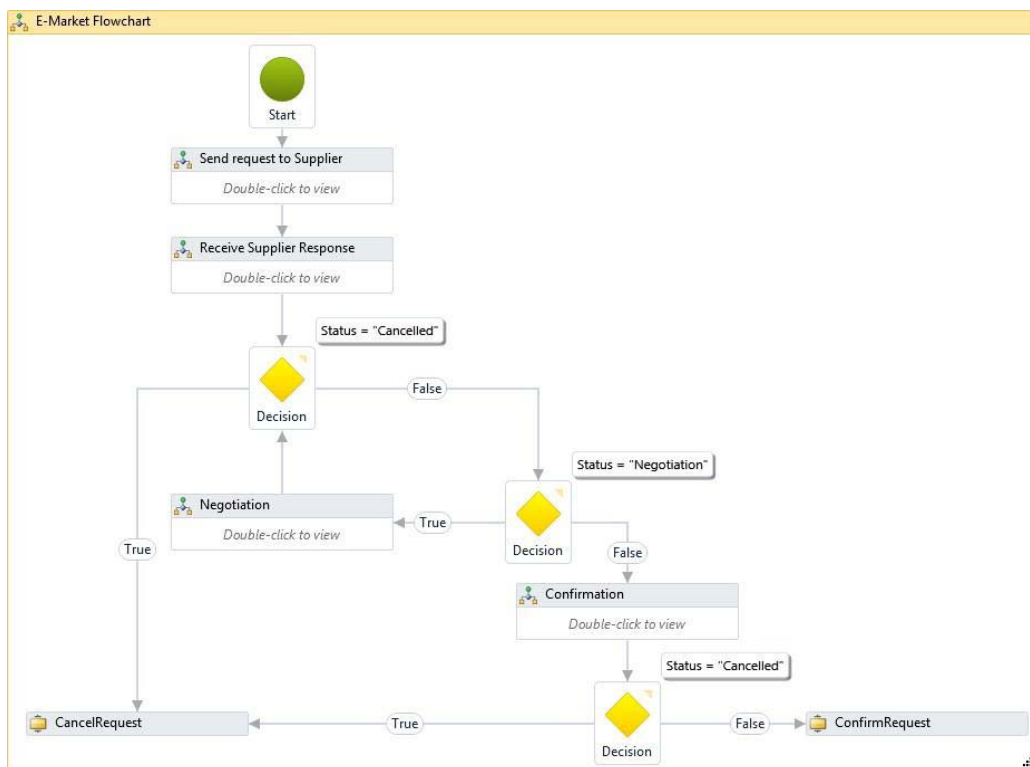
Obr. 3 Negociácia v rámci procesu objednávania

E. E-MARKET.WORKFLOWS

Tento projekt obsahuje všetky biznis procesy, ktoré sú používané v rámci aplikácie. Tieto biznis procesy sú následne volané ako webové služby Silverlight aplikáciou. Momentálne je implementovaný len proces objednávania, ktorý je zobrazený na obr. 4. Tento proces je prvým z UBL 2.0 procesov implementovaných do Windows Workflow Foundation 4 v tejto aplikácii. Ostatné procesy budú implementované v blízkej budúcnosti.

F. E-MARKET.ACTIVITIES

Je projekt, ktorý dopĺňa projekt E-Market.Workflows a poskytuje mu také aktivity pre biznis procesy, ktoré nie sú preddefinované vo Windows Workflow Foundation 4, a teda museli byť naprogramované od základov. Jedná sa najmä o aktivity súvisiace s prístupom do databázy.



Obr. 4 Proces objednávaní vo Windows Workflow Foundation 4

VI. ZÁVER

Cieľom aplikácie bolo, aby sa stala základom pre jednoduchý elektronický trh. Pre splnenie tohto cieľa je nutné dokončiť implementáciu ostatných biznis procesov v čo najkratšom čase. Aplikácia však už teraz poslúžila ako výborný testovací nástroj pri implementácii UBL 2.0 procesov do Windows Workflow Foundation 4.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] DELINA, Radoslav – VAJDA, Viliam: Teória a prax elektronického obchodovania. Košice : TU-EKF, 2008. 172s. ISBN 978-80-969953-3-2

Predikcia predajnosti na knižnom trhu prostredníctvom vybraných metód objavovania znalostí

Ludmila LIZÁKOVÁ, František BABIČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

ludmila.lizakova@student.tuke.sk, frantisek.babic@tuke.sk

Abstrakt — Cieľom tohto článku je predstaviť postup pri riešení predikčnej úlohy prostredníctvom vhodných metód dolovania znalostí z databázy. Obchodným cieľom je predpoved' predajnosti vybraných knižných titulov na nemeckom knižnom trhu na základe zozbieraných historických dát popisujúcich celkovú predajnosť knižných titulov na tomto trhu. Na praktické experimenty bol využitý softvérový nástroj SPSS Clementine, jednotlivé kroky boli realizované v súlade s metodológiou CRISP-DM.

Kľúčové slová — data, dolovanie znalostí, predikcia, lineárna regresia

I. ÚVOD

Knižný trh v Nemecku je jedným z najväčších trhov v súčasnosti. Pre predajcov je výzva kúpiť správny titul na sklad a ponúkať ho vo vhodnom množstve a čase. Ročne sa v Nemecku vydá cez 96 000 titulov. Prvoradé je teda vybrať správne tituly a ponúkať ich čo najlacnejšie. V spolupráci s najväčším nemeckým predajcom kníh sa podarilo zmerať a uložiť veľké množstvo dát na základe, ktorého je možné vytvoriť predpoveď predajnosti kníh.

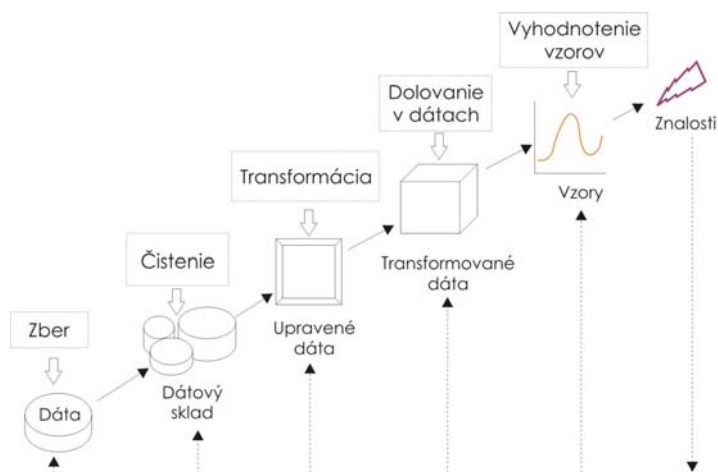
Cieľom tejto práce je predpovedať predajnosť 8 knižných titulov na 2418 miestach. Táto predikcia bude realizovaná prostredníctvom modelu, ktorý bude vytvorený na základe predajnosti v rámci 2394 lokalít, ktoré poznáme. Všetky dáta pochádzajú z jedného časového úseku a úlohou je predpovedať predajnosť tak presne ako to bude možné.

II. DOLOVANIE ZNALOSTÍ Z DATABÁZ

Pojem Dolovanie znalostí z databáz (v preklade Knowledge Discovery in Databases, skr. KDD) zatiaľ nie je jednoznačne definovaný, rôzni autori ho popisujú rôzne a rovnako aj v prekladoch dochádza k značnému miešaniu pojmov. Podľa SAS inštitútu je KDD proces výberu, prehľadávania a modelovania veľkých objemov dát na objavenie doteraz neznámych vzorov vhodných pre podnikateľské účely. Podľa iného zdroja: KDD je proces semi-automatické extrakcie znalostí z databáz. Znalosti musia byť štatisticky platné, doposiaľ neznáme a potenciálne užitočné. (Samotné dolovanie v dátach väčšina autorov považuje len za časť procesu KDD) [1]. Pojmovo jasná je aj definícia podľa Fayyad a kolektívu, kde sa uvádza, že dolovanie znalostí v databáze je interaktívny a iteratívny proces s niekoľkými krokmi a dolovanie dát je časťou tohto procesu. Samotný proces KDD definujú ako: netriviálny proces rozpoznávania platných, nových (neobvyklých), potenciálne užitočných a jednoznačne zrozumiteľných vzorov z dát [1].

Proces KDD je tvorený 6 podprocesmi (viď. obr. 1):

- čistenie dát – odstránenie „hľuku a nekonzistentností
- integrácia – spájanie dátových zdrojov (kroky 1. a 2. sa v niektorých literatúrach uvádzajú ako podprocesy, ktorých výsledkom je vytvorenie data warehous-u)
- výber – dáta relevantné k cieľu skúmania sú vybraté z databázy
- transformácia – úprava dát do podoby vyžadovanej algoritmami alebo nástrojmi
- dolovanie – fáza, v ktorej sú aplikované samotné algoritmy dolovania znalostí
- vyhodnotenie vzorov – identifikácia užitočných vzorov, ktoré reprezentujú znalosti
- prezentácia znalostí – aplikácia získaných znalostí do praxe



Obr. 1 Proces KDD a jeho jednotlivé fázy

Pri riešení problémovej úlohy budeme postupovať podľa metodiky CRISP-DM.

A. CRISP-DM

Skratka CRISP-DM (Cross Industry Standard Process For Data Mining) predstavuje štandardizovaný proces pre objavovanie znalostí, ktorý je použiteľný pre rôzne typy úloh. Popisuje postup spracovania projektu objavovania znalostí v dátach, príklad spracovania komerčného zadania projektu od zákazníka. Uľahčuje riešiteľom vypracovanie projektu a pomáha im vyhnúť sa problémom pri vypracovaní. Metodika popisuje jednotlivé fázy životného cyklu projektu. Postupnosť fáz nie je definitívne určená, dokonca je žiaduce, aby sa po niektorých fázach vracalo naspäť do predchádzajúcich a došlo ku korektúram.

Pochopenie biznis cieľa: pre správne riešenie zadanej úlohy (projektu) je potrebné vykonať analýzu čo klient chce samotným projektom dosiahnuť, čo chce získať. Dôkladné preskúmanie podnikateľského pozadia a prostredia v ktorom bude projekt spracovaný. Zákazník má často veľa doplňujúcich a aj protichodných požiadaviek či obmedzení. Cieľom tejto fázy je všetky tieto požiadavky zjednotiť, vytvoriť logickú a dosiahnuteľnú štruktúru cieľov. Pochopenie biznis cieľov neznamená to isté čo ciele KDD. Preto je dobré na základe podnikateľských požiadaviek sformulovať aj dokument s cieľmi KDD, podľa, ktorého budeme pracovať. Na záver tejto fázy je pripravený podrobný projektový plán.

Pochopenie dát: pozostáva zo štyroch hlavných činností od zberu samotných dát až po ich prvú analýzu. Zbieranie dát a ich ukladanie do požadovaných formátov uľahčí ďalšiu prácu a prinesie to rýchlejšie porozumenie. Vytvoríme popisnú správu dát, skladá sa zo základných údajov, veľkosť, počet dát, počet tabuliek. Tento dokument by mal priniesť odpoveď na otázku či dáta vyhovujú základným požiadavkám projektu. Overovanie kvality dát v tejto fáze znamená, či sú dáta úplne, či neobsahujú chyby. Ak existuje problém s kvalitou, je potrebné spísať možné riešenia.

Predspracovanie: táto fáza projektu je časovo najnáročnejšia. Obsahuje operácie ako výber dát: vyhodnotenie ktoré dáta budú používané ďalej, v závislosti od cieľov projektu, kvality a technologických obmedzení; čistenie dát: má za úlohu zvýšiť kvalitu používaných dát. Ide o odstránenie chýbajúcich a výrazne odchylených hodnôt, identifikovanie prázdnych hodnôt; vytváranie dát: využitie rôznych metód, ktoré majú priniesť zlepšenie výsledku, môžu tu vzniknúť aj nové vlastnosti (atribúty), ide o derivácie, transformáciu pôvodných atribútov, atď.; integrácia: spájanie dát z rôznych zdrojov. Je nevyhnutná ak sú dáta zbierané do rôznych tabuliek a databáz. Tu je potrebné dať pozor na vznik redundancií; formátovanie dát: úprava, formálna zmena dát, nemení sa ich význam, len vonkajší zjav.

Modelovanie: výber konkrétnej metódy na modelovanie. Predtým ako vygenerujeme model si navrhujeme prístup, pomocou ktorého budeme merať kvalitu a pravdivosť dosiahnutých výsledkov. Najčastejšie používame delenie dát na dve množiny, trénovaciu a testovaciu. Pri budovaní modelu sa používa trénovacia množina. Reálne sa v tejto fáze nevytvorí jeden ale viacero modelov a postupne sa porovnáva ich kvalita a parametre.

Vyhodnotenie: neanalyzuje presnosť modelu ako nástroja ale či sa mu podarilo splniť zadanie biznis cieľov. Sumarizujeme prehľady a hlavné postupy, či nebolo niečo zabudnuté alebo

prehliadnuté, ak je to nutné je potrebné niektoré časti opakovať. Na záver tejto fázy zadáme kroky ako budeme ďalej postupovať. Musíme vedieť kedy ukončiť predchádzajúce fázy a opakovania, kedy sa posunúť do poslednej fázy.

Nasadenie: plán nasadenia znamená nasadenie výsledkov, zostavenie stratégie pre zavedenie výsledku projektu. Zhrnie stratégiu nasadenia a plán ako to dosiahnuť. Fáza nasadenia vyžaduje vytvorenie plánu na kontrolu a údržbu. Dôkladne pripravený plán na údržbu systému pomôže predchádzať nesprávnemu používaniu získaných znalostí. Záverom projektu je vytvorenie finálnej správy. Ide o zhrnutie projektu, môže to byť aj prezentácia o získaných vedomostiach. Často sa na záver projektu organizuje stretnutie kde sa výsledky predstavia zákazníkom, diskutuje sa. [1]

III. POSTUP PRI RIEŠENÍ ÚLOHY

A. Pochopenie biznis cieľa

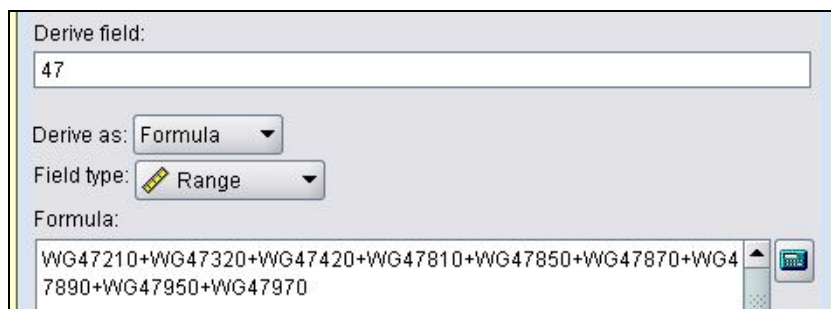
Knižný trh v Nemecku je jedným z najväčších trhov v súčasnosti. Pre predajcov je výzva kúpiť správny titul na sklad a vo vhodnom čase a množstve ho ponúknuť na predaj. Ročne sa v Nemecku vydá cez 96 000 titulov. Cieľom je vybrať správne tituly a ponúkať ich čo najlacnejšie. To je biznis cieľ našej úlohy, ktorý je možno pretransformovať do oblasti KDD, pôjde o predikciu hodnoty numerického atribútu, kde sa vhodnou metódou javí lineárna regresia.

B. Pochopenie dát

V tejto fáze ide o viaceré činnosti: zbieranie dát a ich ukladanie do požadovaných, overovanie kvality dát (či sú dáta úplne, či neobsahujú chyby) Ak existuje problém s kvalitou je potrebné spísať možné riešenia. Samotné dáta pre riešenie projektu boli už zozbierané a uložené vo formáte txt. Množina train, na ktorej budeme vytvárať model obsahuje 2418 riadkov a 1856 stĺpcov. Riadky znamenajú počet lokalít, kde sa jednotlivé knihy predávali, stĺpce sú atribúty. Atribúty v našom prípade sú rôzne typy kníh. Číselné označenie je zvolené preto lebo na určenie knihy nie je zvolený autor ale charakteristika knihy. Označenie vlastne vytvára hierarchiu konceptov. Na najvyššej pozícii je určenie vonkajších vlastností knihy. Zvyšné 4 číslce určujú začlenenie knihy do žánrovej oblasti. Názov atribútu potom vyzerá napr. takto WG18230. Cieľové atribúty sú označené ako T1...T8.

C. Predspracovanie

Táto fáza procesu dolovania znalostí je časovo najnáročnejšia. V rámci nej boli realizované výber dát: vyhodnotenie, ktoré dáta budú používané ďalej, v závislosti od cieľov projektu, kvality a technologických obmedzení - použijeme všetkých 2418 riadkov. Čistenie dát: má za úlohu zvýšiť kvalitu používaných dát. Ide o odstránenie chýbajúcich a výrazne odchýlených hodnôt, identifikovanie prázdnych hodnôt. Dáta neobsahujú prázdne hodnoty. Sú kompletné. Nulové hodnoty neodstraňujeme, pretože nesú dôležitú informáciu a o knihách, ktoré sa vôbec nepredávajú. Vytváranie dát: využitie rôznych metód, ktoré majú priniesť zlepšenie výsledku, môžu tu vznikať aj nové vlastnosti (atribúty). Ide o derivácie, transformáciu pôvodných atribútov atď. Nové atribúty vytvárame derivovaním z pôvodných. Keďže atribútov v úlohe je 1856 je dôležité ich správne pochopenie. Každý atribút popisuje jednu knihu. Číselné označenie je zvolené preto lebo na určenie knihy nie je zvolený autor ale charakteristika knihy. Označenie vlastne vytvára hierarchiu konceptov. Na najvyššej pozícii je určenie vonkajších vlastností knihy. Zvyšné 4 číslce určujú začlenenie knihy do žánrovej oblasti.

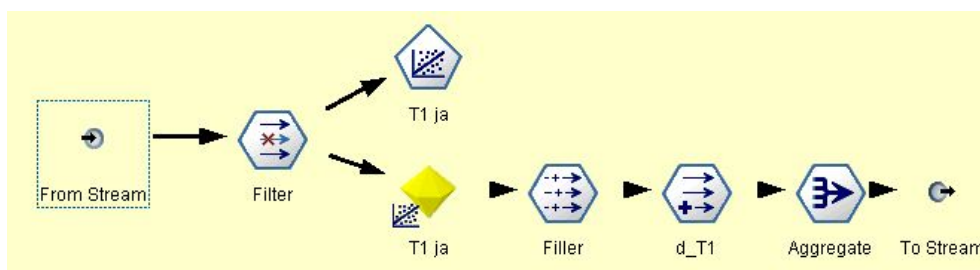


Obr. 2 Príklad vytvorenia nového atribútu v SPSS Clementine postupom na vyššiu úroveň v hierarchii konceptov

Tieto atribúty potrebujeme generalizovať, aby sme zredukovali ich počet. Nové atribúty vznikli spojením pôvodných podľa hierarchie konceptov, ktorá je vytvorená popisnými číslami v názvoch. Generalizáciou nám vzniklo 95 atribútov. Označených číslami od 00 až po 88. Na ukončenie generalizácie som používala priemernú hodnotu predajnosti v novom atribúte v intervale $<0,2000>$. V prostredí Clementine sme vytvorili osem uzlov Super Node, pre každú skupinu nových atribútov. SuperNode obsahuje uzly DERIVE so súčtovými výrazmi. Takýto uzol s novým atribútom a jeho zloženie vidíme na obrázku 2.

D. Modelovanie

Ako model sme si vybrali model lineárnej regresie (*regression*), metódu Forward. Regresia začína s najjednoduchším modelom bez vstupných polí a postupne prechádza polia. Každé nové pole je testované o koľko by zlepšili model a tie najlepšie sú k modelu pridané. Keď už žiadne ďalšie polia nie sú alebo najlepšie pole neprodukuje dostatočné zlepšenie vtedy dôjde k samotnému vytvoreniu modelu. Keďže máme 8 výstupných hodnôt, je teda potrebné naučiť 8 samostatných modelov, vid' obr.3.



Obr. 3 Príklad implementácie algoritmu lineárna regresia v prostredí SPSS Clementine

E. Vyhodnotenie

Kvalita výsledných predpovedí je daná sumou, (absolútnou hodnotou) všetkých rozdielov medzi predpovedanou a skutočnou hodnotou.

p_{ij} - predpovedaná suma i-tého titulu na j-tej lokalite

r_{ij} - skutočná suma i-tého titulu na j-tej lokalite

d - kvalita

$$d = \sum_{i=1}^8 \sum_{j=1}^{2418} |p_{ij} - r_{ij}|$$

Tento vzťah sme použili na zistenie presnosti vytvorených modelov, nazýva sa chybová funkcia. Výsledky chybových funkcií pre jednotlivé modely vidíme v nasledujúcej tabuľke. Celkové predajnosti jednotlivých kníh (T1 až T8) sú v státisícoch, takže chyby niekoľko tisíc kusov sú relatívne presnými predpoveďami:

Tabuľka 1
 KVALITA PREDIKCIE JEDNOTLIVÝCH MODELOV PRE PRÍSLUŠNÉ KNIŽNÉ TITULY

Knižný titul	Kvalita
T1	1646,43
T2	3251,91
T3	1155,49
T4	1429,14
T5	6813,22
T6	12203,52
T7	2183,19
T8	1615,89

Táto úloha bola riešená aj v rámci súťaže Data Mining Cup 2009¹. Cieľom tejto súťaže je podporiť mladých výskumníkov a študentov v oblasti objavovania znalostí prostredníctvom práce na reálnych dátach z rôznych domén. Každoročným favoritom sú tímy študentov z nemeckých univerzít ako Karlsruhe alebo Dortmund. V tomto roku víťaz dosiahol celkovú chybu 17 260, náš výsledok 30 295 by zodpovedal umiestneniu okolo 30. miesta, čo nie je zlý výsledok. Každopádne existuje ešte priestor na potenciálne zlepšenie vytvorených modelov, ktoré bude predmetom ďalšej práce. Napr. aplikácia iných algoritmov predikcie s vhodným

¹ <http://www.data-mining-cup.de/en/review/dmc-2009/>

nastavením parametrov s čím súvisia aj potenciálne zmeny vo formáte dát a predspracovaní; využitie iných výpočtových nástrojov pre oblasť dolovania znalostí, ktoré boli získané len nedávno (SAS² alebo Statistica³); aplikácia zmien v rámci vytvorenej hierarchie konceptov, ktorá by lepšie zodpovedala štruktúre knižného trhu.

F. Nasadenie

V našom prípade by nasadenie vytvorených modelov prinieslo poznanie o predpokladanej predajnosti každej novej knihy, ktorá je k dispozícii na trhu, čo by umožnilo predajcovi posúdiť rentabilitu jej predaja. Tieto modely by bolo nutné kontinuálne vyvíjať a upravovať vzhľadom na meniacu sa situáciu na trhu, aby boli schopné reagovať na nové podmienky.

IV. ZÁVER

Pri samotnom spracovaní projektu dolovania znalostí je vždy dôležité sledovať hlavne obsah, význam dát a biznis cieľ, nie len slepo zostavovať modely. Práve táto schopnosť je kľúčová pre úspešné ukončenie projektu. Riešenie takýchto úloh je v súčasnosti nevyhnutné. Firmy musia optimalizovať svoje náklady aby dosahovali zisky, alebo dokonca aby neklesli do straty. Samotný proces KDD prináša odpoveď na viacero typov problémov, preto znalosť problematiky KDD je do budúcnosti veľmi perspektívna zručnosťou.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] PARALIČ, Ján : Objavovanie znalostí v databázach. Košice : TUKE-FEI, 2003. 80 s. ISBN 80-89066-60-7
- [2] Fayyad, U.M. - Piatetsky-Shapiro, G. - Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery & Data Mining, Fayyad, U.M. - Piatetsky-Shapiro, G. - Smyth, P. - Uthurusamy, R., Eds. AAAI/MIT Press, Cambridge, Massachusetts, 1996.
- [3] www.twocrows.com: Introduction to Data Mining and Knowledge discovery. Two Crows Corporation. Potomac, 2005.
- [4] Chen, M.S. - Han, J. - Yu, P.S.: Data Mining: An Overview from a Database Perspective. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Dec. 1996, Vol.8, No.6, pp.866-883.
- [5] Marbán, Ó – Mariscal, G. – Segovia, J.: A Data Mining Knowledge discovery process model: Universidad Politécnica de Madrid, 2009. 16 s. ISBN 978-3-902613-53-0
- [6] Chapman, P. a kol. CRISP-DM 1.0 [online]. Version 1.0. Dostupné na internete: <<http://www.crispdm.org/>>

² <http://www.sas.com/technologies/analytics/index.html>

³ <http://www.statsoft.com/#>

Vytvorenie modelu pre predikciu úspešnosti internetových aukcií

Vladimír GAŠPAR, Adrián DRINGUŠ, František BABIČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

vladimir.gaspar@student.tuke.sk, adrian.dringus@student.tuke.sk,
frantisek.babic@tuke.sk

Abstrakt — Tento článok popisuje riešenie problému predikcie úspešnosti internetových aukcií prostredníctvom vhodných metód objavovania znalostí. Cieľom je vytvoriť klasifikačný model, ktorý bude rozdeľovať novo zriadené internetové aukcie do dvoch kategórií: aukcia so ziskom vyšším ako priemer alebo aukcia so ziskom nižším ako priemer. Tento model bol vytvorený na základe historických dát, ktoré reprezentujú už ukončené internetové aukcie na portály Ebay.com. Celý proces bol navrhnutý a realizovaný na základe metodológie CRISP-DM, ktorá predstavuje všeobecne používaný štandard pre proces dolovania znalostí. Jednotlivé experimenty boli uskutočnené v prostredí SPSS Clementine.

Kľúčové slová — internetová aukcia, zisk, klasifikácia, rozhodovacie stromy

I. ÚVOD

V dnešnej dobe sa medzi používateľmi internetu stále viac rozširuje používanie aukčných platforiem ako Ebay¹, či Amazon, atď. Svoje služby poskytujú súkromným aj právnickým osobám. Ich spoločnou črtou je dosiahnuť čo najvyššiu predajnú cenu. Otázka teda znie: „Ako dosiahnuť optimálnu (maximálnu) predajnú cenu?“ Mnohí experti na internetové aukcie radia rôzne postupy, dlhšie/kratšie trvanie aukcie, rozličné počiatočné ceny, prerušenie aukcií počas víkendov, či pracovných dní a mnohé iné varianty, ktoré majú zaručiť úspech. Jednou z možností je využiť vhodné metódy objavovania znalostí na vytvorenie modelu pomocou zhromaždených historických dát, ktorý bude ďalej použitý na klasifikáciu nových prípadov (elektronických aukcií) do vopred pripravených kategórií: zisk vyšší ako priemer, zisk nižší ako priemer.

Historické dáta popisujú internetové aukcie predajcu elektroniky prostredníctvom portálu Ebay.com. Tento predajca zaznamenal, že jeho aukcie dosahujú dosť odlišné tržby, a snažil sa nájsť nejaké typické vzory spôsobujúce túto situáciu. Tento prístup však nepriniesol úspech, preto sa rozhodol pre využitie metód objavovania znalostí. S cieľom maximalizácie zisku má byť vytvorený model popisujúci optimálne nastavenie aukcie (čas začiatku a konca, doba trvania, ďalšie atribúty, atď.), ktorý bude predpovedať, či skutočné tržby z predaja budú vyššie ako priemer z predaja výrobkov danej kategórie.

Existuje niekoľko metodológií pre implementáciu celého procesu objavovania znalostí, my sme sa rozhodli pre CRISP-DM², ktorý definuje tento proces ako životný cyklus zložený zo 6 základných fáz, ktoré v sebe ďalej ukrývajú jednotlivé kroky a operácie.

II. CRISP-DM

A. Pochopenie cieľa

Na základe trénovacej množiny (historické dáta, je jasné či daná aukcia bola zisková alebo nie) je potrebné navrhnuť model na klasifikáciu internetových aukcií do dvoch kategórií: či daná aukcia dosiahne vyššiu alebo nižšiu hodnotu ako priemerná cena v danej kategórii, do ktorej prislúšná položka aukcie patrí. Tento model bude použitý nové aukcie (aukcie nie sú zadelené do tried), kde cieľom je správne predpovedať čo najväčší počet aukcií do jednotlivých tried.

¹ <http://www.ebay.com/>

B. Pochopenie dát

Pre tento účel boli k dispozícii dáta popisujúce jeho aukcie na portály Ebay.com za niekoľko posledných mesiacov. Konkrétne ide o vzorku 8 000 online aukcií z kategórie "Audio&Hi-Fi:MP3-Player:Apple iPod". Dôležité údaje o týchto aukciách sú uložené v súbore train.txt, ktorý predstavuje trénovaciu množinu. V ňom, okrem iných atribútov, nájdeme atribúty category_avg_gms, ktorý obsahuje priemernú hodnotu predajných cien pre každú z kategórií iPod-ov a tiež atribút gms_greater_avg, ktorý indikuje, či dosiahnutá cena bola vyššia ako priemer v danej kategórii. Tržby nad priemer označujeme ako vysoká cena a tržby nižšie ako priemer označujeme ako nízka cena. K dispozícii je ďalej súbor class.txt, ktorý obsahuje údaje o nových aukciách a bude slúžiť na otestovanie modelu. Cieľovým atribútom je gms_greater_avg, ktorý určuje či sú tržby z danej aukcie vyššie ako priemer. Popis jednotlivých atribútov:

- auct_id (celé číslo) – ID priradené aukcii
- item_leaf_category_name – produktová kategória daného tovaru
- listing_title – názov alebo titulok aukcie - tento textový atribút obsahuje popis aukcie, ktorý má zaujať kupujúceho. Ide teda o najvhodnejšie dáta pre vytvorenie modelu alebo zvyšovanie jeho presnosti. Keďže obsahuje mnohokrát rovnaké slová, je veľmi vhodný na odvádzanie nových atribútov.
- listing_subtitle – podtitulok aukcie - podobne ako predchádzajúci textový atribút, môže byť nápomocný pre opis daného tovaru. Problémom je však jeho výskyt, keďže je definovaný len pre menší počet aukcií.
- listing_start_date – Dátum začiatku aukcie vo formáte MM/DD/RRRR
- listing_end_date – Dátum ukončenia aukcie vo formáte MM/DD/RRRR
- listing_durtn_days (celé číslo) – trvanie aukcie v dňoch
- listing_type_code (1,2,7,9) – typ aukcie (napr. normálna, multiaukcia, atď.)
- feedback_score_at_listing_time (celé číslo) – skóre predávajúceho zo spätnej väzby nakupujúcich v čase prezerania aukcie
- start_price (číslo) – vyvolávacia cena aukcie v €
- buy_it_now_price (číslo) – cena pre okamžitú kúpu v €
- buy_it_now_listed_flag (T/F) – aukcia spustená s voľbou okamžitej kúpy (binárny atribút) - pri tejto možnosti sa dá tovar v aukcii okamžite kúpiť za cenu uvedenú v atribúte buy_it_now_price
- bold_fee_flag (T/F) – aukcia uvedená tučným písmom (binárny atribút)
- featured_fee_flag (T/F) – aukcia uvedená ako top ponuka na domovskej stránke portálu (binárny atribút)
- category_featured_flag (T/F) – aukcia uvedená ako najlepšia ponuka v kategórii (binárny atribút)
- gallery_fee_flag (T/F) – aukcia uvedená s obrázkom (binárny atribút)
- gallery_featured_fee_flag (T/F) – aukcia uvedená s možnosťou náhľadu obrázka (binárny atribút)
- ipix_featured_fee_flag (T/F) – aukcia uvedená s ipix (binárny atribút) - balíčky obrázkov produktu (dodatočné, xxl, pic.show, pack)
- reserve_fee_flag (T/F) – aukcia uvedená s minimálnou cenou (binárny atribút)
- highlight_fee_flag (T/F) – aukcia uvedená s farebným pozadím (binárny atribút)
- schedule_fee_flag (T/F) – aukcia načasovaná dopredu a spustená automaticky
- border_fee_flag (T/F) – aukcia zadaná s hranicou (binárny atribút)
- qty_available_per_listing (celé číslo) – množstvo ponúkaného tovaru v multiaukcii
- gms (reálne číslo) – dosiahnutá predajná cena v € - tento atribút je smerodajný pri určovaní vysokej alebo nízkej ceny. Podľa neho je totiž určený atribút gms_greater_avg.
- category_avg_gms (reálne číslo) – priemerné tržby v € danej produktovej kategórie
- gms_greater_avg (1/0) – 0 ak je predajná cena nižšia ako priemer produktovej kategórie a 1 ak je predajná cena vyššia ako priemer produktovej kategórie

Ďalším krokom bolo načítanie trénovacej množiny dát do nástroja SPSS Clementine³ (momentálne vyvíjaný a predávaný pod obchodnou značkou IBM SPSS Modeler), v ktorom je

² <http://www.crisp-dm.org/>

³ <http://www.spss.com/software/modeler/>

možné realizovať jednotlivé kroky procesu dolovania znalostí. Na popis atribútov je možné využiť funkcie Data audit, Quality (viď. obr. 1) alebo Statistics. Zamerali sme sa na úplnosť dát, čiže na prázdné, nezadané, prípadne chybné zadané hodnoty atribútov a na pomer v akom je rozdelený cieľový atribút `gms_greater_avg`. Po preskúmaní oboch spomínaných záležitostí, sme boli pozitívne prekvapení, keďže všetky atribúty okrem `gms` vykazovali 100% kompletnosť a 8 000 správne zadaných údajov. Atribút `gms` vykazoval 98.525% kompletnosť, keďže 112 položiek v ňom nebolo zadaných. Neskôr sme však zistili, že to nebolo podstatné, lebo tento parameter do analýzy zahrnutý nebol, tak aj kvôli tomu, sme sa týmito prázdnyimi hodnotami ďalej nezaoberali. Cieľový atribút `gms_greater_avg` obsahoval hodnoty 1 a 0 vo vyhovujúcom pomere 47,51% ku 52,49 %.

C. Predspracovanie

Predspracovanie predstavuje najdôležitejšiu fázu v celom procese dolovania znalostí, keďže pripravuje dáta pre vybrané algoritmy. V tomto prípade sme použili viacero prístupov na dosiahnutie kvalitnej vzorky dát.

Na úvod sme odfiltrovali všetky atribúty typu flag, ktoré mali v názve `fee_flag`. K tomuto kroku sme dospeli preskúmaním dát. Ich rozloženie medzi aukciami v súbore `train.txt` tak ako aj v súbore `class.txt` je veľmi podobné (takmer na 100%) a nie len to. Ich rozloženie medzi aukciami s vysokými cenami a aukciami s nízkymi cenami je takisto skoro totožné. Rozdiely sú v rozmedzí 0,002% až 1,01%. V streame však filtrovanie uskutočňujeme v jednom kroku na konci pred samotným algoritmom. Transformácia atribútov priniesla zmiešané výsledky. Zistili sme, že odvodenie mesiaca nám znižuje úspešnosť, no odvodenie nového atribútu `weekday` nám zvýšilo úspešnosť modelu o 0,6%. Potom prišiel na rad binning pre atribút `start_price` ktorý sme rozdelili na 10 rovnako veľkých intervalov nahradených rovnakou hodnotou. Takto sme dosiahli vyhladenie atributu `start_price`. A nakoniec sme sa pustili do derivovania textových atribútov. Skúšali sme slová vyskytujúce sa najčastejšie v atribúte „`listing_title`“. Vytvorili sme nový atribút `kapacita_GB`, ktorý sme naplnili veľkosťou jednotlivých iPodov v GB.

Odvodili sme nový atribút vytvorením pomeru medzi atribútmi `buy_it_now_price` a `start_price`. Tento pomer v praxi predstavuje možnosti bidovania. Napríklad ak je `buy it now` cena 500€ a štartovacia cena je 50€, koeficient = 10. Takýmto spôsobom vieme povedať, aké možnosti stanovenia ceny dáva predávajúci kupujúcemu. Čím je tento koeficient nižší (bližšie k 1), tým je menší predpoklad, že bude aukcia úspešná. Ak teda je `buy_it_now_price` 500€ a štartovacia cena je 489€ = 1,02, je zrejmé, že predajca nedal veľký priestor na úpravu ceny pre kupujúceho.

Tento atribút je typu flag (binárny 0, 1) a popisuje či aukcia obsahuje popis – subtitle (ak áno tak 1 ak nie tak 0). Z pohľadu kupujúceho je popis veľmi dôležitý, keďže sa dozvedá o stave, veku, vlastnostiach tovaru a o výhodách, ktoré súvisia s kúpou tohto tovaru.

`Neu_neuwertig` je typu flag (binárny) a nadobúda 1 (true) ak `listing_title` obsahuje slová ako NEU (nový) alebo NEUWERTIG (nový tovar), ak tieto slová neobsahuje, tak nadobúda hodnotu 0 (false). Je jednou z hlavných súčastí popisu tovaru. Nehovorí o tom či bol tovar používaný ale popisuje stav tovaru. Podľa toho sa zákazník rozhoduje, či je cena adekvátna tovaru. Ak je cena za nový tovar rovnaká ako za mierne poškodený resp. s poškodeným balením atď., je väčšia pravdepodobnosť kúpi nového tovaru.

`ZUBEHOR` atribút je typu flag (binárny) a nadobúda 1 (true) ak `listing_title` obsahuje slovo ZUBEHOR (príslušenstvo), ak toto slovo neobsahuje tak nadobúda hodnotu 0 (false). Tovar môže byť aj nový ale ak existuje aj iná aukcia, kde sa predáva tovar s príslušenstvom, je väčšia šanca predaja zariadenia s príslušenstvom. O jednotlivých kombináciách nového/použitého/poškodeného VS s príslušenstvom/bez príslušenstva by sa dalo polemizovať.

`OVP` Atribút je typu flag (binárny) a nadobúda 1 (true) ak `listing_title` obsahuje slovo OVP alebo ORIGINALVERPACKUNG alebo ORIGINALVERPACKT čo znamená originálne balenie, ak toto slovo neobsahuje, tak nadobúda hodnotu 0 (false). Originálne balenie, znamená vo veľkej väčšine prípadov kompletný tovar s príslušenstvom, bez toho aby bol tovar používaný resp. vybalený.

`DEFEKT` Atribút je typu flag (binárny) a nadobúda 1 (true) ak `listing_title` obsahuje slovo DEFEKT (poškodený) alebo slovo GEBRAUCHT (použitý), ak neobsahuje ani jedno z týchto slov tak nadobúda hodnotu 0 (false). Aj keď tieto slová nie sú synonymá, spojili sme ich lebo samostatne by tvorili veľmi malé kategórie. Existuje všeobecný predpoklad, že použité alebo poškodené zariadenia by sa mali pohybovať v nižších cenových reláciách. Samozrejme cena závisí aj od vážnosti poškodenia resp. opotrebenia pri používaní. Poškodený tovar teda prináša nižšie tržby.

GARANTIE atribút je typu flag (binárny) a nadobúda 1 (true) ak listing_title obsahuje slovo GARANTIE (záruka), ak toto slovo neobsahuje tak nadobúda hodnotu 0 (false). Platnosť záruky, je veľkou výhodou v prípade predávajúceho pretože vie potvrdiť vek tovaru. Podobne aj kupujúci má tendenciu kúpiť tovar v záruke, pre prípad skrytej chyby a možnosti vrátiť tovar ak by došlo k jeho poškodeniu.

RECHNUNG atribút je typu flag (binárny) a nadobúda 1 (true) ak listing_title obsahuje slovo RECHNUNG (faktúra), ak toto slovo neobsahuje tak nadobúda hodnotu 0 (false). Dôležitý faktor pre kupujúceho pretože vie porovnať cenu, za ktorú kúpil tovar predávajúci v určitom čase s cenou za ktorú je tovar ponúknutý na predaj.

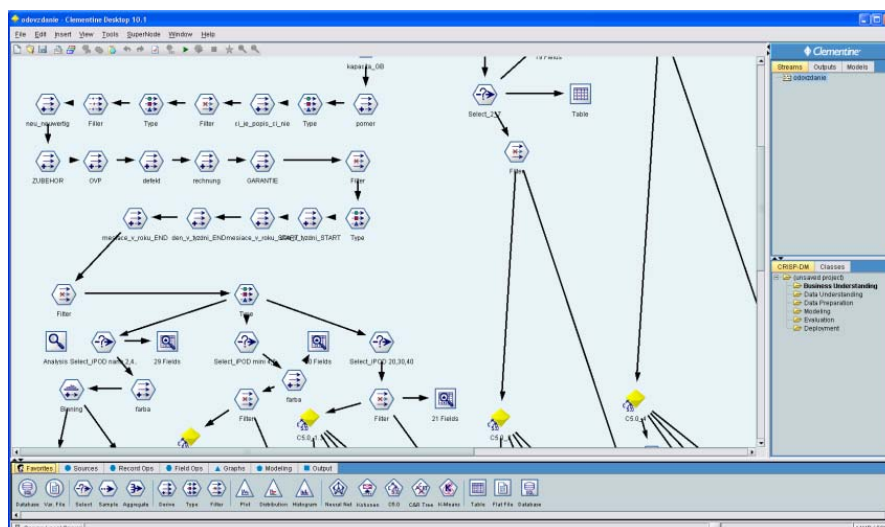
Okrem týchto atribútov sme ešte odvodili nové atribúty z dátumov:

- den_v_tyzdni_START - tento atribút nadobúda hodnoty od 1 (nedeľa) do 7(sobota). Vychádza z atribútu listing_start_date. Pre aukcie je začiatkový deň dôležitý z pohľadu pracovného dňa alebo voľného dňa resp. dňa blízko víkendu (piatok).
- mesiace_v_roku_START - tento atribút nadobúda hodnoty 1 (január) až 12 (december). Vychádza z atribútu listing_start_date. Pri každom predaji akéhokoľvek tovaru existuje predpoklad sezónnosti. Veľký význam majú Vianoce (12 mesiac). Predpokladom je teda začiatok aukcie PRED Vianocami.
- den_v_tyzdni_END - tento atribút nadobúda hodnoty od 1 (nedeľa) do 7(sobota). Vychádza z atribútu listing_end_date. Podobne ako pre START je deň konca aukcie podstatný z hľadiska pracovného dňa alebo dňa pracovného voľna resp. dňa blízko víkendu (piatok)
- mesiace_v_roku_END - tento atribút nadobúda hodnoty 1 (január) až 12 (december). Vychádza z atribútu listing_end_date. Podobne ako START je podstatné aj mesiac ukončenia aukcie.

D. Modelovanie

V našom prípade sme sa v prostredí SPSS rozhodli pre algoritmus C5, kde sme použili krížovú validáciu a metódu boosting pre zvýšenie presnosti klasifikátora. Krížová validácia znamená, že záznamy sa rozdelia do n rovnako veľkých podmnožín. Stále sa použije n-1 podmnožín na tréning a zvyšná podmnožina na testovanie. Toto sa opakuje n- krát, čím získame n rôznych presností klasifikátora. Výsledná presnosť sa určí ako aritmetický priemer n klasifikátorov. My sme použili typickú 10-násobnú krížovú validáciu presnosti klasifikátorov.

Boosting znamená, že každému príkladu v tréningovej množine je priradená určitá váha. Potom sa trénuje séria klasifikátorov K_n ($n = 1, 2, \dots, N$), takým spôsobom, že po vygenerovaní každého klasifikátora K_n sa upravujú váhy príkladov v tréningovej množine tak, aby sa pri učení nasledujúceho klasifikátora K_{n+1} venovala zvýšená pozornosť predtým chybné klasifikovaným príkladom. Výsledný zložený klasifikátor K kombinuje hlasy jednotlivých klasifikátorov tak, že váha hlasu každého z klasifikátorov je priamo úmerná jeho presnosti.



Obr. 1 Časť implementovaného procesu dolovania v prostredí SPSS

V tomto prípade sme použili 30 alebo 35 násobný boosting, čím sme dosiahli zvýšenie presnosti klasifikátora. Pre každú vetvu sme vytvorili osobitný model. V nastaveniach modelu

C5 sme použili buď základné nastavenia alebo pruning (orezávanie), ak mali pravidlá v modeli nízku odhadovanú presnosť kvôli prílišnému rozvetveniu (orezávali sa tie vetvy, ktoré popisovali len veľmi malý počet príkladov tréningovej množiny).

E. Vyhodnotenie

Úspešnosť modelov sme hodnotili podľa počtu správne zatriedených príkladov z testovacej množiny. Následne sme tieto čiastkové výsledky spočítali a dostali sme celkový počet správne zatriedených príkladov. Ten sme predelili číslom 8000 (počet príkladov v množine) a vynásobili číslom 100, aby sme dostali percentuálnu úspešnosť. Priemerná úspešnosť vytvoreného modelu sa pohybovala okolo 74%. Tento výsledok je možné porovnať s dosiahnutými hodnotami v rámci súťaže Data mining Cup (DMC) 2006⁴, ktorej cieľom je poskytnúť možnosť pracovať na reálnom probléme využitím metód a výpočtových prostriedkov procesu objavovania znalostí. V tomto prípade boli výsledky hodnotené na základe cenovej matice, ktorá ohodnocuje výsledky klasifikácie:

Tabuľka 1
CENOVÁ MATICA

	Dosiahnutý vysoký zisk	Dosiahnutý nízky zisk
Predpovedaný zisk aukcie ako vysoký	1	-1
Predpovedaný zisk aukcie ako nízky	-1	1

Cieľom je dosiahnuť maximálny počet bodov, čo znamená vysokú presnosť vytvoreného klasifikačného modelu. V našom prípade sme sa pohybovali okolo hodnoty 3 800 bodov, kde víťaz tejto edície DMC dosiahol 5 020 bodov. Porovnanie použitých prístupov ukazuje odlišnosti najmä z pohľadu výkonnosti, prístup je v oboch prípadoch porovnateľný. Napr. v oboch prípadoch boli extrahované príznaky/atribúty popisujúce jednotlivé predmety aukcie, viď. kapitola Predspracovanie (tento fakt bol aj základným stavebným kameňom úspešnosti, lebo len jednoduchá klasifikácia na základe dodaných atribútov nepriniesla požadovanú presnosť). Víťaz extrahoval viac ako 4 tis. kľúčových slov, z ktorých vytvoril binárne atribúty, čiže či popis príslušného predmetu obsahuje toto kľúčové slovo alebo nie. Ako vhodná metóda boli v oboch prípadoch použité rozhodovacie stromy, kde víťazný model obsahoval 1000 rozhodovacích stromov (naš model 4), avšak samotné modelovanie trvalo skoro celý deň. Tento prístup sme v minimálnej miere testovali aj v našom prípade, kde už 8 rozhodovacích stromov prinieslo zlepšenie v počte bodov o 200. Tento progres predstavuje motiváciu v ďalšej práci s cieľom priblížiť sa dosiahnutým výsledkom víťaza ale s ohľadom na dostupné výpočtové kapacity (nami dosiahnutý výsledok bol vygenerovaný v reálnom čase).

III. ZÁVER

Dolovanie znalostí z databázach predstavuje dôležitú oblasť umelej inteligencie, ktorá nachádza uplatnenie v rôznych doménach. Jednou z nich je aj ekonomická sféra, kde hlavným cieľom je maximalizácia ziskov. Tento cieľ je možné dosiahnuť pomocou vhodne zvolených metód objavovania znalostí, ktoré nám umožnia napr. predikciu ziskovosti internetových aukcií, predikciu vývoja kurzu akcií alebo mien, identifikáciu zaujímavých položiek v rámci nákupov, identifikáciu zaujímavých kombinácií služieb pre zákazníkov, ale aj analýzu podvodného správania, podozrivých transakcií, atď. Tento článok ukazuje možný postup ako realizovať riešenie reálneho problému s využitím SPSS Clementine ako vhodného nástroje pre akademické ale aj komerčné účely v tejto doméne.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

⁴ <http://www.data-mining-cup.de/en/review/dmc-2006/>

Aplikácia DDE a OPC komunikačných protokolov v identifikácii a riadení reálneho systému

Štefan JAJČIŠIN, Anna JADLOVSKÁ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

stefan.jajcisin@tuke.sk, anna.jadlovska@tuke.sk

Abstrakt — Článok sa zaoberá opisom reálneho hydraulického systému z hľadiska technického a programového vybavenia, spôsobom jeho ovládania, ako aj komunikáciou s počítačom, na ktorom je spustený algoritmus identifikácie alebo riadenia v prostredí Matlab. V rámci spojenia reálneho systému s počítačom je bližšie rozobraný spôsob vytvorenia a realizácie komunikácie s využitím DDE a OPC protokolov. V rámci článku sú uvedené aj zjednodušené vývojové diagramy základných programov, ktoré sme vytvorili v prostredí Matlab za účelom systémovej identifikácie a riadenia reálneho hydraulického systému, ktoré využívajú DDE a OPC protokoly.

Kľúčové slová — komunikačný protokol DDE, komunikačný protokol OPC, PLC automat, reálny hydraulický systém, simulačný jazyk Matlab

I. ÚVOD

V článku sa zaoberáme dvoma rôznymi spôsobmi komunikácie reálneho systému s Matlabom, a to pomocou DDE a OPC protokolov.

V prvej časti sa venujeme opisu samotného reálneho systému. Uvádzame technické vybavenie systému a vzájomné prepojenia jednotlivých častí, ale aj programovú výbavu, ktorá je súčasťou reálneho systému. V ďalšej časti sa zaoberáme komunikáciou reálneho systému s počítačom, na ktorom je spustený identifikačný alebo riadiaci algoritmus vytvorený v prostredí Matlab. Zaoberáme sa základným postupom pri vytváraní a udržiavaní komunikácie prostredníctvom DDE a OPC protokolov. V rámci článku sa tiež venujeme implementácii vytvorených komunikačných programov pri identifikácii a riadení reálneho hydraulického systému.

Súčasťou článku je vývojový diagram programu pre zber potrebných dát pri identifikácii, ako aj vývojový diagram programu pri uskutočňovaní riadenia hydraulického systému pomocou PSD algoritmov.

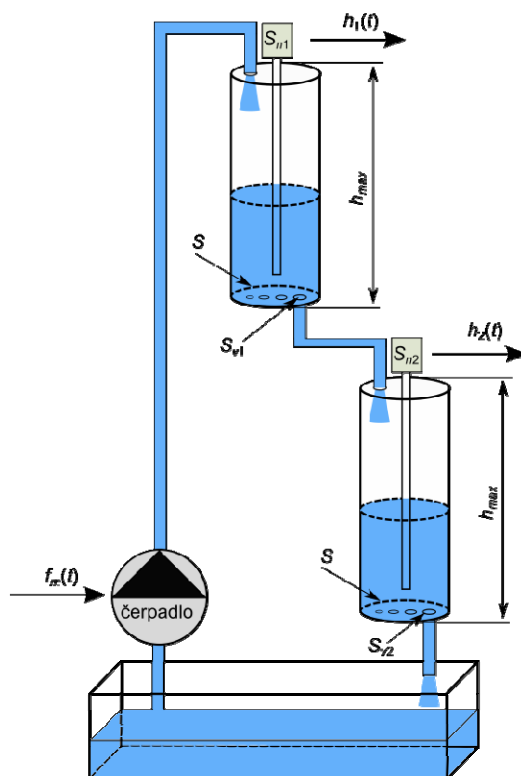
II. REÁLNY MODEL HYDRAULICKÉHO SYSTÉMU

Reálny model hydraulického systému, ktorý vlastní Katedra kybernetiky a umelej inteligencie je tvorený dvoma valcovými nádobami umiestnenými pod sebou. Kvapalina, v našom prípade voda, je do prvej nádoby privádzaná prostredníctvom čerpadla, ktoré ju vytláča z vaničky cez privodnú trubicu do vstupného otvoru prvej nádoby. Nádoby sú prepojené hadičkou, ktorou kvapalina vytekajúca z výstupného otvoru v prvej nádoby preteká cez vstupný otvor do druhej nádoby. Z výstupného otvoru druhej nádoby kvapalina vyteká pod vplyvom hydrostatického tlaku späť do vaničky. Veľkosť prierezu výstupných otvorov je možné vybrať zo štyroch rôznych hodnôt pomocou škrtiacich skrutiek.

Tento reálny systém predstavuje hydraulický systém dvoch nádob bez vzájomnej interakcie.

Schematické znázornenie fyzikálnej podstaty reálneho hydraulického systému je na Obr. 1, pričom jednotlivé symboly majú nasledujúci význam:

- S - prierez nádob,
- S_{v1}, S_{v2} - prierez výstupných ventilov prvej, resp. druhej nádoby,
- h_{max} - výška nádob (maximálna výška hladiny),
- S_{n1}, S_{n2} - snímače merajúce aktuálnu výšku hladiny v oboch nádobách,
- $f_m(t)$ - frekvencia motora čerpadla generovaná frekvenčným meničom,
- $h_1(t), h_2(t)$ - aktuálna výška hladiny kvapaliny v nádobách.



Obr. 1 Hydraulický systém – fyzikálny princíp

A. Technické vybavenie

Z hľadiska technického vybavenia sú, okrem už spomenutých častí, súčasťou tohto systému aj:

- membránové čerpadlo *Spring* [1];
- frekvenčný menič *Allen-Bradley PowerFlex 40* [2], ktorý na základe hodnoty vstupného napätia v rozsahu 0 – 10V mení frekvenciu napätia motora čerpadla v definovanom rozsahu;
- programovateľný logický automat (PLC) *Allen-Bradley Micrologix 1200* so vstupno-výstupnou analógovou kartou [3];
- kapacitné snímače *Dinel CLM-36* [4], ktoré sú umiestnené v oboch nádobách.

B. Programové vybavenie

Programové vybavenie, ktoré je súčasťou reálneho hydraulického systému je softvér v rámci frekvenčného meniča, ktorým sa nastavujú jednotlivé parametre, napríklad:

- doba nábehu motora na požadovanú frekvenciu,
- minimálna a maximálna frekvencia motora,
- parameter určujúci, ktorý zo vstupov frekvenčného meniča má byť aktívny (prúdový, napäťový, potenciometer, atď.).

Podrobné informácie o jednotlivých parametroch a ich nastavovaní sú uvedené v [2].

Za programové vybavenie reálneho systému sa dá tiež považovať aj program spustený v PLC, keďže naším cieľom je návrh riadiacich algoritmov, ktoré budú fungovať v Matlabe. Avšak program spustený v PLC, ktorý zabezpečuje zápis na výstupy PLC, načítavanie príslušných údajov zo vstupov PLC a ošetrenie havarijných stavov sme museli taktiež navrhnuť a naprogramovať pomocou rebríkových schém v prostredí RSLogix 500 [5].

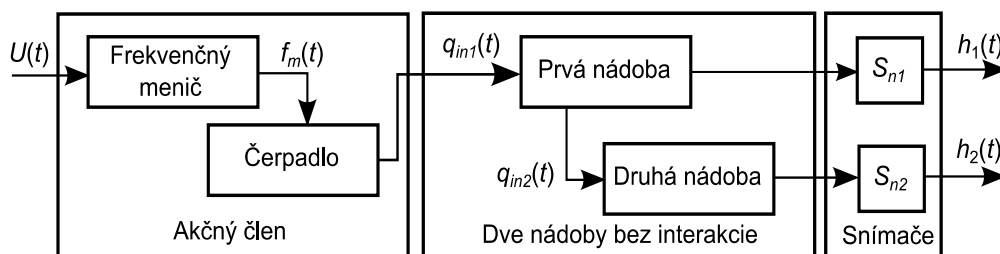
C. Princíp ovládania reálneho hydraulického systému

Z hľadiska systémového prístupu predstavuje frekvencia motora čerpadla $f_m(t)$ vstup do systému. Frekvenciu $f_m(t)$ je možné meniť rôznymi spôsobmi, napr. prúdovým signálom v rozsahu 4 – 20mA, napäťovým signálom v rozsahu 0 – 10V, pomocou potenciometra na paneli meniča alebo ďalšími inými spôsobmi. Ktorý z uvedených spôsobov sa má použiť je potrebné nastaviť vo frekvenčnom meniči. V našom prípade využívame napäťový vstup, na ktorý privádzame napätie $U(t)$ generované v rozsahu 0 – 10V PLC automatom prostredníctvom výstupov vstupno-výstupnej analógovej karty.

Dynamiku hydraulického systému predstavuje meniaci sa výška hladiny kvapaliny v oboch nádobách $\dot{h}_1(t)$ a $\dot{h}_2(t)$.

Výstupmi systému sú aktuálne výšky hladín kvapaliny v oboch nádobách $h_1(t)$ a $h_2(t)$, ktoré sú snímané kapacitnými snímačmi S_{n1} a S_{n2} . Údaje namerané snímačmi sú prostredníctvom vstupov analógovej karty spracovávané v PLC, t.j. preškálované a prostredníctvom komunikačného kanála odoslané do programu spustenom v Matlabe na osobitnom počítači. Po spracovaní týchto údajov je v Matlabe na základe príslušného algoritmu vygenerovaný vstup (akčný zásah) hydraulického systému, ktorý je opäť odoslaný do PLC, kde sa údaje znova preškálujú a odošlú prostredníctvom analógovej karty na vstup frekvenčného meniča.

Systémový rozbor hydraulického systému je na Obr. 2, kde okrem už spomenutých veličín je $q_{in1}(t)$, resp. $q_{in2}(t)$ prítok do prvej, resp. druhej nádoby.

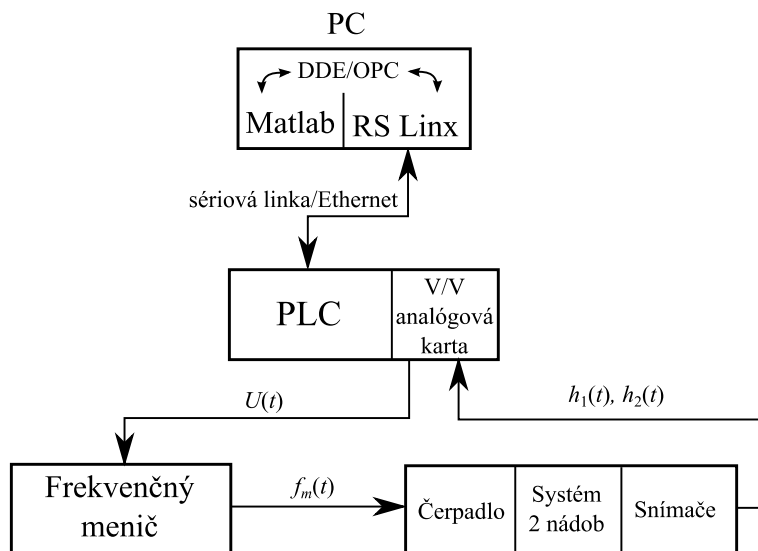


Obr. 2 Hydraulický systém – systémový pohľad s jedným vstupom a dvomi výstupmi

III. KOMUNIKAČNÉ PROTOKOLY DDE A OPC

Komunikáciu medzi PLC a počítačom zabezpečuje v našom prípade softvérový produkt firmy Rockwell Automation RSLinx [6], ktorý je potrebné do počítača vopred nainštalovať. RSLinx vykonáva úlohu DDE, resp. OPC servera, pričom DDE, resp. OPC klientom je Matlab. Takáto konštelácia nám umožňuje pristupovať k jednotlivým premenným v PLC pomocou DDE a OPC nástrojov priamo z Matlabu.

Na Obr. 3 je znázornená kompletná štruktúra reálneho hydraulického systému s pripojením na počítač. V ďalšom sa zameriame na stručný opis a spôsob uskutočnenia komunikácie medzi RSLinx-om a Matlabom na báze DDE a OPC protokolov.



Obr. 3 Kompletná štruktúra reálneho hydraulického systému s pripojením na počítač

A. DDE protokol

DDE (*Dynamic Data Exchange*) je komunikačný protokol založený na architektúre typu *klient - server*. Protokol bol vytvorený firmou *Microsoft* a umožňuje výmenu údajov medzi aplikáciami spustenými v operačnom systéme *Windows*. Aplikácia podporujúca DDE protokol môže slúžiť ako server pre viac klientov, klient pre viac serverov alebo byť zároveň klientom aj serverom.

Základné prvky, ktoré musia byť v procese komunikácie známe sú:

1. *Service* - meno aplikácie DDE servera (väčšinou názov aplikácie alebo služby),
2. *Topic* - téma rozhovoru (zvyčajne názov súboru, cesta),
3. *Item* - identifikátor údajov, napr. bunka (rozsah buniek) tabuľky, názov premennej.

Je potrebné zdôrazniť, že názvy pre *Service*, *Topic* a *Item* závisia od komunikujúcich aplikácií a ich presné označenie si je vždy potrebné overiť.

DDE protokol umožňuje uskutočniť tri typy komunikácie medzi DDE serverom a klientom:

1. *Hot link* - pri zmene hodnoty údajov dochádza k ich automatickému prenosu,
2. *Warm link* - pri zmene hodnoty údajov je najprv serverom odoslaná informácia o tom, že došlo k zmene, prenos údajov sa vykoná až na základe potvrdenia prenosu klientom,
3. *Cold link* - prenos údajov sa vykonáva iba na základe požiadavky klienta.

B. OPC protokol

OPC (*OLE for Process Control*) protokol je založený na COM (*Component Object Model*) technológii, kde komponenty, ako sú aplikácie, dátové objekty, ovládacie prvky, služby a pod. môžu navzájom komunikovať ako objekty. Taktiež má architektúru typu *klient – server*.

OPC je štandardný protokol pri komunikácii rôznych dátových zdrojov s klientskymi aplikáciami riadiacich systémov.

Na úrovni OPC servera sú dáta definované v skupinách (*Group*), ktoré obsahujú jednotlivé položky (*Item*). Položky, podobne ako v DDE komunikácii, predstavujú konkrétne bunky alebo premenné dátových zdrojov. Skupiny a ich položky môžu byť verejné alebo lokálne. Verejné môžu byť zdieľané viacerými klientmi, pričom lokálne sú určené iba pre konkrétneho klienta. V každej skupine musí byť definovaná minimálne jedna položka. Dôležité si je uvedomiť, že položky nie sú dátovými zdrojmi, iba ich reprezentujú, adresujú.

Okrem toho, OPC špecifikácia dovoľuje uskutočňovať komunikáciu aj na základe vzniknutých udalostí alebo výnimiek, nielen v presne stanovených intervaloch [7].

IV. APLIKÁCIA KOMUNIKAČNÝCH PROTOKOLOV V IDENTIFIKÁCII A RIADENÍ REÁLNEHO SYSTÉMU

V tejto časti bližšie popíšeme postup vytvorenia a uskutočňovania komunikácie medzi RSLinx-om (server) a Matlabom (klient). V oboch prípadoch, či už s použitím DDE alebo OPC protokolu, pozostáva komunikácia z troch základných krokov:

1. vytvorenie komunikačného kanála, resp. klienta,
2. čítanie/zápis údajov cez komunikačný kanál,
3. zrušenie komunikačného kanála.

V RSLinx-e nie je potrebné nastavovať žiadne parametre, automaticky plní úlohu DDE, resp. OPC servera. Za zmienku snáď stojí možnosť overenia aktívnych tém komunikácie (*topic-ov*) prostredníctvom položky *DDE/OPC* v hlavnom menu okna RSLinx-u. Čo sa týka nastavení Matlabu, uvádzame syntax príkazov, ktoré umožňujú realizovať vyššie uvedené kroky. V prípade OPC komunikácie je potrebné mať nainštalovaný *OPC Toolbox* v prostredí Matlab.

A. DDE komunikácia

1. Vytvorenie komunikačného kanála

```
service = 'rslinx';  
topic = 'riadenie'; (Topic musí byť zhodný s názvom súboru, programu, ktorý je nahraný  
v PLC, teda s ktorým je nadviazaná komunikácia.)  
kanal = ddeinit(service,topic);
```

2. Načítanie hodnoty premennej PLC na adrese F8:1 do premennej Matlabu h2

```
h2_adr = 'F8:1';  
h2 = ddereq(kanal,h2_adr);
```

Zápis hodnoty premennej Matlabu u do premennej PLC na adrese F8:2

```
u_adr = 'F8:2';  
u = 5;  
b = ddpoke(kanal,u_adr,u); (Premenná b nadobudne hodnotu 0/1 podľa toho, či bol zápis  
neúspešný/úspešný.)
```

3. Zrušenie komunikačného kanála

```
b = ddeterm(kanal);
```

B. OPC komunikácia

1. Vytvorenie OPC klienta s príslušnými skupinami pre vstupy, výstupy a ich položkami

```
opc_client = opcda('localhost', 'RSLinx OPC Server');  
gInputs = addgroup(opc_client, 'Inputs');  
gOutputs = addgroup(opc_client, 'Outputs');
```

```
itemI = additem(gInputs,{'[riadenie]F8:2'}); (V hranatých zátvorkách je názov súboru,  

    programu, ktorý je nahraný v PLC, teda s ktorým je nadviazaná komunikácia.)  

itemO = additem(gOutputs,'[riadenie]F8:0','[riadenie]F8:1');
```

Okrem zadefinovania skupín a ich položiek je potrebné nastaviť aj ich dátové typy, napr.
`set(itemO(1), 'DataType', 'single');`
 Taktiež je možné nastaviť aj rôzne parametre skupín a ich položiek. Okrem iného napr.
`set(gInputs, 'UpdateRate', 1);`
`set(itemO(1), 'Tag', 'h1');`
 Detailnejší popis a možnosti nastavenia jednotlivých parametrov skupín a ich položiek je
 uvedený v [8].

Nadviazanie komunikácie

```
connect(opc_client);
```

2. Načítanie hodnoty premennej PLC na adrese F8:1 do premennej Matlabu h2

```
h2_struct = read(itemO(2));
```

```
h2 = h2_struct.Value;
```

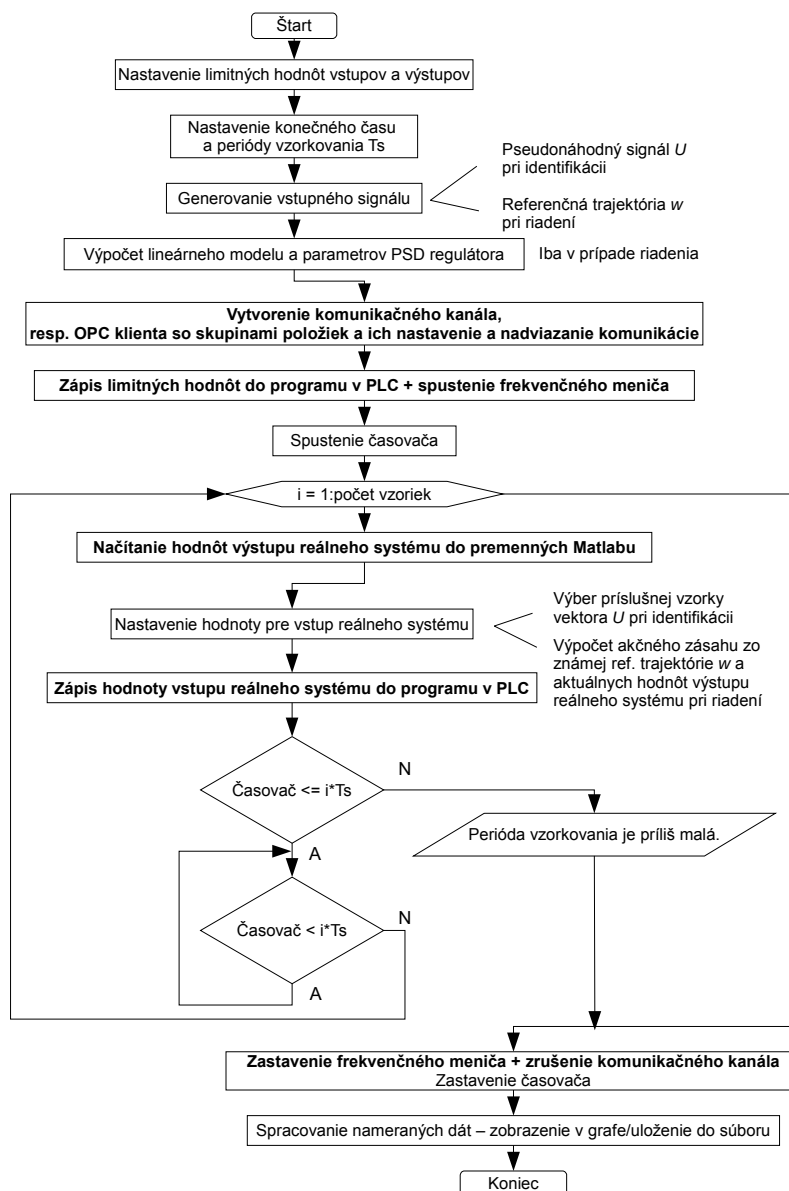
Zápis hodnoty premennej Matlabu u do premennej PLC na adrese F8:2

```
u = 5;
```

```
write(itemI(1),u);
```

3. Zrušenie komunikačného kanála

```
disconnect(opc_client);
```



Obr. 4 Vývojový diagram aplikácie DDE a OPC komunikačných protokolov v identifikácii a riadení reálneho systému

Uvedené príkazy jazyka Matlab sme aplikovali na všetky premenné programu v PLC, ktoré sú potrebné pre správnu funkčnosť reálneho hydraulického systému. Okrem vstupného napätia U a výstupných hodnôt h_1 a h_2 sú to aj dva bity nastavujúce povolenie a štart frekvenčného meniča, ako aj hodnoty maximálnej výšky hladiny v oboch nádobách.

Príkazy realizujúce komunikáciu sme potom implementovali do algoritmov identifikácie a riadenia reálneho systému v súlade s vývojovým diagramom na Obr. 4. Takto doplnené algoritmy umožňujú posilať na vstup reálneho systému hodnotu napätia akčného zásahu, či už ako vstupný signál v prípade identifikácie alebo akčný zásah generovaný riadiacim algoritmom. Zároveň sú v rámci algoritmov načítavané hodnoty výstupných veličín hydraulického systému, ktoré v prípade riadiaceho algoritmu slúžia pre výpočet akčného zásahu v ďalšom kroku.

V. ZHODNOTENIE

V článku sme prezentovali reálny hydraulický systém z pohľadu fyzikálnej štruktúry, technického a programového vybavenia, ktorý vlastní Katedra kybernetiky a umelej inteligencie. Uviedli sme stručný opis DDE a OPC komunikačných protokolov, ako aj ich použitie v komunikácii medzi počítačom a PLC ovládajúcim reálny hydraulický systém v rámci algoritmu identifikácie a riadenia reálneho systému.

Hoci realizácia komunikácie na báze DDE protokolu je jednoduchšia, v podstate vyžaduje znalosť a použitie štyroch jednoduchých príkazov jazyka Matlab, OPC toolbox predstavuje výhodnejší komunikačný prostriedok. Zložitejšie vytváranie OPC klienta a definovanie skupín s príslušnými položkami je zdanlivo nevýhoda, avšak v rámci OPC toolboxu existuje grafické používateľské prostredie uľahčujúce túto procedúru. Navyše tento prístup ponúka oveľa viac možností ako nastaviť vlastností jednotlivých položiek a komunikácie, okrem iného možnosť reagovať na udalosti a definované výnimky v rámci komunikácie. OPC toolbox Taktiež poskytuje nástroje pre kompletné vytvorenie a realizáciu komunikácie v prostredí Simulink, čo umožňuje vyhnúť sa písaniu akéhokoľvek zdrojového kódu v jazyku Matlab. Ďalšou výhodou OPC protokolu je jeho otvorenosť a fakt, že predstavuje medzinárodný štandard v komunikácii riadiacich systémov.

Z hľadiska časovej náročnosti počas prijímania a odosielania dát boli v našom prípade oba protokoly vyhovujúce. Je však potrebné zároveň podotknúť, že pri identifikácii a riadení reálneho hydraulického systému úplne postačuje perióda vzorkovania 1s. Zaujímavé by bolo porovnať tieto dva protokoly pri nasadení na systémy s rýchlo sa meniacou dynamikou, kde perióda vzorkovania je rádovo ms.

POĎAKOVANIE

Tento článok bol vytvorený realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Bia-Verder, *SPRING membránové provedení*, technická dokumentace. Dostupné na internete: http://bia-verder.verder.ams3 footsteps.nl/Cerpadla/Davkovaci_cerpadla/Mechanicka/Seko_SPRING_MS.
- [2] Rockwell Automation, *Allen-Bradley PowerFlex 40 Adjustable Frequency AC Drive*, User Manual. Dostupné na internete: <http://www.ab.com/support/abdrives/webupdate/literature.html>.
- [3] Rockwell Automation, *MicroLogix 1200 Programmable Controllers*, User Manual. Dostupné na internete: http://literature.rockwellautomation.com/idc/groups/literature/documents/um/1762-um001_-en-p.pdf.
- [4] Dinel, *Kapacitní hladinoměry CLM-36*, Návod k použití. Dostupné na internete: <http://www.dinel.cz/uploads/pdf/090916081606-clm-36-nav-cz.pdf>.
- [5] Rockwell Automation, *RSLogix 500*, Product Profile. Dostupné na internete: http://literature.rockwellautomation.com/idc/groups/literature/documents/pp/lg500-pp001_-en-e.pdf.
- [6] Rockwell Automation, *RSLinX Classic*, Getting Results Guide. Dostupné na internete: http://literature.rockwellautomation.com/idc/groups/literature/documents/gr/linux-gr001_-en-e.pdf.
- [7] Michal Blaho, „OPC – štandard priemyselnej komunikácie II“ v *Posteris.sk*, Ročník 2, číslo 5, ISSN 1338-0087.
- [8] Mathworks, *OPC Toolbox 2*, User Guide. Dostupné na internete: http://www.mathworks.com/help/pdf_doc/opc/opc.pdf.

Automatická analýza sentimentu textových zdrojov

Peter KONCZ, Ján PARALIČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

peter.koncz@tuke.sk, jan.paralic@tuke.sk

Abstrakt — Veľké množstvo subjektívneho obsahu publikovaného na internete je významným faktorom ovplyvňujúcim názory jeho používateľov. Automatická analýza sentimentu textov sa preto stáva dôležitou výskumnou oblasťou. Cieľom práce je poskytnutie prehľadu existujúcich prístupov k analýze sentimentu, ako aj overenie možností použitia vybraných postupov na slovenských textoch. V práci bol použitý korpus textov s používateľskými hodnoteniami internetových obchodov. Použitými metódami boli metóda podporných vektorov a dvojica klasifikačných algoritmov s použitím boostingu. Presnosť použitých metód je porovnateľná s presnosťou uvádzanou v prácach venovaných analýze sentimentu anglických textov.

Kľúčové slová — analýza sentimentu, klasifikácia textu, kontrolované učenie, strojové učenie

I. ÚVOD

Veľká časť obsahu publikovaného na internete je tvorená subjektívnymi názormi používateľov, pričom tieto názory významne ovplyvňujú názory iných používateľov [1; 2; 3; 4; 5; 6; 7; 8]. Vzhľadom na množstvo a význam tohto typu internetového obsahu je potrebné vyvíjať metódy automatickej analýzy subjektívnych aspektov textu. Skúmaním týchto metód sa zaoberá analýza sentimentu, ktorá sa stáva dôležitou výskumnou oblasťou [8; 9; 10; 11; 12; 13; 14; 7; 15; 16]. Zvyšujúci sa záujem o túto vednú oblasť pritom začal byť zrejmý až v minulom desaťročí [13; 3; 17; 18]. Počet prác venovaných analýze sentimentu v slovenských textoch je však doposiaľ malý. V súčasnosti ide predovšetkým o práce Budinská a kol. [19] Budinská a Kadora [14], prácu Rajzák a kol. [20] a Machová a Krajč [51]. Overenie použiteľnosti metód použitých prevažne v anglickom a čínskom jazykovom prostredí v rámci slovenských textov je preto jedným z hlavných cieľov tejto práce. Zvolenými metódami sú metóda podporných vektorov a dvojica klasifikačných algoritmov s použitím boostingu. Výsledok práce je členený nasledovne. Druhá kapitola je venovaná súvisiacim prácam, v tretej sú popisované možnosti využitia analýzy sentimentu, vo štvrtej kapitole je popisované experimentálne overenie zvolených metód analýzy sentimentu, napokon posledná časť je venovaná zhrnutiu a diskutovanému sú možnosti ďalšieho výskumu.

II. SÚVISIACE PRÁCE

Za počiatky analýzy sentimentu je možné považovať niektoré staršie práce venované príbuzným oblastiam výskumu. Prehľad niektorých z nich, venovaných oblastiam ako presvedčenia, interpretácia metafor, či uhol pohľadu, je možné nájsť v práci Pang a Lee [3]. Medzi prvé práce venované explicitne problematike analýzy sentimentu býva radená práca Hatzivassiloglou a McKeown [21], popisujúca metódu pre odhad orientácie sentimentu prídavných mien. Vrstávajúci záujem o analýzu sentimentu sa však prejavil až v minulom desaťročí [13; 3; 17; 18]. Rozsiahly prehľad prác, publikovaných väčšinou do roku 2005 je pritom možné nájsť v prácach Prabowo a Thelwall [11] ako aj Abbasi a kol. [22]. Analýze sentimentu sa tak v súčasnosti venuje veľké množstvo prác, líšiacich sa cieľmi, použitými vzorkami údajov, metódami, či mierami výkonnosti týchto metód.

Hoci cieľom veľkej časti prác je vytvorenie postupov na určenie sentimentu jednotlivých dokumentov [23; 24], nevyhnutne využívajú analýzu sentimentu na úrovni menších textových jednotiek. Väčšinou sú pritom metódy analýzy sentimentu aplikované na samostatné vety [12] a výsledky sú následne agregované vo výpočte sentimentu dokumentu.

Výskumy sa značne odlišujú aj charakteristikami vzorky. Vzhľadom k typu a tematickej doméne textov, sú tvorené filmovými recenziami [11; 9; 10; 7; 25; 14; 15; 26], hodnoteniami

produktov a služieb [11; 10; 27; 26], príspevkami z fór [8, 51], ale aj menej homogénnymi textami, ako napríklad tlačovými správami [23; 15]. Pre účely porovnávania výsledkov metód medzi jednotlivými prácami je pritom vhodné používanie korpusov. Korpusmi určenými pre oblasť analýzy sentimentu sú napríklad MPQA Opinion Corpus¹, Chinese Opinion Analysis Evaluation [28], voľne dostupný korpus použitý v rámci prác Pang a kol. [25; 7], či Multi-Domain Sentiment Dataset². Škála použitá na ohodnotenie dokumentov resp. viet v rámci vzorky je jednou z jej základných charakteristík. Voľba vhodnej škály pritom závisí od faktorov akými sú charakteristiky meraného konštruktú, respondentov, či jazyka v ktorom je škála prezentovaná. Vo väčšine analyzovaných prác bola polarita dokumentov hodnotená iba ako pozitívna, negatívna či neutrálna. Je však zrejme, že je možné použiť presnejšiu diferenciáciu subjektívnej orientácie textu. Vzhľadom na charakter premennej je vhodné použitie bipolárnej škály s vhodne zvoleným delením. Podľa výsledkov metaanalýz sa vhodnou voľbou pre bipolárne škály ukazuje sedembodová škála [29; 30]. Možnosti použitia presnejších škál v kontexte analýzy sentimentu sú pritom skúmané napríklad v práci Pang a Lee [7]. Medzi charakteristiky vzorky patrí aj jazyk textov. Vo väčšine analyzovaných prác boli vzorky tvorené anglickými [11; 10; 7; 23; 25; 14; 15; 26], ako aj čínskymi textami [8; 9; 10; 27; 12]. Z malého počtu prác, venovaných analýze sentimentu v iných jazykoch, je možné uviesť prácu Ahmad a Almas [31], kde bola vzorka tvorená arabskými textami, prácu Kieu a Pham [18], so vzorkou tvorenou vietnamskými textami, ako aj vyššie zmienené práce Budinská a Kadora [14], Rajzák a kol. [20] a Machová a Krajč [51], aplikujúce analýzu sentimentu na slovenské texty.

Charakteristiky vzorky údajov ovplyvňujú aj výber metódy analýzy sentimentu. Často sa pritom vyskytuje delenie týchto metód na metódy vychádzajúce zo strojového učenia resp. z analýzy prirodzeného jazyka [11], ako aj delenie na metódy využívajúce kontrolované resp. nekontrolované učenie [12; 27]. V ďalších častiach práce je však preferované delenie metód na endogénne a exogénne.

Pri endogénnych metódach analýzy sentimentu je odhad sentimentu funkciou algoritmu a vzorky údajov. V prácach využívajúcich endogénne metódy analýzy sentimentu sú aplikované metódy strojového učenia, akými sú metóda podporných vektorov (Support vector machines) [11; 9; 7; 25; 32; 26], naivný bayesovský klasifikátor [9; 25; 26], či metóda *k*-najbližších susedov [9]. K tejto skupine prác patria aj práce Hu a Li [10], využívajúca vlastný pravdepodobnostný model, Wilson a kol. [23], v ktorej boli použité klasifikátory na báze boostingu, Bai [15], využívajúca klasifikáciu pomocou tzv. Markov blanket, či práca Zhu a kol. [32], využívajúca umelé neuronové siete. Metódy strojového učenia sú pritom aplikované samostatne, alebo navzájom kombinované, ako napríklad v práci Xia a kol. [26].

V prípade exogénnych postupov je odhad sentimentu funkciou algoritmu, vzorky údajov, ako aj externého zdroja znalostí. Väčšina týchto metód vychádza z analýzy prirodzeného jazyka. Vonkajší zdroj znalostí je zvyčajne reprezentovaný slovníkmi, obsahujúcimi slová s priradenou hodnotou emocionálnej polarity, prípadne aj slová negujúce či modifikujúce polaritu iných slov. Príkladom takýchto slovníkov je anglicko-čínsky slovník pre analýzu sentimentu vychádzajúci z HowNetu³, MPQA subjectivity lexicon⁴, appraisal lexicon⁵, SentiWordNet⁶, obsahujúci množiny synonym z WordNetu⁷ a im priradené hodnoty sentimentu, či NTU Sentiment Dictionary⁸. Vzhľadom na časovú náročnosť anotácie slov ľuďmi, existujú snahy o automatické vytváranie týchto slovníkov. Kamps a kol. [33] popisujú metódu pre určenie orientácie prídavných mien na základe ich vzdialenosti od slov dobrý (good) a zlý (bad) v grafe vytvorenom na základe WordNetu. V práci Kim a Hovy [34] bol WordNet použitý na rozšírenie množiny emocionálnych slov prostredníctvom ich synonym a antonym. Ďalšie z postupov vychádzajú z využitia štatistických mier vzájomného výskytu slov resp. skupín slov v rámci korpusov. Turney [35] pre určenie orientácie fráz použil hodnoty tzv. bodovej vzájomnej súvislosti (pointwise mutual information; ďalej len PMI) medzi frázou a slovami excelentný (excellent) resp. biedny (poor). Prabowo a Thelwall [11] použili PMI ako súčasť metódy analýzy sentimentu spoločne s ďalšími mierami asociácie slov a slovných spojení. Výsledky experimentov so spájaním slovníkov popisované v Zhai a kol. [27] pritom poukazujú na kladnú závislosť medzi veľkosťou slovníka a presnosťou analýzy sentimentu. Z prác využívajúcich slovníky pre analýzu sentimentu je možné spomenúť práce Miao a kol. [36], Li a Wu [8], Zhai a kol. [27]. Pričom v prácach Li a Wu [8], Zhai a kol. [27] a Machová a Krajč [51] sú využívané aj slovníky negujúcich slov a intenzitu orientácie modifikujúcich slov.

¹ <http://www.cs.pitt.edu/mpqa/databaserelease/>

² <http://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

³ www.keenage.com

⁴ www.cs.pitt.edu/mpqa

⁵ lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz

⁶ sentiwordnet.isti.cnr.it

⁷ <http://wordnet.princeton.edu/>

⁸ nlg18.csie.ntu.edu.tw:8080/lwku

Vo väčšine analyzovaných prác boli pri analýze sentimentu použité sémantické a syntaktické atribúty, čo korešponduje so zistením Abbasi a kol. [22]. Ide o atribúty ako sú samostatné slová, n-gramy slov, štruktúry slov a ich vzory. Pri ich tvorbe a výbere môžu byť použité metódy ako part of speech analýza [26; 25] filtrovanie na základe hodnoty vzájomnej súvislosti, informačného zosilnenia, chí-kvadrátu, či dokumentovej frekvencie [9]. V niektorých prácach sú však používané aj menej časté atribúty, ako sú štruktúry odkazov a citácií [37], či tzv. emotikony (tiež „smajlíky“) [38; 39]. Veľká časť metód pritom využíva k analýze sentimentu slová a ich štruktúry iba v rámci jednotlivých viet. Zhao a kol. [12] však vo svojej práci dokazujú význam tzv. interdokumentových a intradokumentových závislostí pri identifikácii sentimentu jednotlivých viet. Pri analýze sentimentu viet tak môžu byť využité odhady hodnôt sentimentu okolitých viet, či podobných viet v rámci iných dokumentov.

Práce sa odlišujú aj použitými mierami výkonnosti metód pre analýzu sentimentu. Hodnota sentimentu je pritom zvyčajne reprezentovaná ako nominálna premenná a používané sú štandardné miery presnosti klasifikácie, s častým využitím n-násobnej krížovej validácie. Vo väčšine analyzovaných prác bola použitá priemerná presnosť klasifikácie (accuracy) [10; 7; 23; 25; 27; 15; 32; 26], ako aj presnosť, návratnosť, či z nich odvodená F miera [9; 23; 27; 11; 51]. Porovnanie mier výkonnosti klasifikácie je možné nájsť napríklad v práci Sokolova a Lapalme [40]. Priemerná presnosť klasifikácie sa v prácach venovaných analýze sentimentu pohybuje zväčša v rozmedzí 70 až 90 percent, pričom iba zriedka prekračuje hranicu 90 percent. Diverzita mier výkonnosti, ako aj vzoriek by však mala viesť k opatrnosti pri porovnávaní výsledkov metód naprieč prácami, ako na to upozorňujú aj Prabowo a Thelwall [11]. Jedným zo spôsobov ako čiastočne eliminovať skreslenie vyplývajúce z faktorov, ako sú rozdiely v počte kategórií a distribúcii textov resp. viet v týchto kategóriách, je porovnávanie výsledkov voči výkonnosti náhodného klasifikátora. V niektorých prácach je presnosť metód pre odhad sentimentu určovaná vzhľadom k paralelnému bodovému hodnoteniu, t.j. pokiaľ sa k slovnému hodnoteniu viaže aj bodové hodnotenie, potom je na základe modelu analýzy sentimentu a slovného hodnotenia získaný odhad sentimentu, ktorý sa následne porovná voči bodovému hodnoteniu. V tejto súvislosti uvádzajú Pang a kol. [25] doménu filmových recenzií za „experimentálne vhodné“, práve vzhľadom na vyššie zmienený spôsob hodnotenia filmov používateľmi. Závažným nedostatkom popísaného prístupu je, že bodové hodnotenia pochádzajú iba od autorov slovných hodnotení. Presnosť klasifikácie sentimentu dokumentov je podľa Tan a Zhang [9] výrazne ovplyvňovaná aj konkrétnou doménou resp. témou. Zhai a kol. [27] v tejto súvislosti píšú o doménovej závislosti tzv. citovo nejednoznačných slov (sentiment noise words), ktoré sú citovými slovami sťažujúcimi výpočet polaritu sentimentu v konkrétnej doméne.

III. OBLASTI VYUŽITIA

Medzi perspektívnymi oblasťami využitia analýzy sentimentu zvyknú byť uvádzané predovšetkým identifikácia spotrebiteľských preferencií a podpora rozhodovania pri kúpe produktov. Výsledky štúdie vedenej The Kelsey Group a comScore uvádzajú, že spotrebiteľia sú ochotní zaplatiť, v závislosti od typu tovaru resp. služby, o 20 až 99 percent viac v situácii, keď je produkt označený ako excelentný („5 hviezdčiek“) v porovnaní so situáciou, kedy je rovnaký produkt označený ako dobrý („4 hviezdčiky“) [4]. V spomenutej štúdii sa tiež uvádza, že 73 až 87 percent respondentov používajúcich internet, taktiež v závislosti od typu tovaru resp. služby, zohľadňuje on-line recenzie kupovaných produktov. Uvedené tvrdenia korešpondujú s výsledkami uvádzanými Horriganom [2], podľa ktorých až 81% používateľov internetu v skúmanej vzorke uvádzalo, že pri kúpe produktov využívajú internet pri získavaní informácií o danom produkte. Potreba manažmentu on-line hodnotení je zdôrazňovaná aj v práci Lee a kol. [5]. Autori sa zamerali na negatívne hodnotenia, ktoré majú na hodnotenie spotrebiteľov väčší vplyv než pozitívne hodnotenia, pričom sa odvolávajú na viaceré práce, ktoré tento názor potvrdzujú. Qiu a kol. [41] vo svojej práci popisujú webový reklamný systém, v ktorom je reklama prispôbovaná nielen téme, ale aj sentimentu textu na stránke. Niektoré z prác však vyššie popisované závislosti v určitých kontextoch popierajú. Duan a kol. [1], ktorí sa venovali vplyvu on-line hodnotení filmov na ich predajnosť uvádzajú, že tento vplyv nie je významný. Zároveň však uvádzajú, že význam on-line hodnotení produktov spočíva v šírení povedomia o produkte a presvedčavom vplyve, pričom význam majú aj v rámci predikcií predaja. Vplyv, ktorý majú on-line hodnotenia tovarov na spotrebiteľov je závislý aj od autorstva hodnotenia, pričom sa ukazuje, že hodnotenia bežných konzumentov služieb sú spotrebiteľmi vnímané ako dôležitejšie než hodnotenia editorov resp. „profesionálov“ [4; 6].

V posledných rokoch sa princípy analýzy sentimentu začínajú uplatňovať aj v rámci reálnych, verejne prístupných aplikácií. Wright [42] ako príklady takýchto služieb uvádza služby poskytované spoločnosťami Lithium⁹, Apinions¹⁰, či služby ako Tweetfeel, Twendz and Twitr-

⁹ <http://www.lithium.com>; pôvodne Scout Labs

ratr. Spoločnosť Lithium používateľom svojich služieb umožňuje monitorovať názory na služby, produkty, ako aj témy v rámci blogov, online správ, fór, či sociálnych sietí. Podobne Appinions poskytuje služby, ktoré používateľom umožňujú sledovať názory uverejňované na internete. Trojica služieb Tweetfeel¹¹, Twendz¹² a Twitrratr¹³ pre podobné účely využíva analýzu sentimentu príspevkov v rámci služby Twitter¹⁴. Príkladmi služieb využívajúcich analýzu sentimentu sú aj Heartbeat a Media Analysis Platform, poskytované spoločnosťou Sysomos¹⁵, ktoré analýzu sentimentu využívajú ako súčasť služieb pre analýzu sociálnych médií. Niektoré z týchto služieb pritom poskytujú možnosti analýzy sentimentu aj prostredníctvom svojho API (Application Programming Interface), ako je to aj v prípade Evri¹⁶ [43].

IV. EXPERIMENT

Cieľom experimentálnej časti práce je overenie použiteľnosti vybraných metód strojového učenia pre účely analýzy sentimentu slovenských textov.

A. Vzorka

Použitá vzorka údajov bola tvorená 400 vetami obsahujúcimi hodnotenia internetových obchodov, získanými z portálu heureka (www.heureka.sk). Vety tvoriace vzorku boli expertom ohodnotené ako pozitívne resp. negatívne orientované. Vyberané pritom boli iba vety s jednoznačnou orientáciou, ktoré boli zmysluplné aj mimo kontextu hodnotiaceho príspevku. Vzorka bola tvorená 200 pozitívnymi vetami a 200 negatívnymi vetami.

B. Postup

V rámci predpracovania bola v textoch najskôr odstránená interpunkcia, následne bola vykonaná lematizácia slov, odstránenie diakritiky a bola získaná vektorová reprezentácia textov. Pre lematizáciu textu bol použitý algoritmus, využívajúci slovník poskytnutý Jazykovedným ústavom Ľudovíta Štúra SAV, obsahujúci slová v základnom tvare, ich morfológické tvary, ako aj morfológickú anotáciu. Lematizácia slov teda spočívala v nájdení slova a jemu priradeného základného tvaru v slovníku. Disambiguácia nebola súčasťou lematizácie. Vektorová reprezentácia textov bola získaná aplikáciou *tf-idf* schémy, s využitím termovej frekvencie normalizovanej na maximálnu hodnotu a inverznej dokumentov frekvencie, pričom výsledný vektor bol normovaný na jednotkovú dĺžku vektora. Texty boli rozdelené do štyroch skupín vo veľkosti 100 viet, pričom bola zachovaná rovnaká distribúcia tried ako v pôvodnej vzorke. Toto rozdelenie do skupín bolo použité v rámci rozvrstvenej krížovej validácie, kde v každej iterácii bola jedna zo skupín testovacou vzorkou, kým zvyšné tri skupiny boli trénovacou vzorkou. Vytvorené vektory boli následne klasifikované s využitím knižnice JBowL (Java Bag-of-words Library) [44]. Použitými metódami strojového učenia boli metóda podporných vektorov (support vector machines) a dvojica algoritmov s použitím boostingu, Boosting a BoosTexter. Metóda podporných vektorov bola najčastejšie používanou metódou v rámci analyzovaných prác a jej popis je možné nájsť napríklad v práci Prabowo & Thelwall [11]. Pri tejto metóde boli použité jadrová funkcia „LinearKernel“, penalizačný parameter *C* pre pozitívne aj negatívne príklady rovný 1 a požadovaná presnosť rovná 0,001. Algoritmus BoosTexter bol pre analýzu sentimentu použitý napríklad v práci Wilson a kol. [23]. Pri algoritme Boosting boli ako základné klasifikátory použité jednoduché klasifikátory, klasifikujúce na základe výskytu jedného termu a maximálny počet klasifikátorov bol nastavený na 1000. Algoritmy použitých postupov sú voľne dostupné v rámci už spomenutej knižnice JBowL.

Tabuľka 1
 POROVNANIE PRESNOSTI POUŽITÝCH METÓD

Metóda	Presnosť v iterácii				Priemerná presnosť
	1.	2.	3.	4.	
SVM	90	95	90	89	91
BoosTexter	78	89	86	82	83,75
Boosting	76	72	82	69	74,75

SVM = Metóda podporných vektorov (Support Vector Machines), hodnoty sú uvádzané v %.

¹⁰ <http://appinions.com>; pôvodne Jodange

¹¹ <http://www.tweetfeel.com>

¹² <http://twendz.waggeneratedstrom.com>

¹³ <http://twitrratr.com>

¹⁴ <http://twitter.com/>

¹⁵ <http://sysomos.com>

¹⁶ <http://www.evri.com>

C. Výsledky

Zvolené klasifikačné metódy boli porovnávané na základe hodnoty priemernej presnosti s použitím rozvrstvenej štvornásobnej krížovej validácie. Získané hodnoty presností v rámci jednotlivých iterácií a priemerné presnosti sú uvedené v tabuľke 1. Teoretická presnosť náhodného klasifikátora je vzhľadom na charakteristiky vzorky 50%. V rámci porovnávaných metód sa ako najpresnejšia ukázala metóda podporných vektorov s priemernou presnosťou 91%, menej presnými sa ukázali algoritmy BoosTexter s priemernou presnosťou 83,75% a Boosting s priemernou presnosťou 74,75%.

V. ZÁVER

Presnosť získaných výsledkov je porovnateľná s presnosťou výsledkov prezentovaných v prácach venovaných analýze sentimentu anglických textov, pričom priemerná presnosť pre metódu podporných vektorov je v kontexte analyzovaných prác relatívne vysoká. Za hlavný nedostatok experimentu je možné považovať nedostatočnú veľkosť korpusu. Metódy overované v experimentálnej časti pritom predstavujú iba zlomok z metód použiteľných pre analýzu sentimentu. Použitelnosť týchto metód by mohla byť predmetom ďalšieho výskumu. Pre realizáciu výskumov v oblasti analýzy sentimentu slovenských textov je tiež dôležité vytvorenie dostatočne veľkých polytematických paralelných korpusov, anotovaných pre tieto účely.

POĎAKOVANIE

Slovník použitý v experimentálnej časti práci, bol poskytnutý Jazykovedným ústavom Ľudovíta Štúra SAV, za čo autor ďakuje. Táto práca bola podporovaná Vedeckou grantovou agentúrou Ministerstva školstva Slovenskej republiky na základe projektu č. 1/0042/10 (50%) a realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (50%).

LITERATÚRA

- [1] Duan, Wenjing, Bin Gu, a Andrew B. Whinston. „Do online reviews matter? — An empirical investigation of panel data.“ *Decision Support Systems*, Volume 45, Issue 4. vyd. (2008): 1007-1016.
- [2] Horrigan, John A. *Pew Internet & American Life Project Report*. 13. February 2008. http://www.pewinternet.org/~media/Files/Reports/2008/PIP_Online%20Shopping.pdf (cit. 2. 1. 2010).
- [3] Pang, Bo, a Lillian Lee. „Opinion Mining and Sentiment Analysis.“ *Foundations and Trends in Information Retrieval* (Now Publishers), Vol. 2: No 1–2. vyd. (2008): 1-135.
- [4] comScore; The Kelsey Group Reveals t. *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. 29. November 2007. http://www.comscore.com/Press_Events/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior (cit. 1. 1. 2011).
- [5] Lee, Jumin, Do-Hyung Park, a Ingoo Han. „The effect of negative online consumer reviews on product attitude: An information processing view.“ *Electronic Commerce Research and Applications*, Vol. 7, Issue 3. (2008): 341-352.
- [6] Zhang, Ziqiong, Qiang Ye, Rob Law, a Yijun Li. „The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews.“ *International Journal of Hospitality Management*, Volume 29, Issue 4. vyd. (2010): 694-700.
- [7] Pang, Bo, a Lillian Lee. „Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.“ *Proceedings of the ACL*, 2005: 115-124.
- [8] Li, Nan, a Desheng Dash Wu. „Using text mining and sentiment analysis for online forums hotspot detection and forecast.“ *Decision Support Systems*, Volume 48, Issue 2. vyd. (2010): 354-368.
- [9] Tan, Songbo, a Jin Zhang. „An empirical study of sentiment analysis for chinese documents.“ *Expert Systems with Applications*, Volume 34, Issue 4. vyd. (2008): Volume 34, Issue 4.
- [10] Hu, Yi, a Wenjie Li. „Document sentiment classification by exploring description model of topical terms.“ *Computer Speech & Language*, Volume 25, Issue 2. vyd. (2011): 386-403.
- [11] Prabowo, Rudy, a Mike Thelwall. „Sentiment analysis: A combined approach.“ *Journal of Informetrics*, Volume 3, Issue 2. vyd. (2009): 143-157.
- [12] Zhao, Yan-Yan, Bing Qin, a Ting Liu. „Integrating Intra- and Inter-document Evidences for Improving Sentence Sentiment Classification.“ *Acta Automatica Sinica*, Volume 36, Issue 10. vyd. (2010): 1417-1425.
- [13] Tang, Hui Feng, Songbo Tan, a Xueqi Cheng. „A survey on sentiment detection of reviews.“ *Expert Systems with Applications*, Volume 36, Issue 7. vyd. (2009): 10760-10773.
- [14] Budinská, Ivana, a Andrej Kadora. „New Aspects on Sentiment Analysis of Slovak Texts.“ *6th International Workshop on Grid Computing for Complex Problems (GCCP 2010)*. Bratislava, Slovakia: Institute of Informatics SAS, 2010. 110-114.
- [15] Bai, Xue. „Predicting consumer sentiments from online text.“ *Decision Support Systems*, Volume 50, Issue 4. vyd. (2011): 732-742.
- [16] Neviarouskaya, Alena, Helmut Prendinger, a Mitsuru Ishizuka. „SentiFul: Generating a Reliable Lexicon for Sentiment Analysis.“ *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*. Amsterdam, Netherlands: IEEE, 2009. 363-368.
- [17] Mejova, Yelena. „Sentiment Analysis: An Overview.“ [www.cs.uiowa.edu](http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf). 16. November 2009. <http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf> (cit. 2011. 1. 23).
- [18] Kieu, Binh Thanh, a Son Bao Pham. „Sentiment Analysis for Vietnamese.“ *Second International Conference on Knowledge and Systems Engineering*. Hanoi, Vietnam: IEEE Computer Science Society, 2010. 152-157.

- [19] Budinská, Ivana, Zoltán Balogh, a Emil Gatiaľ. „Monitorovanie správ rozšírené o metódy analýzy sentimentu.“ Editor Pavol Návrat a Valentino Vranič. *Proceedings of 3rd Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2008)*. Smolenice, Slovakia: Nakladateľstvo STU, 2009. 74-77.
- [20] Rajzák, Pavol, Peter Koncz, Mojmir Pridavok, Radoslav Blahovec, a Miloš Pridavok. „Systém pre hodnotenie online reputácie bánk.“ *Proceedings of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice*. Košice, Slovakia: Fakulta elektrotechniky a informatiky, 2010. 652 - 657.
- [21] Hatzivassiloglou, Vasileios, a Kathleen R. McKeown. „Predicting the Semantic Orientation of Adjectives.“ *Proceedings EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Stroudsburg, USA: Association for Computational Linguistics, 1997. 174-181.
- [22] Abbasi, Ahmed, Hsinchun Chen, a Arab Salem. „Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums.“ *ACM Transactions on Information Systems*, Volume 26, Issue 3. vyd. (2008): 12:1-12:34.
- [23] Wilson, Theresa, Janyce Wiebe, a Paul Hoffmann. „Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.“ *Proceeding of the conference on empirical methods in natural language processing (EMNLP 2005)*. Vancouver: Association for Computational Linguistics Morristown, 2005. 347-354.
- [24] Choi, Yejin, Claire Cardie, Ellen Riloff, a Siddharth Patwardhan. „Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns.“ *Proceeding of the conference on empirical methods in natural language processing (EMNLP 2005)*. Vancouver: Association for Computational Linguistics Morristown, 2005. 355-362.
- [25] Pang, Bo, Lillian Lee, a Shivakumar Vaithyanathan. „Thumbs up?: sentiment classification using machine learning techniques.“ *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP 2002)*. Philadelphia: Association for Computational Linguistics Morristown, 2002. 79-86.
- [26] Xia, Rui, Chengqing Zong, a Shoushan Li. „Ensemble of feature sets and classification algorithms for sentiment classification.“ *Information Sciences*, Volume 181, Issue 6. vyd. (2011): 1138-1152.
- [27] Zhai, Zhongwu, Hua Xu, a Peifa Jia. „An Empirical Study of Unsupervised Sentiment Classification of Chinese Reviews.“ *Tsinghua Science & Technology*, Volume 15, Issue 6. vyd. (2010): 702-708.
- [28] Zhao, J., H. Xu, X. Huang, S. Tan, K. Liu, a Q. Zhang. „Overview of Chinese Opinion Analysis Evaluation 2008.“ *Proceedings of the First Chinese Opinion Analysis Evaluation*. 2008.
- [29] Malhotra, Neil, Jon A. Krosnick, a Randall K. Thomas. „Optimal Design of Branching Questions to Measure Bipolar Constructs.“ *Public Opinion Quarterly*, Volume 73, Issue 2. vyd. (2009): 304-324.
- [30] Krosnick, Jon A., a Alexander M. Tahk. „The Optimal Length of Rating Scales to Maximize Reliability and Validity.“ *Unpublished manuscript* (Unpublished manuscript. Stanford University), dátum neznámy.
- [31] Ahmad, Khurshid, a Yousif Almas. „Visualising Sentiments in Financial Texts?“ *Proc. of the Ninth International Conference on Information Visualisation (IV'05)*. London, UK: IEEE Computer Society, 2005. 363-368.
- [32] Zhu, Jian, Chen Xu, a Han-shi Wang. „Sentiment classification using the theory of ANNs.“ *The Journal of China Universities of Posts and Telecommunications*, Volume 17, Supplement 1. vyd. (2010): 58-62.
- [33] Kamps, Jaap, Maarten Marx, Robert J. Mokken, a Maarten de Rijke. „Using WordNet to Measure Semantic Orientations of Adjectives.“ *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal, 2004. 1115-1118.
- [34] Kim, Soo-Min, a Eduard Hovy. „Automatic Detection of Opinion Bearing Words and Sentences.“ *Proceedings of the Second International Joint Conference on Natural Language Processing (JCNLP 2005)*. Jeju Island, Korea, 2005. 61-66.
- [35] Turney, Peter D. „Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002).“ *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*. Philadelphia: Association for Computational Linguistics, 2002. 417-424.
- [36] Miao, Qingliang, Qiudan Li, a Ruwei Dai. „AMAZING: A sentiment mining and retrieval system.“ *Expert Systems with Applications*, Volume 36, Issue 3. vyd. (2009): 7192-7198.
- [37] Efron, Miles. „Cultural orientation: Classifying subjective documents by cocitation analysis.“ *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*. AAAI Press, 2004. 41-48.
- [38] Pak, Alexander, a Patrick Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining.“ *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), 2010. 1320-13626.
- [39] Read, Jonathon. „Using emoticons to reduce dependency in machine learning techniques for sentiment classification.“ *Proceedings of the ACL Student Research Workshop (ACLstudent 2005)*. Michigan: Association for Computational Linguistics, 2005. 43-48.
- [40] Sokolova, Marina, a Guy Lapalme. „A systematic analysis of performance measures for classification tasks.“ *Information Processing & Management*, Volume 45, Issue 4. vyd. (2009): 427-437.
- [41] Qiu, Guang, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, a Chun Chen. „DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis.“ *Expert Systems with Applications*, Volume 37, Issue 9. vyd. (2010): 6182-6191.
- [42] Wright, Alex. *Mining the Web for Feelings, Not Facts*. 23. August 2009. http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=2&partner=rss&emc=rss (cit. 7. 1. 2011).
- [43] Perez, Sarah. *How Does the Web Feel? Evri's New Sentiment API Tells You*. 14. August 2009. http://www.readwriteweb.com/archives/how_does_the_web_feel_evri_tells_you.php (cit. 8. 1. 2011).
- [44] Bednár, Peter, Peter Butka, a Ján Paralič. „Java Library for Support of Text Mining and Retrieval.“ *Proceedings of the 4th annual conference Znalosti 2005*. Ostrava: VŠB TU, 2005. 162-169.
- [45] Li, Gang, a Fei Liu. „A Clustering-based Approach on Sentiment Analysis.“ *International Conference on Intelligent Systems and Knowledge Engineering (ISKE, 2010)*. Hangzhou, China: IEEE, 2010. 331-337.
- [46] Blaga, Lucian. „Web Document Classification and its Performance Evaluation.“ *9th WSEAS International Conference on Evolutionary Computing*. Sofia, 2008. 105-110.
- [47] Zhan, Jiaming, Han Tong Loh, a Ying Liu. „Gather customer concerns from online product reviews – A text summarization approach.“ *Expert Systems with Applications*, Volume 36, Issue 2. vyd. (2009): 2107-2115.
- [48] Li, Dun, Yong-tao Ma, a Jian-li Guo. „Words semantic orientation classification based on HowNet.“ *The Journal of China Universities of Posts and Telecommunications*, Volume 16, Issue 1. vyd. (2009): 106-110.
- [49] Li, Si, Hao Zhang, Weiran Xu, Guang Chen, a Jun Guo. „Exploiting Combined Multi-level Model for Document Sentiment Analysis.“ *Pattern Recognition (ICPR), 2010 20th International Conference on*. Istanbul, 2010.
- [50] Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, a Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: The MIT Press, 1966.
- [51] Machová K. a Krajč M. *Klasifikácia názorov vo vláknových diskusiách na webe. Proceedings of the 10th annual conference Znalosti 2011*. Stará Lesná: VŠB TU, 2011. 136-147.

Návrh automatického dopĺňania Wordnetu

Viliam ROČKAI

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

viliam.rockai@gmail.com

Abstrakt – Slovníky wordnetového typu už dlhé roky nachádzajú široké uplatnenie nie len v doméne spracovania textov v prirodzenom jazyku. Ich využitiu pri spracovaní NLP je venované nepočtené množstvo prác, no oblasť ich automatického vytvárania stále naráža na mnoho problémov. Najbežnejšími postupmi bývajú automatické preklady podobných slovníkov v iných jazykoch, prípadne prístupy so silnými znalosťami o jazyku, nad ktorým pracujú. Keďže pojmy sú v slovníku zoradené hierarchicky podľa rôznych sémantických relácií (hypernymá, hyponymá, príbuznosť...) a s využitím teórie asociatívneho učenia pojmov je možné podobné stromy pojmov vytvárať, existuje predpoklad na jej využitie pri budovaní wordnetových slovníkov. Táto práca sa zaoberá možnosťou využitia teórie asociatívneho učenia pojmov v procese budovania slovníku wordnetového typu, na základe správneho zaradenia nových pojmov do podstromu hierarchicky zoradených pojmov v slovníku.

Kľúčové slová – asociatívne učenie pojmov, Wordnet, spracovanie prirodzeného jazyka

I. ÚVOD

Asociatívne učenie pojmov je prístup založený na neuro-fyziologickom modeli spracovania talamo-kortikálnej informácie od R. Hecht Nielsena [1]. Model predpokladá existenciu fixného lexikónu **symbolov** v ľudskom talame, ktorý sa vytvorí v skorom vývinovom štádiu jedinca. Učenie potom pozostáva z vytvárania asociácií medzi neurónmi, odrážajúc tak neurónové spojenia medzi kortikálnymi regiónmi. Najnovšie znalosti o teórii talamu sa nachádzajú v [2]. Dynamika učenia je namodelovaná podľa hebbovskej metódy. Asociatívne učenie pojmov je teda nekontrolované učenie nad prúdom symbolov. Cieľom je naučenie sa znalostnej reprezentácie štruktúry do podoby sémantických sietí, prípadne hierarchií konceptov. Metóda je založená na indukcii asociácií medzi symbolmi vzhľadom na ich spoločné výskyty v kontextovom okne, cez ktoré samotné učenie prebieha.

WordNet je sémantický lexikón anglického jazyka. Zoskupuje anglické slová do množín synonymým, nazývaných synsety a poskytuje krátke definície a záznamy rôznych sémantických relácií medzi týmito synsetmi. Má to dvojaký účel: poskytnúť slovník a teaurus, ktorý by bol viac intuitívne použiteľný a poskytnúť podporu automatickej analýzy textov a aplikácií pre UI. Databáza a softwarové nástroje sú vydané pod BSD style licenciou a môžu byť voľne stiahnuté a používané. Databázu je možné prezerať si aj on-line.

II. ASOCIATÍVNE UČENIE KONCEPTOV

Nutná podmienka pre vytvorenie asociácie medzi párom tokenov je daná ich štatisticky nenáhodným spoločným výskytom. Implementuje sa podľa vzorca na výpočet signifikancie:

$$S(i, j) = \frac{p(i, j)}{p(i) \cdot p(j)} \quad (1)$$

kde i a j sú diskkrétne náhodné premenné. Vzájomná signifikancia $S(i, j)$ je potom definovaná ako podiel vzájomnej pravdepodobnosti symbolov (diskrétnych premenných) i, j a pravdepodobnosti ich výskytu. Ak hodnota $S(i, j)$ prekročí vopred určenú prahovú hodnotu, považujeme dané symboly za asociované.

Vzájomná informácia (MI) je mierou prekrytia informácie dvoch náhodných premenných. MI dvoch náhodných premenných X a Y , ktorých hodnoty majú apriorne pravdepodobnosti $p(x)$ and $p(y)$ a pravdepodobnosť spoločného výskytu $p(x, y)$, je definovaná ako:

$$MI(X; Y) = \sum_{x, y} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

Bodová vzájomná informácia (PMI) je miera toho, ako veľmi sa líši pravdepodobnosť spoločného výskytu dvoch javov $p(x, y)$ od hodnoty, ktorú by sme očakávali ak by boli dané javy na sebe nezávislé (teda $p(x)p(y)$) [3]. PMI je definovaná ako:

$$PMI(x, y) = \ln \frac{p(x, y)}{p(x)p(y)} \quad (3)$$

A. Výpočet sémantickej príbuznosti

Prístup je založený na Jaccardovom indexe (4) počítania podobnosti alebo diverzity dvoch množín:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

Rovnica 4 popisuje výpočet podobnosti dvoch množín A a B. Obor hodnôt Jaccardovho indexu je $\langle 0, 1 \rangle$, takže nie je potrebné ho ďalej normalizovať. V AUP je každý symbol množinou asociácií, teda svojim asociovaným okolím v rôznych kontextových vzdialenostiach. Okolie symbolu x v kontextovej vzdialenosti i definujeme ako $O_i(x)$ (5).

$$O_i(x) = \{t \in L | Si(x, t) > thresh\} \quad (5)$$

$O_i(x)$ je teda množinou všetkých takých symbolov z množiny známych symbolov, lexikonu L , pre ktoré platí, že sú so symbolom x asociované v kontextovej vzdialenosti i . Podobnosť dvoch symbolov a a b vzhľadom na kontextovú vzdialenosť i môžeme definovať ako:

$$Rel_i(a, b) = \frac{|O_i(a) \cap O_i(b)|}{|O_i(a) \cup O_i(b)|} \quad (6)$$

Váňovanou sumou podobností symbolov cez všetky kontextové vzdialenosti dostávame vzorec na výpočet podobnosti dvoch symbolov $Rel_w(a, b)$:

$$Rel_w(a, b) = \sum_{i=0}^n w_i Rel_i(a, b) \quad (7)$$

Oborom hodnôt $Rel_w(a, b)$ je množina $\langle 0, \sum_{i=0}^n w_i \rangle$. Ak sú váhy pre jednotlivé kontextové vzdialenosti volené tak, že ich súčet dáva hodnotu 1, nie je nutné ďalej nutné výsledné hodnoty normalizovať.

Výpočtom podobností vybraného symbolu so všetkými symbolmi z lexikónu L dostávame zoradený zoznam podobností.

III. NÁVRH PROCESU DOPLŇANIA WORDNETU

A. Wordnet

WordNet vznikol v roku 1985 a je udržiavaný na pôde Princetonskej univerzity v Laboratóriu kognitívnych vied pod vedením profesora psychológie George A. Millera. v roku 2003 obsahovala databáza okolo 140 000 slov organizovaných do viac ako 110 000 synsetov s celkovým počtom 195 000 slovných párov.

WordNet rozlišuje medzi podstatnými menami, slovesami, prídavnými menami a príslovkami na predpoklade, že aj v ľudskom mozgu sa tieto štruktúry ukladajú rozdielne. Každý synset obsahuje množinu synonymických slov alebo slovných spojení, ktoré produkujú špecifický význam (napr. Vysoká škola). Slová sa zvyčajne vyskytujú vo viacerých synsetoch naraz. Význam synsetov je ďalej doladený krátkymi poznámkami.

Typickým príkladom synsetu s poznámkou je: správny, vhodný, akurátny – (najvhodnejší pre dané špecifické účely; **správny čas orať, vhodný moment na ústup**, atď.).

Každý synset je relačne prepojený na iné synsety. Tieto relácie závisia od typu slova. Jedná sa o tieto typy relácií.

- Synonymá – synsety s podobným významom.
- Hypernymá - Y je hypernymom k X, ak každé X je (ako) Y.
- Hyponymá - Y je hyponymom k X, ak každé Y je (kind of) X.
- Holonymá - Y je holonymom k X, ak X je časťou Y.
- Meronymá - Y je meronymom k X, ak Y je časťou X.
- Antonymá – slová opačného významu.
- Rovnocenné pojmy – slová zdieľajúce bežné hypernymum.

Ako príklad nám môže poslúžiť anglické slovo **string** myslené vo význame **struna** (napríklad na gitare). Použitím on-line nástroja [5] po zadaní daného slova dostaneme zoznam možných

výsledkov. Pre možnosť **a tightly stretched cord of wire or gut, which makes sound when plucked, struck, or bowed** je možný dotaz **domain category**, ktorý doménovou reláciou dáva strunu do súvislosti s **music (an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner)**, čiže hudbou.

WordNet poskytuje hodnotu **polysemy count**, ktorá vyjadruje počet synsetov, ktoré dané slovo obsahujú. Ak sa dané slovo vyskytuje vo viacerých synsetoch (t.z. má viacero významov), tak je zrejmé, že niektoré významy sú bežnejšie ako iné. WordNet to kvantifikuje pomocou **frequency score**. V niekoľkých vzorkách textov sú všetky slová sématicky označené korešpondujúcim synsetom a neskôr je spočítané koľkokrát sa vyskytli s daným špecifickým významom. Databázové rozhranie vie z užívateľského vstupu dedukovať koreňovú formu slova a iba do databázy sa ukladajú iba tieto formy [4].

B. Využitie asociatívneho učenia pojmov

Keďže s využitím teórie asociatívneho učenia pojmov je možné podobné stromy pojmov vytvárať, existuje predpoklad na jej využitie pri budovaní wordnetových slovníkov. Táto práca predstavuje postup, ktorý by mohol slúžiť pri procese budovania slovníku wordnetového typu, na základe správneho zaradenia nových pojmov do podstromu hierarchicky zoradených pojmov v slovníku. Predpokladá využitie teórie asociatívneho učenia pojmov ako poskytovateľa pomocných znalostí v tomto procese. Proces správneho zaradenia pojmu by prebiehal v štyroch základných krokoch:

- 1) Výpočet podobných pojmov k pojmu, ktorý chceme zaradiť do stromu Wordnetu.
- 2) Vyhľadanie podoných pojmov vo Wordnete.
- 3) Výpočet spoločného podstromu.
- 4) Zaradenie dotazovaného pojmu do stromu Wordnetu.

V prvom kroku sa využije implementácia AUP k vytvoreniu množiny podobných slov k dotazovanému slovu, ktoré chceme správne zaradiť do podstromu Wordnetu. Vzhľadom na sémantickú príbuznosť takto vygenerovaných slov existuje predpoklad, že prienikom ich podstromov relácie hyponým vo Wordnete bude podstrom, do ktorého bude patriť aj dotazované slovo.

V druhom kroku sa vyhľadajú pozície všetkých podobných slov k dotazovanému pojmu za účelom získania informácie o ich podstromoch.

V treťom kroku nasleduje výpočet samotného podstromu, do ktorého bude v poslednom kroku dotazovaných pojem zaradený. Táto časť je kľúčovým bodom procesu. Bude pozostávať ako z filtrácie irelevantných podstromov (ktoré mohli vzniknúť vďaka vágnosti podobných pojmov poskytnutých AUP), tak aj zo samotného výpočtu výslednej pozície dotazovaného pojmu, ktorá bude listovým uzlom najkratšieho spoločného podstromu.

Vo štvrtom kroku dojde k zaradeniu dotazovaného pojmu na vypočítanú pozíciu.

IV. ZÁVER

Táto práca sa zaoberala možnosťou využitia teórie asociatívneho učenia pojmov v procese budovania slovníku wordnetového typu, na základe správneho zaradenia nových pojmov do podstromu hierarchicky zoradených pojmov v slovníku. Popísala základné body postupu, ktorý by mohol slúžiť k takémuto účelu s využitím výpočtu podobných slov pomocou AUP, a ich vyhľadaním v strome hyponým Wordnetu. Do budúcnosti je potrebné vykonať spresnenie postupu filtrácie podstromov podobných slov, ako aj experimentálne overenie presnosti takéhoto modelu.

PREHLÁSENIE

Práca prezentovaná v tomto článku bola podporená Vedeckou grantovou agentúrou Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky v rámci 1/0042/10 projektu Metódy identifikácie, anotácie, vyhľadávania a kompozície služieb, za použitia sémantických metadát, ako podpora vybraných procesných typov.

LITERATÚRA

- [1] Nielsen R. H.: *A Theory of the Cerebral Cortex*, ICONIP, 1998
- [2] Nielsen R. H.: *Confabulation Theory: The Mechanism of Thought*, Springer, 2007
- [3] Bouma G., Normalized (Pointwise) Mutual Information in Collocation Extraction, In Proceedings of the Conference of the German Society for Computational Linguistics 2009, GSCL-2009, Potsdam, Germany, 2009
- [4] V. Ročkai, 2005. *Mining of Concepts and Semantic Relations from Texts in Natural Language* Diplomová práca, Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita, Košice, 2005
- [5] Wordnet, *Wordnet online*, Dostupné online: <http://wordnet.princeton.edu/perl/webwn>

Structural Analysis of Social Groups in Collaboration Network

Gabriel TUTOKY, Martin REPKA, Ján PARALIČ

Dept. of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and
Informatics, Technical University of Košice, Slovak Republic

gabriel.tutoky@tuke.sk, martin.repka@tuke.sk, jan.paralic@tuke.sk

Abstract — This article describes approaches of positional analysis and community detection in collaboration network. We discuss positional analysis background, data types and community detection methods focusing on evolution and properties of collaboration network. Next we shortly describe transformation and processing of available data and possible inclusion of other types of data into structural analysis. As types of data we suppose unstructured data consisting of compositional and temporal attributes.

Keywords — social network analysis, structural analysis, positional analysis, community detection, collaboration networks

I. INTRODUCTION

Collaboration networks have a lot of common with affiliation networks, which is a special type of two-mode social networks. One mode is a set of actors, and the second mode is a set of events which are affiliated to the actor. The ties between actor and event are created, if actor participates on particular event. Affiliation networks describe collections of actors rather than simply ties between pairs of actors and connections among members of one of the modes are based on linkages established through the second mode [1]. These networks can be expressed by bipartite graphs as the most used representation for affiliation networks.

As we presume that specific actors attend to the certain event, they should form some kind of positions or roles in this network. Since the position, role or social category is reflected in relations between actors we can empirically identify and define the social position by using network data. The question is, if this is satisfactory enough for a precise identification of a position or role.

In this article we discuss analysis of larger components in the network – social groups instead of focusing on individual actors. General opinion in research in this field indicates that it is more important studying of relations between sets of actors instead of relations between individual actors [1], [2]. In our case we are analyzing social groups by approaches of positional analysis and community detection algorithms.

II. POSITIONAL ANALYSIS

Intuitively we could say that two actors have the same position or role to the extent that the profile of their relations with other actors is the same. There is not simple way to find out which are the essential relations within the actors and indeed we cannot determine the relevant set of actors. It all depends on the purpose of our analysis, used perspective and set of players. Analytical methods are not very helpful in answering these policy issues. The second problem is with our intuitive definition of the role or position. It works only if we have a collection of actors and sets of relationships that are meaningful to human perception of the problem, such as study specific issues [15].

Same idea of equivalence (with respect to position), or similarity (as somewhat equivalent) must be defined, however, quite precise. Again there is no definite and clear procedure for such a definition. But there are several approaches for relatively accurate analysis of what it means to be similar to actually examine the data and define social roles or social positions empirically. What does it mean that two players have “similar” profiles (patterns) relations, and thus are both members of the same role or social status? Analysis of networks usually defines two nodes (or other complex structures) to be similar when they fall into the same “equivalence class”.

Usually we need to specify why two actors (or other structures) are members of the same class, which and how is different from other classes. The question remains, which features

(structural, compositional, temporal attributes) of actor positions are essential for assigning the actor to an equivalence class and in what way they are equal.

There are many ways to define equivalence of actors based on their relationships with others. For example, we could create two classes of equivalence of actors with zero outdegree and outdegree of actors with more than zero. In fact a very large number of algorithms, that analyze a collection of actors using categories, determine their category on the basis of certain pattern of their position in the graphs.

One of the used approaches is to simplify the analysis of positional information in the data network. This simplification is generally made up of representations of the network using the term (name) of positions and the mathematical expressions defining the relations of these positions among themselves. Positions are usually defined by the equivalence of actors. Reorganization of the same data can create blocks or clusters in networks [1].

III. UNCONVENTIONAL DATA IN STRUCTURAL ANALYSIS

As main part of the network data we consider structural data. But except structural variables there are also compositional and temporal variables. Structural data are obtained from the structural links between the actors of underlying social network. This information represents specific ties, e.g. business transactions between companies, friendship between people or trade between countries.

Compositional data represent the attributes of the actors. These are data of individual objects, e.g. gender, ethnicity, geographic location, number of employees etc. Apart from the direct attributes (compositional data) derived attributes of actors can be used (for example, some computed data from compositional and temporal attributes) in the network. In addition to these static data it is possible to learn more about the dynamic behavior of data during the life of the network (temporal data). Dynamic analysis of social networks is trying to describe the dynamics of probabilistic values. The problem of the expansion of data sometimes does not allow to store and process the full historical data of large social networks. Here arises the possibility of using the attributes describing the dynamic properties of networks.

Temporal data (time data) are either part of or complement to traditional network data. They can be integrated into the network data as a part by adding one dimension (to each attribute of actor, or actors have represented their history in some way) or as a complement in the form of transactions representing any change in the network and its phasing over the information (tracking of the changes). These data are very interesting in terms of dynamics and its analysis. Mostly, data for social network analysis come in the form of static data (almost no dynamic data) they are collected manually in the form of surveys, questionnaires and other methods. Such data are difficult to obtain, therefore there are only a few published datasets and also these are of questionable accuracy.

Concurrent presence of all of these types of data brought us to an idea of “complete data model” construction. All structural, compositional and temporal data are present in this data model [3]. By intuition we can presume, that these data have some kind of correlation and some of the attributes may depend on each other in some way. Compositional and structural data are properties of individual actors and sets of actors in structural positions indicate similarities in compositional data. This idea supports also recent research in paradigm data mining - relational classification. This classification uses attribute-oriented classification and group of classification methods supported by information about structural ties. Due to the added information this methods are able to classify more precisely [4].

IV. COMMUNITY DETECTION

Community detection is a fundamental problem in the analysis of social networks [6]. A “network community” (also sometimes referred to as a module or cluster) is often defined as a group of network members with stronger ties to members within the group than to members outside of the group [7]. In other words a community is a densely connected subset of nodes that is only sparsely linked to the remaining network [9], [10]. Modular structure introduces important heterogeneities in complex networks. For example, each module can have different local statistics; some modules may have many connections, while other modules may be sparse [9].

A number of recent results suggest that networks can have properties at the community level that are quite different from their properties at the level or the entire network, so the analyses that focus on the whole network and ignore community structure may miss many interesting features [11]. For instance, in some social networks one finds individuals with different mean number of contacts in different groups; the individuals in one group might be gregarious,

having many contacts with others, while the individuals in another group might be more reticent. Detection of communities and important persons within communities in collaboration network described below should lead up to detection of unknown and potentially important actors of particular communities.

V. DAK COLLABORATION NETWORK

A. Data description

DAK – community network is a collaboration network of members of a non-profit organization dealing with organizing educative-pedagogic workshops for young people [12]. Usually there organized around 10 workshops annually with 150 – 700 participants on single workshop. The count of participants depends on workshop type and duration. All participants of a single workshop are partitioned into smaller groups, usually with 8 – 12 members. Each group member cooperates with each other group member and so there are established new social relations between group members. Generally there are two types of group members – *group participants* and *leader(s) of a group*. Participants are “isolated” inside group (usually they do not establish social relations outside group), but leader(s) cooperate with other leaders and so they create another social relations. Additionally we recognize two types of groups – *participants group* and *organizers group*. In summary we have two types of groups and four types of group members: *base participant*, *base organizer*, *leader of participants group* and *leader of organizers group*.

B. Accessible compositional and temporal attributes

Compositional attributes in DAK data set are available for both, actors and events. Actors are described by attributes such as date of birth (age) or gender; and by geographical attributes – city or area of living. Events are described by their type. We can recognize two main types of events – events for base participants and events for organizers. Events for organizers are next categorized by their types of activity like registration, security or accommodation event.

Moreover, temporal attributes are available together with compositional attributes, e.g. start and end date of events. From these data we can derive several other attributes such as “length of event” or “time of first visit” for particular actor. In our case it means moment when the actor visited any event for the first time.

C. Data structure and transformation

We analyzed and preprocessed obtained data of DAK. As first we cleaned the data and transformed them from single data matrix into three separate matrixes: actors, events and cross data matrix which connects actors and events. Actors are all members of organization above who participated at least in one event. Each group on the single workshop we represent as a single event, so for one workshop we obtain several events. Additionally we added two more events for representation of cooperation between leaders of participants and leaders of organizers.

This real data of DAK collaboration network can be expressed for single workshop by a bipartite graph as representation of two-mode network (see Fig. 1). The first mode is a set of actors, and the second mode is a set of events which affiliate the actors. One of the advantages of DAK data set is availability of temporal information in the data. We are able to track events in the time line and recognize which events were organized in parallel and which sequentially. Also we are able to track participation of single actors on particular events.

VI. FUTURE WORK

A. Clusterisation and positions in community network

On Fig. 1 we have a network presenting a single workshop full of events. Events are represented by rectangles and actors by circles. The color of actors separates them by attendance to events. It only shows, if actor attended to single event (maroon) or multiple events (green) and has for our purpose no more important meaning, but is very helpful in understanding of positions. It simply distinguishes more and less important actors. To form several positions in our network we use an idea of a generalized “local role equivalence”. Practically it means, that two actors or events are in the same position if they are similarly connected to other actors in particular positions. Using this definition we can create seven positions (four positions for actors and three for events). Definition of A_1 position is quite simple: actor in this position is attached to exactly one event in position E_1 . Position A_2 is formed of actors, who are connected precisely to one event in position E_2 . Let us call these

actors (A_2) as ordinary actors and events (E_2) as ordinary events. Actors in position A_3 are connected to one ordinary event (E_2) and one event in position E_1 . This actor (A_3) could be leader of ordinary events (E_2) and then participates in their common “leader’s” event (E_1). On the figure we have two leader’s events and they are connected through an interesting position of actors A_4 . This two actors (A_4) participate in multiple events of E_2 , but do not participate in ordinary events. This position along with A_1 could be some kind of organizers or technical stuff. As we create an image of positions, we can obtain a hierarchy of actors and events in one workshop.

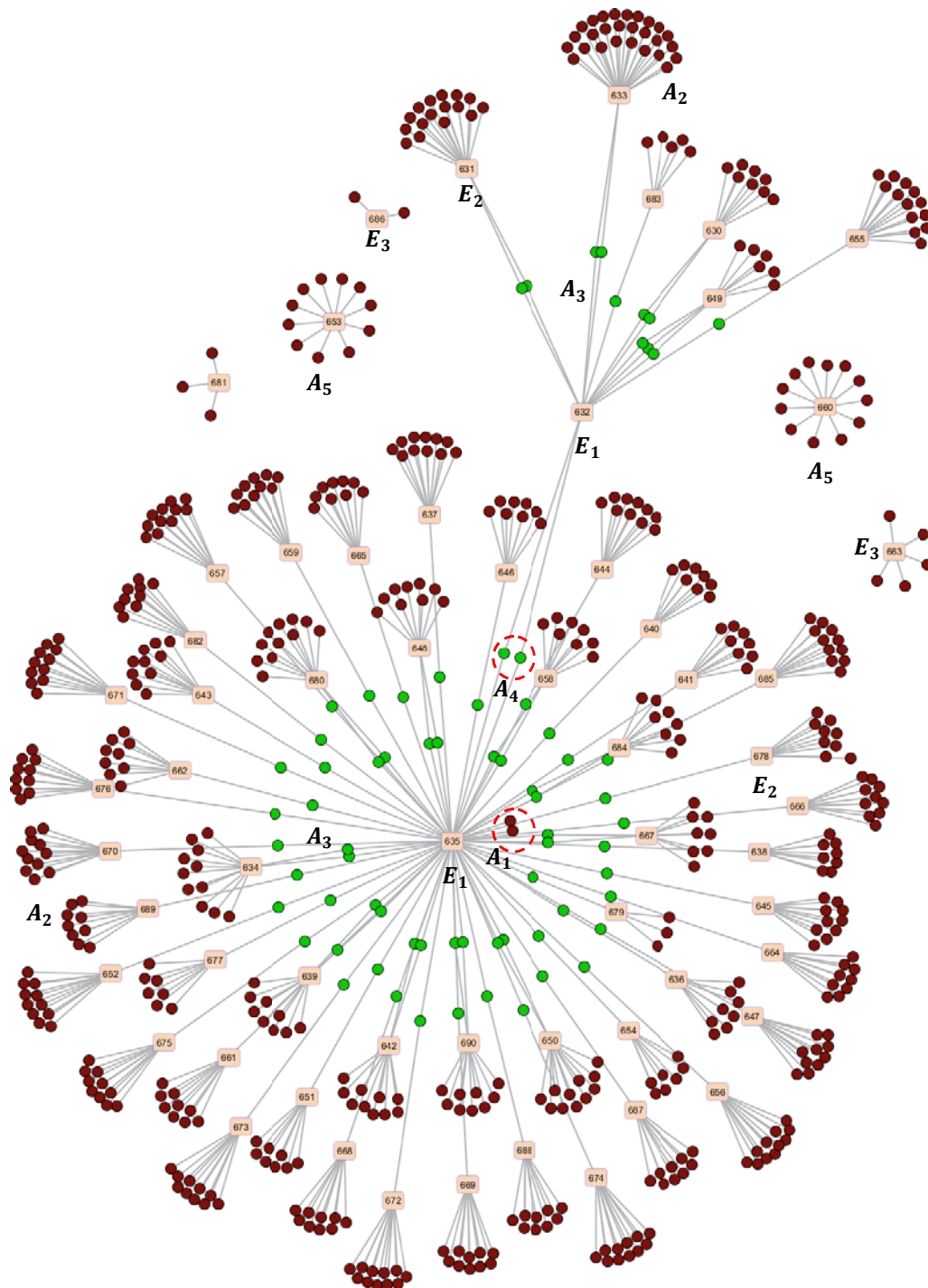


Fig. 1 DAK collaboration network of a single workshop - analysis of positions

To create an intuitive definition “role equivalence” of actors and events is quite simple if we have only one workshop. Considering full DAK community network (containing more than 20 large workshops) it is not very easy to define all positions in network. Now the process of positional analysis can be performed in four steps. *Definition of equivalence (1)* is to be done using two-mode data. Mainstream clusterisation methods are using one-mode data and use of

two-mode data brings a lot of problems to create simple representation. Next step is to define *measure of equivalence* (2), which practically means to define some similarity between actors in positions. After these tasks we create a *representation of assignment* (3) of position to actor and *adequacy of equivalence* (4) is examined [1].

This process is iterative, we find out some equivalence then we test it and see the results. If equivalence is not adequate, we have to do some changes in definition of equivalence (or in other steps of process) and try it again. Our ambition is to create a semi-automatized methodology for positional analysis.

B. Improving positional analysis using unstructured data

Analysis of positions in collaboration network could be improved, if supported by compositional or temporal attributes. Research has proven that compositional data have some correlation with structural data [4]. In the following work positional analysis will be enriched with above mentioned “unconventional data”. The modified methodology for positional analysis will be based on complete data model. Compositional data will be helpful in the process of finding positions in the network. There will be designed a more detailed representation, performed the technical implementation of the complete data model, and methodology for the positional analysis with complete data will be designed as well.

C. Temporally based community detection

Previous studies on the community structure of social networks have typically focused on static networks. With temporal data we are able to analyzing dynamic network in time series, so the natural extension of community detection to dynamic networks is community tracking, which makes it possible to observe how communities grow, shrink, merge, or split with time [7].

In [7] Xu et al. proposed a method for tracking communities in dynamic social networks. Their method makes use of an evolutionary clustering framework that detects communities at each time step using an adaptively weighted combination of current and historical data. The method incorporates temporal smoothing to stabilize the variation of communities over time. They were able to track communities where the members were continually changing or perhaps assuming multiple identities, which suggests that the proposed method may be a valuable tool for tracking communities in networks of illegal activity.

Greene et al. in [8] described an algorithm for tracking of communities across time steps. They discussed question how best to map step communities at each time to existing set of dynamic communities.

In our future work we will perform community detection for DAK collaboration network in all time steps of network evolution. We will track communities and analyze positions and roles inside detected communities. Then we will be able to track particular actors and changes in their roles with evolution of the whole network.

VII. CONCLUSIONS

In this publication we introduced two approaches to structural analysis of social groups in collaboration network. We discussed a possibility of improved positional analysis with the use of compositional and temporal data. Then we discussed another approach of analysis – community detection. These approaches have some dissimilarities. Positional analysis can be performed at two-mode data, which allows us to preserve duality of actors and events [5]. On the other hand community detection uses one-mode data. This data needs to be transformed from two-mode data and this process is quite difficult, if we want to preserve as much as possible of original information. The advantage of community detection is that properties at the community level could be quite different from properties at the level of the entire network. Analyses that focus on community structure can bring more precise results in positional analyses of the network.

ACKNOWLEDGMENT

The work presented in this paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0042/10 (50%). This work is also the result of the project implementation Development of Centre of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Programme funded by the ERDF (50%).

REFERENCES

- [1] S. Wasserman, K. Faust, "Social Network Analysis," Cambridge University Press, 1994. ISBN: 978-0-521-38707-1
- [2] J. Scott, "Social Network Analysis" (A handbook, second edition). SAGE publications, 2000. ISBN: 978-0-7619-6339-4
- [3] M. Repka, "Analysis of specific types of social networks," Written part of dissertation exam. Technical university of Košice, Košice, 2011.
- [4] P. Vojtek, "Príspevok k relačnej klasifikácii s využitím predpokladu homofilie," Report of solution in dissertation project V, Institute of Informatics and Information Technologies. Bratislava 2010.
- [5] S. Field, K. A. Frank, "Identifying Positions from Affiliation Networks: Preserving the Duality of People and Events," on-line at <<https://www.msu.edu/~kenfrank/papers/identifying%20positions%20from%20affiliation%20networks.pdf>>
- [6] M. E. J. Newman, "Modularity and community structure in networks," in *Proc. of the Nation Academy of Sciences* 103(23), 8577-8582, 2006.
- [7] K. S. Xu, M. Kliger, A. O. Hero, „Tracking Communities in Dynamic Social Networks“ on-line at <http://www.eecs.umich.edu/~xukevin/xu_evolution_clust_sbp_2011.pdf>
- [8] D. Greene, D. Doyle, p. Cunningham, "Tracking the Evolution of Communities in Dynamic Social Networks," in *Proc. of the International Conference on Advances in Social Networks Analysis and Mining*, 2010, on-line at <<http://www.cliquecluster.org/system/files/greene10tracking.pdf>>
- [9] N. Gulbahce, S. Lehman, "The art of community detection," Wiley Periodicals, Inc. 2008 on-line at < http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200809-17_BioEssays-CommunityDetection/200809-17_BioEssays-CommunityDetection.pdf>
- [10] J. Leskovec, K. J. Lang, M. W. Mahoney, „Empirical Comparison of Algorithms for Network Community Detection," in *Proc. of the WWW 2010*, Raleigh, North Carolina, USA.
- [11] M. E. J. Newman, "Finding community structure in networks using the eigenvectors," University of Michigan, 2006.
- [12] DAK – Collaboration network, data set of non-profit organization, on-line at www.domcek.org
- [13] R. A. Hanneman, M. Riddler, M., "Introduction to social network methods." Riverside, CA, University of California, Riverside. 2005, on-line at <<http://faculty.ucr.edu/~hanneman/nettext/>>
- [14] K. A. Frank, "Identifying cohesive subgroups." *Social Networks*, 17, 27-56. 1995.
- [15] H. T. Welser, E. Gleave, D. Fisher, M. Smith, "Visualizing the Signatures of Social Roles in Online Discussion Groups." *Journal of Social Structure* Vol.8 No.2. 2007

Využitie agentových simulácií v socioekonomickej analýze a ich možné rozšírenia

¹Peter BUTKA, ²Martin SARNOVSKÝ

¹Centrum pre informačné technológie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

²Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

peter.butka@tuke.sk, martin.sarnovsky@tuke.sk

Abstrakt — V tomto príspevku sa venujeme základným aspektom (multi)agentových simulácií a ich využitiu v rôznych analýzach z dôrazom na socioekonomické aplikácie, ako aj možným rozšíreniam štandardného prístupu v rámci našej ďalšej práce. Multiagentové simulácie poskytujú oblasti tzv. sociálnych agentových simulácií (Agent-Based Social Simulation) možnosť analyzovať komplexitu a emergenciu socioekonomických modelov vytvorenú vďaka kombinácii veľkého počtu jednoduchých prvkov (agentov), a študovať tak javy, ktoré klasickým prístupom nie je možné (jednoducho alebo vôbec) popísať v analytickej forme. Pre tento účel popíšeme prehľad príkladov využitia tohto prístupu. Následne sa venujeme aj možnosti rozšírenia klasického prístupu modelovania a simulácie v rámci znalostného modelovania a distribuovania simulačných výpočtov.

Kľúčové slová — (multi)agentové simulácie, socioekonomická analýza, znalostné modelovanie, distribuované simulácie

I. ÚVOD

V praxi existuje mnoho príkladov systémov prvkov, ktoré vykazujú (občas nečakanú) komplexitu a emergentné správanie na základe pomerne jednoduchých rozhodovacích pravidiel. Jedným z príkladov je skúmanie sociálnych a ekonomických fenoménov pohľadom "zdola-nahor", a teda sledovanie dynamiky systémov zložených z (často veľkého počtu) pomerne jednoduchých elementov pri rôznych nastaveniach ich vlastností a prostredia, pričom ich považujeme za "tvorcov" správania systému a ich súbor tvorí emergentne vytvorený "model" zvoleného problému. Snahou je teda študovať vývoj týchto modelov v čase a jeho zmeny s cieľom porozumieť problému, a to aj v prípadoch, kedy by tvorba modelu "zhora-nadol" (odpovedajúca napr. ekonometrickým modelom, či všeobecne makromodelom) bola zložitá alebo dokonca nemožná. Logickým prostriedkom takéhoto skúmania je prepojenie modelovania a simulácie s agentovými technológiami. Cieľom multiagentových sociálnych simulácií je potom popísať jednotlivé prvky systému a ich správanie v reči agentov, ich spustenie v simulačnom prostredí, ako aj analýza výsledkov a priebehu simulácií.

Cieľom tohto článku je popísať základné pojmy oblasti agentových simulácií (v kapitole II), špeciálne najmä oblasť (multi)agentových sociálnych simulácií (ABSS - Agent-Based Social Simulation), ako aj podať prehľad použitia simulácií v rôznych oblastiach socioekonomickej analýzy (kapitola III). V závere sú spomenuté niektoré možnosti na rozšírenia simulácií z pohľadu použitia znalostných technológií a distribuovaných výpočtov.

II. AGENTOVÉ SIMULÁCIE A OBLASŤ ABSS

Klasická definícia z oblasti umelej inteligencie autorov Norwig a Russel popisuje agenta ako "čokoľvek čo vníma svoje okolie pomocou senzorov a koná na základe toho pomocou aktuátorov" [7]. Ďalšia a často používaná definícia pochádza od autorov Wooldridge a Jennings [9]. Títo rozlišujú dva typy agentov: slabý a silný. Slabý agent predstavuje hardvérový alebo softvérový počítačový systém, ktorý je autonómny (koná bez cudzieho zásahu, kontroluje svoj

stav a akcie), sociálny (komunikuje s ďalšími agentmi pomocou nejakého jazyka), reaktívny (vníma svoje okolie a vo vhodnom čase reaguje na zmeny) a proaktívny (môže mať cieľovo orientovanú odozvu). Silný agent dopĺňa tieto vlastnosti o ďalšie vystihujúce mentálne a emocionálne stavy. Je pochopiteľné, že využitie slabý agent predstavuje rozšírenejší prípad v praxi. Existujú však rôzne iné typy vlastností, ktoré vedú k inak pomenovaným typom agentov, ako spolupracujúci agenti (kooperujúci), agenti rozhrania, hybridný agent, deliberatívny agent, a pod. Celkový pohľad na rôzne druhy agentov podrobne popisuje oblasť tzv. typológie agentov zameraná na rôzne subtypy podľa kombinácie vlastností autonómnosti, kooperácie a schopnosti učenia sa agentov.

Ďalším dôležitým aspektom je *prostredie*, ktoré predstavuje všetko to, s čím agent prichádza do styku počas svojej činnosti. Prostredie z hľadiska agenta môžeme rozdeliť na plne pozorovateľné (agent môže senzormi sledovať jeho kompletný stav) alebo čiastočne pozorovateľné (agent nie je schopný sledovať všetky stavy prostredia). Prostredie je pre agenta statické, ak sa môže meniť iba jeho akciami. Prostredie je pre agenta dynamické, ak sa môže meniť aj bez zásahu agenta. Prostredie je deterministické ak je jeho stav po vykonaní nejakej akcie daný iba touto akciou a predchádzajúcim stavom tohto prostredia. Prostredie je diskrétné vtedy ak má konečne alebo spočítne mnoho stavov. Pre agenta založeného na číslcových počítačoch je každé prostredie diskrétné. Prostredie je spojité, ak má nekonečný počet stavov.

V multiagentových systémoch je zoskupená množina agentov za určitým cieľom, medzi ktorými sú definované konkrétne vzťahy a väzby. Každý agent má určité prostriedky, s ktorými môže disponovať a zároveň aj obmedzenia, v hraniciach ktorých sa musí pohybovať pri dosiahnutí cieľa. Každý z agentov sa snaží o čo najkvalitnejšie uskutočnenie danej úlohy, čo sa v konečnom dôsledku prejaví v celkovom riešení. V takejto "spoločnosti agentov", agenti spoločne koordinujú svoje schopnosti, znalosti, ciele a plány za účelom vykonania nejakej akcie alebo kolektívneho vyriešenia nejakého problému. Títo agenti môžu pracovať buď spoločne na jednej úlohe alebo každý samostatne na inej alebo rovnakej úlohe. Multiagentová *simulácia* spočíva vo využití tzv. agentov, ktorí predstavujú autonómne entity s relatívne jednoduchým chovaním, reprezentujúce reálne jednotky sledovaného systému, situované do definovaného kontextu prostredia, v ktorom jednájú a reagujú.

ABSS (Agent-Based Social Simulation) je vedecká oblasť zaoberajúca sa simuláciami sociálnych a socioekonomických fenoménov pomocou multiagentových modelov. Ľubovoľný jedinec (reprezentujúci napr. človeka, ale aj firmu, zákazníka, atď.) je modelovaný ako agent, implementovaný zväčša ako softvérový agent v simulačnom prostredí. Cieľom skúmania pomocou ABSS z praktického pohľadu multiagentových simulácií je podľa [5] najmä *porozumieť* študovanému javu a *predikovať* konkrétne hodnoty dát modelu pre budúci vývoj analyzovaného systému. Klasické prístupy sociálnych simulácií sa zameriavajú najmä na skúmanie samoorganizujúcich sa procesov vývoja kooperácie, konfliktov, trhov, dynamiky dopravy, či vývoja sociálnych podskupín.

Praktická realizácia simulácií zahŕňa proces modelovania a nástroje používané na modelovanie a simuláciu. Proces modelovania sa skladá z formulácie riešeného problému, výberu prostredia, identifikácie agentov a ich charakteristík, implementácie modelu, experimentovania, verifikácie modelu a analýzy výsledkov. Ako nástroje používané pre tieto potreby uveďme niekoľko príkladov. Jednoduchý, praktický a názorný (vhodný najmä pre vzdelávacie účely, s dobrou a obširnou dokumentáciou jazyka pre popis agentov a existujúcich modelov) je Netlogo (<http://ccl.northwestern.edu/netlogo/>). Rýchly a populárny nástroj pre tvorbu multiagentových systémov s voliteľnou sadu vizualizačných nástrojov v oblasti 2D a 3D predstavujú MASON (<http://cs.gmu.edu/~eclab/projects/mason/>). Výborné prostredie pre tvorbu modelov vo viacerých jazykoch (C++, Java, Groovy) so silným vývojovým prostredím poskytuje Repast (existuje v rôznych verziách, <http://repast.sourceforge.net/>). Voľne dostupné Java prostredie pre vývoj multiagentových systémov vo všeobecnejšom ponímaní poskytuje JADE (<http://jade.tilab.com/>). Mnohé z týchto nástrojov poskytujú integrované prostriedky pre rýchly návrh a vývoj simulačných modelov, ako aj ich samotné spúšťanie a analýzu dosiahnutých výsledkov.

III. PREHĽAD POUŽITIA SIMULÁCIÍ V SOCIOEKONOMICKÝCH ANALÝZACH

Agentové simulácie majú uplatnenie v rôznych oblastiach života. Využitie môžeme nájsť v ekonomike, medicíne, biológii, spoločnosti, kultúre, náboženstve, armáde, infraštruktúre, biznise a riadení spoločnosti, pri fyzikálnych a chemických javoch, ale aj pri modelovaní evakuačných simulácií. V oblasti ABSS je snahou preniesť pozornosť od technického aspektu modelov ku sledovaniu a vyhodnocovaniu modelov ako celku, nakoľko komplexná dynamika

môže vzniknúť aj vďaka agentom riadeným veľmi jednoduchými pravidlami (emergencia). Vhodne navrhnutý model, ktorý je zjednodušením reality, môže odhaliť a následne zovšeobecniť vybranú vlastnosť chovania systému. Multiagentové modely v oblasti ABSS pomáhajú vďaka tomu porozumieť procesom ako evolúcia kooperatívneho a koordinovaného správania sa, vznik koalícií, šírenie inovácií, alebo tvorba trhových cien.

V [1] autor identifikuje tri oblasti pre využitie ABSS v predikcii konkrétnych systémov, a to modelovanie:

- *tokov* – modelovanie dopravy a evakuácia osôb,
- *trhov* – modelovanie burzy a identifikácia aspektov nestability, zmeny regulačných parametrov,
- *organizácií* – snaží sa riešiť dynamiku závislosti jednotlivcov, ako napr. modelovanie operačného rizika (chyby zamestnancov) a vzťahov ľudí.

Vo všeobecnosti platí, že multiagentové sociálne simulácie sa hodia najmä v prípadoch, kde populácia agentov je rôznorodá, ich správanie sa a interakcie sú komplexné, nelineárne alebo nespojité, pričom významnú úlohu hrajú aj priestorová a sociálna štruktúra agentov. Uvedieme si teraz niekoľko príkladov ako prehľad využitia agentových simulácií v rôznych socioekonomických analýzach.

A. Maximalizácia zisku

Klasická úloha maximalizácie zisku firmy výberom optimálnej stratégie pri uspokojení požiadaviek zákazníkov, kde veľkú úlohu zohráva kapacita výroby. Štandardne je možné riešiť takéto úlohy napríklad pomocou lineárneho programovania. Môžeme sa však na tento problém pozeráť aj z pohľadu multiagentového modelu. Podľa [3] tu vystupujú dva typy agentov: $A = \{A_1, A_2, \dots, A_i\}$ – zákazníci, $B = \{B_1, B_2, \dots, B_j\}$ – výrobcovia. Agent A_i má požiadavku na výrobu n kusov výrobkov, ktorú pošle v čase t_1 agentovi B_j . Tento agent buď môže splniť požiadavku, alebo nemôže, pretože jeho voľná výrobná kapacita m kusov výrobkov je menšia ako požiadavka na n výrobkov. Ak nemôže splniť požiadavku odošle agentovi A_i zamietnutie požiadavky a informáciu o svojej voľnej kapacite m . Na základe tejto informácie môže agent A_i buď znížiť svoju požiadavku, alebo sa pokúsiť o realizovanie svojej požiadavky u iného agenta typu B v čase t_2 , alebo môže realizovať svoju požiadavku objednaním výrobkov u dvoch agentov, alebo viacerých agentov, prípadne prehodnotí svoju požiadavku a zníži objednávku. V prípade, že agent typu B predstavuje pružný výrobný systém, je možné po úpravách zvýšiť kapacitu výroby a tak vyhovieť agentovi A_i . Predpokladajme, že agenti typu A majú informácie o agentoch typu B a agenti typu B majú informácie o agentoch typu A ohľadom cien výrobkov a výrobných kapacít. V praxi sú tieto ceny predmetom zmlúv medzi partnermi a teda cena, za ktorú kúpi výrobok jeden odoberateľ je iná pre iného odoberateľa od toho istého výrobcu. Na formovanie tejto ceny má vplyv množstvo faktorov. Za predpokladu, že poznáme všetky faktory, môžeme odhadnúť, ako sa budú jednotliví agenti správať. Vďaka tomu je tento model flexibilný a umožňuje skúmať aj nelineárne vzťahy nad rámec klasickej úlohy.

B. Predaj komodít a určenie ceny

Tento ekonomický model obsahuje agentov, ktorí budú predávať určité množstvo výrobkov denne. V [8] autori navrhli model využívajúci genetický algoritmus GALCS na základe ktorého agent určí: 1) či sa ceny komodít v poslednom čase zvýšili alebo znížili, 2) či boli predaje v poslednom čase rastúce alebo klesajúce, 3) či boli zisky v poslednej dobe rastúce alebo klesajúce, 4) či sú ceny vyššie alebo nižšie, než je priemer v odvetví. Na základe odpovedí na jednotlivé body potom agent priradí pravdepodobnostný vektor toho, či sa má cena zvýšiť, znížiť, alebo zostať konštantná. Po vstupe do určitého stavu, sa agent rozhodne ako zmeniť danú cenu pomocou zodpovedajúcej pravdepodobnosti vektora a vyberie náhodné číslo. Agent potom upraví vektor podľa toho, ako zmena ceny ovplyvnila zisky.

C. Finančný trh

V tomto príklade každý agent predstavuje účastníka trhu. Agenti môžu sledovať aktuálne informácie o vývoji trhu, chovanie agentov v okolí a chovať sa náhodne. V rámci modelu sa obchoduje len s jedným aktívom, agenti sú reaktívni, nemajú finančné obmedzenie, obchodujú v každom kroku s rovnakým množstvom aktív. V každom kole sa agent na základe získaných správ chovania kolegov v minulom kole a náhodného faktoru rozhoduje, či bude predávať, či kupovať aktíva alebo nebude nič robiť. Transakčné náklady sú modelované parametrom, ktorý

ovplyvňuje, či sa agent zdrží aktivity. Zvýšenie transakčných nákladov vedie k zníženiu nestability trhu a objemu obchodu. Mení sa pomer medzi dopytom a ponukou. Cieľom je potom sledovať objem obchodu, nestabilitu trhu (smerodajnú odchýlku zisku a straty), minimum a maximum zisku a straty, a rovnováhu medzi dopytom a ponukou. Príklad použitia agentových simulácií na finančné trhy je možné nájsť v [2].

D. Producent a konzument

Je to skupina ekonomických modelov, v ktorých producenti čerpajú zdroje a produkty, aby mohli vytvárať iné produkty, ktoré ukladajú späť do prostredia. Mohli by sme ich rozdeliť na dva druhy. Na primárnu výrobu, ktorá na svoju činnosť potrebuje iba zdroje a prácu, a na sekundárnu výrobu, ktorá ešte potrebuje navyše aj výstupy iných firiem. Oproti producentom stoja konzumenti, ktorí čerpajú zdroje a produkty z prostredia. Všetci títo agenti žijú nad prostredím, kde ukladajú, čerpajú všetky produkty, zdroje. Producenti majú istú maximálnu a minimálnu produkciu, vyrábať môžu iba v týchto hraniciach. Tento model umožňuje vybudovať hierarchiu trhu (ponuky a dopytu), najmä tzv. dodávateľské reťazce, a skúmať ako sa bude systém vyvíjať [6].

E. Vážňova dilema

Tento príklad patrí do oblasti skúmania dôsledkov teórie hier, ktorá sa uplatňuje v mnohých oblastiach ľudskej činnosti od ekonómie cez politológiu, až po sociológiu či biológiu. Teória hier je disciplína aplikovanej matematiky, ktorá analyzuje široké spektrum konfliktných rozhodovacích situácií, ktoré môžu nastať kdekoľvek, kde dochádza k stretu záujmov. V tomto klasickom prípade sa do rúk polície dostali dve rôzne osoby, ktoré sú držané oddelene. Dôkazy, ktoré má polícia, nie sú dostatočné pre usvedčenie, takže sa musí spoliehať na priznanie respektíve udanie. Pokiaľ sa obidvaja navzájom udajú, budú odsúdení na 5 rokov. Pokiaľ jeden udá druhého a druhý ostane mlčať, bude udavač voľný a druhý odsúdený na 10 rokov. Pokiaľ obidvaja ostanú mlčať, odsúdia obidvoch za drobné priestupky na 2 roky. Vzhľadom k tomu, že ani jeden zadržaný si nemôže byť istý, čo zvolí ten druhý, nastáva dilema: hovoriť alebo mlčať? Ide o klasickú úlohu a jej model je možné skúmať (ako aj iné podobné úlohy rozhodovania z oblasti teórie hier) pomocou metód ABSS.

F. Evolúcia etnocentrity

Model evolúcie etnocentrity je modifikáciou klasického modelu vážňovej dilemy. Agenti sú v prostredí rozdelení do niekoľkých etnických skupín, skupiny sú reprezentované celým číslom, graficky a farbou. Každý agent má schopnosť rozpoznať, či susedný agent je príslušníkom vlastného alebo cudzieho etnika. Agenti sa môžu s istou pravdepodobnosťou v každom kole reprodukovať a prenášajú na svojich potomkov príslušnosť k etnickej skupine a stratégiu chovania. Je daná aj určitá pravdepodobnosť úmrtia. Potomok môže byť umiestnený len na voľné susedné pole agenta. Interakcie medzi agentmi sú lokálne a takisto sú lokálne aj kompetencie medzi agentmi o vzácne zdroje, ktorými sú priestor pre nového potomka a pravdepodobnosť reprodukcie agenta. Zvýšenie, či zníženie tejto pravdepodobnosti zároveň znamená zisk, či stratu v hre. V každom kole hrajú agenti so svojimi bezprostrednými susedmi hru typu dilema vážňa. Každý agent získava základnú pravdepodobnosť rozmnoženia. Pokiaľ agent kooperuje s inými agentmi, znamená to preňho isté zmenšenie tejto pravdepodobnosti, ceny kooperácie. Naopak, pokiaľ iný agent kooperuje s ním, získava niekoľkonásobok ceny kooperácie. Skupina vzájomne kooperujúcich agentov vzájomne posilňuje svoju schopnosť reprodukcie. Nekooperujúci agent nič nestráca a môže len získať, pokiaľ s ním susedia kooperujú. Na druhej strane, pokiaľ spolu susedí niekoľko nekooperujúcich agentov, pravdepodobnosť ich reprodukcie sa nijak nezvyšuje.

G. Šírenie vírusu

Tieto modely simulujú prenos a pretrvávajúce vírusu v ľudskej populácii. Biológovia navrhli rad faktorov, ktoré môžu ovplyvniť prežívanie priamo odovzdávaného vírusu v populácii. V praxi model zvyčajne obsahuje určité množstvo ľudí, z ktorých je niekoľko nakazených. Ľudia sa pohybujú náhodne. V tomto svete existujú 3 stavy: ľudia sú chorý, zdravý, ale náchylný k infekcii, alebo zdravý a imúnny. Ľudia v tomto modeli môžu umrieť od staroby alebo infekcie. Potomkovia sa rodia iba zdravým jedincom. Hustota obyvateľstva ovplyvňuje ako často môžu prísť do kontaktu chorí ľudia s ostatnými jedincami. V každom kole, má každý zdravý jedinec šancu na reprodukciu. V rámci modelu sa zvyčajne dá nastaviť stupeň odolnosti

voči vírusu a nákazlivosť. Niektoré vírusy sa z človeka na človeka šíria ľahko, iné len bezprostredným kontaktom, pričom sa dá nastaviť aj doba trvania infekčnosti.

H. Teória davu – evakuačný model

Modely tohto typu sa zameriavajú na zvládnutie evakuácie tzv. teóriou davu. Uvedme si príklad takéhoto modelu – máme návštevníkov múzea. Model je schopný simulovať správanie návštevníka počas normálnych návštevných hodín v múzeu, pozorovať ich reakcie počas náhleho poplachu, pre rôzne stavebné rozvrhnutia, lokalizácie núdzových východov a evakuačných plánov [3]. Model ukazuje aká dôležitá môže byť simulácia pri skúmaní davového správania, ktoré môže ovplyvňovať bezpečnosť. Agentov predstavujú návštevníci, ktorí navštevujú v múzeu rôzne miestnosti s umeleckými dielami. Každý agent je schopný prispôbiť sa okruhu návštevníkov a pohybovať sa k svojim cieľom, ktorými sú buď umelecké diela alebo dvere. Návštevník je naprogramovaný tak aby neprešiel dvakrát cez tie isté dvere a aby nenavštívil dvakrát to isté umelecké dielo. Každý návštevník má ľubovoľný čas na obdivovanie umeleckého diela. V momente, keď sa spustí poplach, každý návštevník skontroluje, či je pri ňom núdzový východ. Ak áno pohne sa tým smerom, inak nasleduje núdzový znak, ktorý vidí alebo v opačnom prípade ide späť tou cestou ktorou prišiel, až kým nenájde núdzový východ. Dĺžka času potrebného na opustenie múzea je rôzna, pretože závisí od množstva návštevníkov a od vzdialenosti jednotlivých návštevníkov od núdzového východu.

Okrem spomenutých príkladov existuje množstvo ďalších zaujímavých modelov, vymenujme len napr. trh s elektrinou, model trhu práce (vzťah zamestnávateľ a žiadateľ o prácu), vytváranie skupín v rámci komunikácie v uzavretej komunite („party“ model), biologicky inšpirované systémy na riešenie rôznych optimalizačných úloh (napr. kolónie mravcov), modelovanie náboženských prvkov v segregácii a sťahovaní ľudí, simulácia rizík vo firme, analýza vývoja urbanizácie a centier osídľovania, a pod.

IV. ZÁVER A MOŽNOSTI ROZŠÍRENIA KLASICKÝCH MODELOV A SIMULÁCIÍ

V príspevku sme uviedli niekoľko príkladov využitia agentových simulácií pre rôzne socioekonomické analýzy a modely. Pre väčšinu popísaných modelov existujú rôzne implementácie v rôznych prostrediach. V našej ďalšej práci sa chceme zamerať na rozšírenia modelov a ich vlastností, ako aj ich výpočtov (simulácií). Z rôznych možností pripadajú do úvahy najmä:

- využitie znalostných technológií (ontológií) v modelovaní a funkčnosti agentov a ich spoločného prostredia,
- využitie distribuovaného počítania v procese simulácie pre analýzu senzibility regulačných parametrov modelov (pomocou grid/cloud technológií),
- návrh a riešenie informačného systému pre zapojenie rôznych typov agentov, vrátane spätnej väzby ľudí ako aktívnych agentov (pre tzv. "role playing games").

V prvom prípade existujú situácie, kedy je možné agentov popísať pomocou špecifických tried, ktoré medzi sebou dedia vlastnosti a tvoria hierarchické štruktúry v zmysle relácií zovšeobecňovania a špecifikácie, prípadne pre nich existujú akcie a stavy, ktoré sú modelované ako ontologický model z možnosťou logického odvodzovania (napríklad sa vyberú pre akciu také funkcie, ktoré zdieľajú len špecificky blízke triedy agentov, ktorých konkrétne inštancie sa získajú práve odvodzovaním). Takisto v prípade presného popisu obsahu komunikácie je potrebné aby agent dostal informácie v správnom formáte, aj keď sa pohybuje v rôznorodom prostredí. V takýchto prípadoch je užitočné použiť ontológiu (znalostný model) pre spresnenie popisu agentov, ich vlastností a vzťahov.

Samotný proces simulácie môže byť náročný pre výpočet, preto existujú možnosti pre distribuovanie výpočtov modelov na rôznych úrovniach. Jednou z nich je využitie väčšieho počtu výpočtových uzlov pre spustenie simulácie pri rôznych počiatočných podmienkach v procese tzv. analýzy citlivosti (senzibility). Jej cieľom je skúmať (podobne ako v prípade iných nelineárnych dynamických systémov) dôsledky zmien počiatočných nastavení regulačných parametrov na výsledok simulácie. V tomto prípade je niekedy problém najmä z určením dosiahnutia výsledku simulácie. Tu môže ísť o stabilný stav (bod alebo periodický stav) alebo nejakú iteratívnu formu určenia konvergencie na základe napr. odchýlky stavu medzi definovanou množinou po sebe nasledujúcich krokov. Celková analýza častí priestoru vstupných parametrov a im odpovedajúcich výsledkov umožňuje zlepšiť úspešnosť predikcie dosiahnutých výsledkov simulácií.

Veľmi zaujímavým aspektom pre oblasť ABSS je kombinovanie softvérových a ľudských agentov v procese simulácie. Znamená to, že aktívny agent (človek) dostane (zvyčajne v určitých krokoch) možnosť určiť stratégiu ďalšieho postupu jemu prislúchajúceho agenta a vložiť tak svoje znalosti o konkrétnom probléme do simulácie. Samotný model pozostáva tak z čisto softvérových agentov, ako aj z agentov s aktívnym rozhodovaním. Konanie aktívneho agenta je zaznamenané a používa sa ako učiaci prvok pre neaktívne (softvérové) rozhodovanie agentov v danej simulácii, ako aj v ďalších experimentálnych spúšťaniach. Je to zaujímavý fenomén sociálnej simulácie známy ako hry, v ktorých človek preberá rolu agenta ("role playing games").

Naším súhrnným cieľom je potom vytvoriť informačný výpočtový systém na báze existujúceho simulačného prostredia (napr. Repast) z využitím existujúcej grid/cloud infraštruktúry (vytvorenej v rámci Centra informačných a komunikačných technológií), v ktorom budú existujúce socioekonomické modely rozšírené o jednotlivé spomenuté prvky - znalostné popisy agentov a prostredia, distribuovanie výpočtov simulácií a analýza citlivosti, zapojenie aktívnych agentov do simulácií a učenie ich softvérových modelov na základe ich konania.

POĎAKOVANIE

Táto práca bola podporená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences* 99(3), Stanford University's Highwire Press, 2002, pp. 7280–7287.
- [2] J. Burian, "Multiagentní model transakčních nákladů na finančních trzích," dostupné online: <http://is.muni.cz/do/1456/soubory/aktivita/mues/12119801/20100407BurianACEFinancniTrhy.pdf>
- [3] S. Capri, C. Garofalo, M. Ignacollo, et. al., "Agent - based simulation of pedestrian behavior," dostupné online: [http://www.cesaregarofalo.it/pdf/Agent - based simulation of pedestrian behaviour.pdf](http://www.cesaregarofalo.it/pdf/Agent-based%20simulation%20of%20pedestrian%20behaviour.pdf)
- [4] B. Frankovič, I. Budinská, T.T. Dang, "Multiagentové systémy v plánovaní a rozvrhovaní výroby," *AT&P Journal*, zvláštne vydanie, 2001, pp. 3-8.
- [5] N. Gilbert, K. Troitzsch, *Simulation for the social scientists*. 2nd edition. Open University Press, 2005.
- [6] C. Macal, M. North, "Tutorial on agent-based modeling and simulation part 2: how to model with agents," *Proceeding of WSC '06*, 2006.
- [7] S. Russell, P. Norvig, *Artificial Intelligence: A modern approach Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
- [8] D. Schoenwald, D. Barton, M. Ehlen, "An Agent-Based Simulation Laboratory for Economics and Infrastructure Interdependency," *Proceedings of American Control Conference*, 2004, pp. 1295-1300.
- [9] M. Wooldridge, N. R. Jennings, "Intelligent Agents: Theory and Practice," *Knowledge Engineering Review* 10(2), 1995, pp. 115–152.

Analýza množín textových dokumentov v distribuovaných prostrediach

¹Martin SARNOVSKÝ, ²Peter BUTKA, ³Peter BEDNÁR

¹Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

^{2,3}Centrum pre informačné technológie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

`martin.sarnovsky@tuke.sk, peter.butka@tuke.sk, peter.bednar@tuke.sk`

Abstrakt — Narastajúci objem dát predstavuje veľký problém, pretože pri využití klasických výpočtových prostriedkov môže riešenie úloh na takýchto dátach trvať neúnosne dlho. Vhodným riešením tohto problému sa javí byť distribúcia algoritmov a dát s využitím technológií gridu a cloudu. Ich výhoda spočíva v tom, že pre ich nasadenie je možné využiť už existujúce výpočtové kapacity a infraštruktúru (napr. Internet). Technológie pre distribuované objavovanie znalostí sú obzvlášť vhodné pre aplikácie, ktoré zvyčajne spracúvajú veľké objemy dát alebo roztrúsené dátové zdroje (digitálne knižnice, vedecké simulácie, obchodné dáta...), ktoré nie je možné analyzovať v prijateľnom čase na klasických počítačoch. Cloud v sebe integruje distribuované a paralelné počítanie, teda predstavuje vhodný nástroj pre distribuovanú analýzu dát.

Kľúčové slová — cloud computing, gridové počítanie, dolovanie v textoch

I. ÚVOD

V súčasnosti riešenie zložitých či rozsiahlych úloh, kde nestačia klasické prístupy predstavuje závažný problém. Vyriešenie takýchto úloh je príliš časovo náročné, keď je úloha príliš zložitá či rozsiahla na sekvenčné spracovanie. Na druhej strane prostriedky distribuovaného počítania poskytujú platformu, kde je všetko úplne ináč, všetko je rýchlejšie a čas potrebný na riešenie takejto úlohy, tým že úlohu rozdelíme na čiastkové podúlohy, ktoré sa budú riešiť súbežne na viacerých uzloch, je výrazne kratší. Dnešným fenoménom je cloud computing, ktorý je formou distribuovaného počítania.

Cloud computing je model umožňujúci pohodlný sieťový prístup (na vyžiadanie), k zdieľaným konfigurovateľným výpočtovým zdrojom (napr. siete, servere, úložisko dát, aplikácie a služby), ktoré možno rýchlo upravovať a sprístupniť s minimálnym úsilím riadenia či zásahov poskytovateľa služieb. Tento model propaguje dostupnosť a je postavený na troch modeloch služieb a štyroch modeloch rozloženia [1]. Cloud computing môžeme chápať ako dodávanie škálovateľných IT zdrojov prostredníctvom internetu. Pod modelmi služieb rozumieme tri rôzne koncepty a to *softvér ako služba* (SaaS), *platforma ako služba* (PaaS) a *infraštruktúra ako služba* (IaaS). V oblasti modelov rozloženia môžeme hovoriť o *cloud* verejnom, komunitnom, súkromnom a hybridnom.

Táto práca sa zaoberá problematikou dolovania v textoch v distribuovanom prostredí. Hlavným problémom je vysoká časová náročnosť klasického sekvenčného riešenia pri veľkom množstve dát. Z tohto hľadiska je vhodné využiť práve distribuované počítanie na riešenie tejto úlohy. V nasledujúcej kapitole bude popísaný prehľad implementovaných riešení v distribuovanom prostredí výpočtového gridu, pričom ďalej budú predstavené aktivity smerujúce k adaptácii zvolených algoritmov do cloudového prostredia využitím open-source cloudového rámca Gridgain.

II. PREHĽAD IMPLEMENTOVANÝCH RIEŠENÍ PRE ANALÝZU MNOŽÍN DOKUMENTOV V DISTRIBUTIVANÝCH PROSTREDIACH

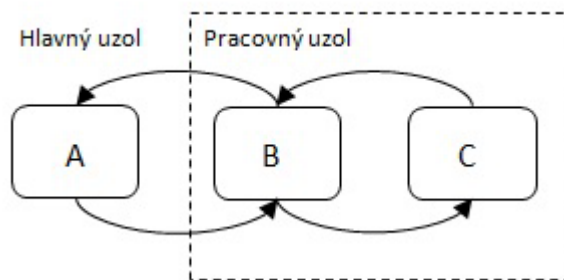
V kontexte dolovania z textov existuje viacero možností ako využiť paralelné resp. distribuované počítanie na zníženie časovej náročnosti celého procesu. V procese dolovania z textov sa stretávame čoraz viac so sekvenčnými úlohami, ktoré vzhľadom na objem niektorých dátových kolekcii vyžadujú dlhý čas na spracovanie. Využitím služieb, ktoré grid poskytuje je možné určité časti týchto úloh paralelizovať a výrazne tak skrátiť priebeh samotného procesu

dolovania. V tejto časti uvedieme niekoľko implementovaných vybraných algoritmov, ktoré boli realizované gridovými službami a ktorých realizácia priniesla výrazné urýchlenie procesu ich behu. Na dôkaz toho, že je to skutočne tak, je možné nájsť detailné analýzy jednotlivých experimentov vo viacerých publikovaných prácach, napríklad [5].

Hlavnou motiváciou bolo v tomto prípade analyzovať existujúce podobné riešenia v rovnakej aplikačnej doméne a na takomto základe navrhnúť služby pre určité čiastkové procesy dolovania v dátach, vhodné pre distribúciu použitím gridových služieb. Cieľom takto modifikovaných algoritmov je na prvom mieste potvrdenie konceptu, že zvyšujúci sa počet zúčastnených pracovných uzlov dokáže výrazne ovplyvniť dĺžku trvania určitých časovo a výpočtovo vysoko náročných úloh. Z predchádzajúcich skúseností v oblasti dolovania v textoch je jednoznačne známe, že najmä na obsiahlych dátových kolekciami pri určitých algoritmoch pre tvorbu klasifikačného modelu dochádza k neúmernému zvyšovaniu časovej náročnosti týchto algoritmov. Problematickou oblasťou v tomto prípade nie je ani samotné dotazovanie modelu a jeho používanie, ale proces jeho tvorby. Ten, v experimentálnych prípadoch, sa samozrejme vytvorí iba raz, na jeho evaluáciu, no v praktickom použití (digitálne knižnice, lekárske databázy), kde sa dátové množiny často menia (doplňujú o nové dáta, atď.) je potrebné tento krok vykonávať opakovane, čo pri spomínaných skutočnostiach nemusí byť postačujúce. Na druhej strane, časová náročnosť nie je jedinou motiváciou pre všeobecnú implementáciu takéhoto typu úloh do distribuovaného, službovo orientovaného prostredia. V posledných rokoch sa webové a gridové služby stali de-facto štandardnými technológiami pre sprístupnenie rôznych typov služieb. Taktiež sa ukazuje, že s podporou technológií sémantického webu je takto možné spomínané nástroje obohatiť o podporu sémantických popisov, čo vytvára podmienky pre implementáciu ďalších podporných metód najmä v oblasti vyhľadávania služieb, automatickej konštrukcie pracovných tokov služieb a iných.

Na vybraných niekoľkých druhoch algoritmov bola otestovaná ich možnosť paralelizácie na klastroch a distribúcia na uzloch gridu a experimentálne overená funkcionálna a výkon. Myšlienka paralelizácie a distribúcie je pri všetkých prístupoch obdobná. Zvolili sme prístup dekompozície na podúlohy so zámerom experimentálne odskúšať implementované služby ako na menších, tak aj rozsiahlejších dátach a overiť tak užitočnosť takejto architektúry aj pre metódy dolovania v textoch. Všetky vybrané algoritmy sú prirodzene vhodné na ich dekompozíciu už len z toho hľadiska, že v každom z nich dochádza k sekvenčnému opakovaniu určitých podprocesov, ktoré ak sú vhodne oddelené, nie sú navzájom závislé. Ako platformu pre dolovanie v textoch sme použili knižnicu JBOWL, keďže podporuje všetky fázy procesu dolovania v textoch a jej použitím sme pripravení analyzovať kolekcie dokumentov a implementovať nové prístupy. Jednotlivé služby dolovania v textoch sú implementované v rámci gridovej služby. Využívajú sa metódy poskytované knižnicou JBOWL. Implementované sú služby realizujúce všetky kroky vrátane fázy predspracovania, filtrov a nástrojov na váhovanie termov, algoritmy klasifikácie (kNN, Support Vector Machine, rozhodovacie stromy, pravidlá), algoritmy zhľukovania (K-Means, GHSOM). V prípade distribuovaných algoritmov, boli text miningové služby implementované na toolkíte Globus. Implementované boli grafické užívateľské rozhrania umožňujúce užívateľom prostredníctvom klientov komunikovať so službami a vytvárať pracovný tok.

Navrhnuté boli distribuované verzie stromového klasifikátora, distribuovanú verziu zhľukovacieho algoritmu GHSOM (Growing Hierarchical Self-Organizing Maps) a aj jeho aplikáciu pri distribuovanom budovaní konceptuálnych zväzov. V prípade rozhodovacích stromov sme použili algoritmus založený na Quinlanovom algoritme C4.5. Rozhodovací strom je štruktúra s vnútornými uzlami, ktoré reprezentujú atribúty (v tomto prípade termy), vetvy z uzlov reprezentujú testy váh atribútov a listy sú kategórie. Takýto prístup je schopný multi-label klasifikácie, iba za použitia konštrukcie množiny binárnych klasifikačných stromov pre každú kategóriu zvlášť. Spojením všetkých čiastkových binárnych klasifikátorov dostaneme výsledný klasifikátor. V prípade obsiahlych textových databáz, sa môže stať proces budovania množstva binárnych klasifikátorov extrémne náročným. Na druhej strane, jednotlivé binárne klasifikátory sú na seba navzájom nezávislé. Fakt, že môžu byť vytvárané nezávisle od seba, znamená, že takýto typ úlohy by bol prirodzený a ideálny pre distribúciu. Metóda buduje multi-label rozhodovací strom distribuovane, rozdelením jednotlivých podúloh na pracovné uzly [2]. Sú to pracovné stanice prepojené sieťou a ich úlohou je iba spustiť konkrétnu inštanciu služby na daných dátach. Druhým typom uzlu použitom v tomto prístupe je hlavný uzol, na ktorom beží samotná služba a ktorý jednotlivé čiastkové podúlohy priradzuje na pracovné uzly. Zjednodušená schéma dátových tokov je zobrazená na obr. 1. Pracovné uzly môžu byť virtualizované a pozostávať z viacerých uzlov (napr. dátový uzol (C), výpočtový uzol (B)).



Obr. 1 Zjednodušená schéma dátových tokov.

Na hlavnom uzle prebehne proces predspracovania dát, spolu s vytvorením modelu dátovej množiny. Táto je vstupom do služby realizujúcej konštrukciu samotného klasifikačného modelu. Zo štatistik dátovej množiny (z počtu kategórií) vypočíta, koľko binárnych klasifikátorov je potrebné skonštruovať na vybudovanie výsledného klasifikačného modelu. Zároveň zistí počet dostupných pracovných uzlov v sieti. Tento počet si volí sám používateľ ako parameter v konfiguračnom súbore, kde zadá počet dostupných uzlov spolu s ich URL adresami. Na základe týchto informácií potom začne pridelovať jednotlivé podúlohy (úlohy vybudovania parciálnych binárnych klasifikátorov) zaradom na jednotlivé pracovné uzly. Proces budovania modelu teda iteruje zoznamom kategórií a pre každú z nich vytvorí zodpovedajúci binárny klasifikátor. Pracovné zaťaženie distribuuje na ďalšie inštancie služby, bežiacie na pracovných uzloch, kde sa vytvárajú čiastkové modely. Tieto sú potom poslané naspäť na hlavný uzol (z C-čka do B-čka, a následne z B-čka do A-čka), kde bola služba spustená, a spojené do výsledného klasifikačného modelu. Tento prístup znižuje zaťaženie siete tým, že nedochádza k stálemu kontaktu uzlov.

Základnou myšlienkou návrhu distribuovanej verzie GHSOM algoritmu bolo paralelné vykonávanie týchto zhukovacích procesov na uzloch gridu [3]. Vstupom prvej fázy je kolekcia dokumentov, na ktorých najskôr prebehne lexikálna analýza. Výstupom prvej fázy sú štatistiky, frekvencie výskytu a slovník termov, ktoré sa uložia do samostatných súborov. Vstupom druhej fázy sú potrebné štatistiky a dokumentové frekvencie indexovaných termov. Po orezaní termov sa vytvorí vektorová reprezentácia a výsledné vektory sa normalizujú. Normalizované vektory sa uložia špeciálnym spôsobom (ukladajú sa len nenulové prvky tak, aby bolo jasné, ktorému termu odpovedajú) do výsledného súboru, ktorý tak obsahuje *tfidf* profily dokumentov a je základným vstupom pre fázu distribúcie. Zo vstupných vektorov dokumentov sa vytvorí mapa GSOM prvej úrovne. V ďalšom kroku sa vytvorí zoznam neurónov, ktoré spĺňajú podmienku expanzie do mapy nižšej úrovne. Následne sa zistí počet momentálne dostupných uzlov a vytvorí sa zoznam ich ukazovateľov. Počet pracovných uzlov je kľúčovým pri vytváraní fronty úloh uzlov.

Každá zhukovacia úloha obsahuje identifikátor neurónu v rámci 1. mapy, zoznam (identifikátorov) vektorov namapovaných na daný neurón a parametre algoritmu GHSOM: minimálny počet vektorov potrebný pre vytvorenie ďalšej podmapy, maximálna hĺbka hierarchie. Priradzovanie zhukovacích úloh uzlom prebieha nasledovne:

Nech n je počet expandovateľných neurónov a u nech je počet dostupných pracovných uzlov, potom :

1. Ak $n \leq u$, do fronty úloh prvých n uzlov sa priradí práve jedna zhukovacia úloha.
2. Ak $n > u$, v prvej iterácii sa prvých u zhukovacích úloh priradí prvým u uzlom, v nasledovných iteráciách sa priradzujú zvyšné zhukovacie úlohy zaradom na jednotlivé uzly dovtedy, kým už nie je čo priradzovať.

Po naplnení fronty úloh, sa postupne z týchto front rozposielajú zhukovacie úlohy na jednotlivé uzly. Na uzloch sa zo vstupných vektorov vytvoria čiastkové hierarchické modely GHSOM použitím definovaných parametrov. Po vytvorení modelu sa tento model odošle späť na hlavný uzol, kde sa uloží a následne sa na ten istý uzol odošle nasledujúca zhukovacia úloha z fronty úloh (ak je fronta prázdna, výpočet na danom uzle končí). Keď sa vyprázdnia všetky fronty úloh a sú prijaté všetky čiastkové hierarchické modely GHSOM, realizuje sa spájanie týchto modelov do výsledného hierarchického modelu. Spájanie modelov prebieha nasledovne: referencie na rodiča mapy uzla 1. úrovne čiastkového modelu sa nastavujú tak, aby odkazovali na mapu 1. úrovne vytvorenej na začiatku procesu. Následne sa zmenia referencie na „potomkov“ mapy uzla 1. úrovne a to tak, aby ukazovali na mapy uzlov 1. úrovne čiastkových modelov.

V prípade distribuovanej konceptuálnej analýzy (Formal Concept Analysis, FCA) je rovnako využívaný princíp rozdelenia problému na menšie časti [4]. Gro metódy spočíva vo formálnom rozbere tabuľky dát a umožňuje identifikáciu zhlukov podobných objektov (reprezentujúcich koncepty) z formálneho konceptuálneho pohľadu. Výstupom tejto metódy sú netriviálne informácie o nájdených konceptoch v zásade dvoch typov – konceptový zväz a atribútové implikácie. Koncept je vlastne zhluk „podobných“ objektov (táto relácia je založená na prítomnosti rovnakých resp. podobných hodnôt atribútov objektov). Koncepty je možné následne hierarchicky zoradiť do podoby zväzu. Samotný moment distribúcie pri tejto úlohe sme zvolili ako východiskový bod možnosti pridelenia jednotlivých konceptových zväzov. Z tohto hľadiska môžeme algoritmus rozdeliť na dve základné vetvy - pridelovanie zhlukov samostatne alebo pridelovanie zhlukov v rámci GSOM mapy. Faktor ktorý do značnej miery ovplyvňuje rýchlosť výpočtu je vstupná GHSOM mapa. V závislosti od jej štruktúry sa mení zložitosť výpočtu konceptových zväzov. Ďalším dôležitým faktorom sú dolované dokumenty. V prípade rozsiahlych dokumentov, ktoré obsahujú značný počet termov, časová náročnosť narastá. Tento algoritmus bol navrhnutý aj pre viacprocesorové systémy.

III. SÚČASNÝ STAV VÝSKUMU A MOŽNOSTI APLIKÁCIE V OBLASTI CLOUD COMPUTING-U

Na základe predchádzajúcich skúseností s využitím gridu ako infraštruktúry pre úlohy dolovania v textoch predstavuje cloud prirodzenú platformu pre implementáciu podobných prístupov využitím cloudových frameworkov. V našom prípade sa ani nejedná o využitie niektorého z verejných cloudov (Amazon EC2, Azure), ale o vytvorenie komunitného cloudu využitím open-sourcových cloud platforiem. Jedným z riešení je využiť platformu GridGain.

GridGain cloudová aplikačná platforma, ktorá umožňuje rýchly vývoj vysoko škálovateľných SaaS, alebo tradičných podnikových webových aplikácií, ktoré sú škálovateľné na akejkoľvek gridovej, či cloudovej infraštruktúre. Od jej prvého vydania v roku 2007 predstavuje GridGain popredný middleware pre distribuované počítanie, založený na technológii Java. V roku 2008 bol GridGain prvý middleware, ktorý bol nezávisle testovaný na lineárne zvyšujúcich sa, až 2048 procesorových jadrách, na Amazon EC2 cloudovej infraštruktúre [6].

Gridgain kombinuje tradičné objektovo-orientované programovanie s komplexnou podporou funkcionálneho programovania. Okrem doménového špecifického jazyka Scala podporuje natívne aj Javu. To výrazne napomáha možnosti adaptovať implementáciu JBOWL algoritmov do distribuovaného cloudového prostredia pomocou tohto frameworku.

Okrem implementácie algoritmov dolovania v textoch je našou úlohou aj vytvoriť systém pre správu vytvorenej infraštruktúry a monitorovanie záťaže jednotlivých uzlov. Navrhovaný systém umožní získavať aktuálne údaje pre pripojeného používateľa a poskytovať tak prostredníctvom webovej aplikácie dôležité informácie o stave pracovných uzlov zapojených v cloude. Rozhranie je vyvíjané v prostredí Java, pričom sú využívané API frameworku GridGain. Každý pripojený výpočtový uzol ponúka svoj diskový priestor a výpočtový výkon. Získané informácie je možné využiť pre monitoring prostredia ale aj pre optimalizáciu rozdelenia jednotlivých úloh. Framework Gridgain umožňuje z jednotlivých pracovných uzlov získať viaceré údaje o ich hardvérovej konfigurácii a ich aktuálnom zaťažení:

- Názov pripojeného počítača
- Meno používateľa
- IP adresa uzla
- MAC adresu uzla
- Operačný systém
- Počet procesorov, ktoré sú k dispozícii
- Vyťaženie procesorov, ktoré sú k dispozícii
- Informácie o bežiacich úlohách
- Stav operačnej pamäte
- Množstvo aktuálne dostupnej pamäte
- Súborový systém
- Voľné miesto na diskovom úložisku

Výsledná webová aplikácia bude samozrejme okrem monitorovacieho rozhrania vyšperkovaná aj ďalšími funkcionalitami, ako sú nástroje pre správu konfigurácie infraštruktúry, vrátane dynamického pridávania nových zdrojov (dátových, alebo výpočtových uzlov), schopnosti nahradiť výpadky v infraštruktúre (failover) a iné. Cieľom je, aby vytvorený

system umožňoval aj jednoduché vytváranie výpočtových úloh, ktoré sa pomocou vytvorenej webovej aplikácie budú môcť spustiť na cloudovej infraštruktúre priamo z rozhrania, pričom používateľ sa nebude musieť zaoberať detailmi infraštruktúry. Vďaka navrhovanému riešeniu budeme pripravení implementovať, monitorovať a analyzovať rôzne úlohy realizované na takejto infraštruktúre.

POĎAKOVANIE

Táto práca bola podporená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (ITMS kód: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] Mell, P. Grance, T.: The NIST Definition of Cloud Computing, ver. 15, NIST, Information Technology Laboratory, 2009.
- [2] Jančiak, I, Sarnovsky, M., Tjoa, A Min, Brezany, P.: Distributed classification of textual documents on the Grid. In: High Performance Computing and Communications : Second international conference, HPCC 2006, Munich, Germany, September 13-15, 2006 : Proceedings. Berlin : Springer, 2006. p. 710-718. ISBN 3-540-39368-4.
- [3] Sarnovský, M., Butka, P., Safko, V.: Distribuované zhľukovanie textových dokumentov v prostredí Gridu. In: Znalosti 2008, Bratislava, Slovakia, pp.192-203, ISBN 978-80-227-2827-0.
- [4] Sarnovský, M., Butka, P., Sakáč, M.: Distribuovaná tvorba hierarchie konceptov z textových dokumentov v gridovom prostredí. In: ITAT 2007 : Information Technologies - Applications and Theory : Conference on Theory and Practice of Information Technologies : Proceedings : 21. 9. - 27. 9. 2007, Poľana, Slovakia. Seňa : PONT, 2007. s. 97-102. ISBN 978-80-969184-7-8.
- [5] Sarnovský, M., Butka, P., Paralič, J. (2009) Grid-based Support for Different Text Mining Tasks. In: Acta Polytechnica Hungarica, vol. 6, no. 4 (2009), pp. 5-27, ISSN 1785-8860.
- [6] GridGain 3.0, White Paper, 2010.

Niektoré vlastnosti sociálnej siete európskych výskumných organizácií

Gabriel LUKÁČ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

gabriel.lukac@tuke.sk

Abstrakt – Predkladaný príspevok ponúka základný prehľad teórie o statických topologických vlastnostiach skúmaných v rámci analýzy sociálnych sietí. Príspevok v rámci teoretického úvodu začína vysvetlením pojmov klasterizácie a klasterizačného koeficientu. Ďalej vysvetľuje vlastnosť Malého sveta a popisuje niektoré distribúcie stupňov vrcholov typické pre sociálne siete. V praktickej časti sa príspevok venuje konkrétnej sociálnej sieti európskych organizácií participujúcich na výskumných projektoch v FP5 - FP7 rámcových programoch. Uvedené sú predbežné výsledky získané analýzou základných topologických vlastností siete. V závere je načrtnutý plán pre ďalší výskum.

Kľúčové slová – sociálne siete, topologické vlastnosti sociálnych sietí, európske projekty

I. ÚVOD

Teória komplexných sietí poskytuje rámec pre výskum prepojeného sveta. Je základom pre popis ich topológie, dostatočne všeobecným na to, aby pokryl potreby charakterizácie sietí z diametrálne odlišných odvetví. Jednu teóriu je takýmto spôsobom možné aplikovať pre analýzu biologických sietí (vzťahy sekvencií génov DNA a metabolizovaných bielkovín), informačných sietí (citačné siete alebo sieť WWW), technologických sietí (distribúcie rôznych komodít, elektrické siete, ...), atď. Posledných desať rokov bolo pre analýzu komplexných sietí kľúčovým obdobím. Objav, že distribúcia stupňov vrcholov u reálnych sietí sleduje mocninný zákon viedla k zmene doterajšieho myslenia o štruktúre a zapôsobila ako katalyzátor pre tvorbu nových topologických modelov. Ich cieľom bolo nájsť čo najväčšiu koreláciu medzi tým, čo poskytovali reálne dáta. Čoskoro sa však prišlo na to, že globálna teória sietí sa nemôže zaoberať bez zapracovania dynamiky do jej základov, pretože ako sa ukazuje, drvivá väčšina príkladov prepojených štruktúr vykazuje určitý druh dynamického správania. Skúmanie dynamiky má kritickú úlohu pri zisťovaní šírenia ideí, inovácií, počítačových vírusov alebo kritických článkov systémov. Na druhej strane dynamické procesy ovplyvňujú zmeny v topológii a teda je vidno, že výskum štruktúry bez dynamiky bude z komplexného hľadiska vždy len čiastočný a naopak.

Skúmanie sociálnych sietí a dynamických procesov, ktoré v nich prebiehajú môže priniesť nové poznatky o formovaní a fungovaní rôznych spoločenských entít. Ak hovoríme o sociálnych sieťach, nejde tu len o známe internetové služby, kde je možné jednoduchým spôsobom komunikovať a virtuálne sa stretávať so svojimi priateľmi. Ich aplikácia má ďaleko širší dopad. Pomáha odborníkom sledovať šírenie inovácií medzi firmami, virológom a epidemiológom ponúka možnosti na vytváranie modelov pre šírenie vírusových ochorení. Psychológovia ju vedia využiť na odhaľovanie šikanovania detí na školách.

Ako ukazujú štúdie realizované na poli rôznych vedeckých disciplín (viď. napr. [1]), grafy reprezentujúce reálne siete majú výrazne inú topológiu ako niektoré historicky populárne teoretické modely. Vlastnosti reálnych sietí sú podstatne iné ako vlastnosti náhodné grafy s uniformnou distribúciou [2], alebo pravidelné mriežkové grafy.

Táto práca ponúka vo svojej praktickej časti náčrt analýzy kolaboratívnej siete medzi európskymi výskumnými organizáciami. Vo svete sa výskumom kolaboratívnych sociálnych sietí vznikajúcimi medzi aktérmi (nech si už pod aktérom predstavíme výskumníka alebo výskumnú organizáciu) zaoberalo viacero autorov. Newman sa v [3] venuje štruktúre kolaboratívnej siete autorov navzájom spolupracujúcich na príprave výskumných príspevkov. Podobnou témou sa zaoberajú autori [4] pri podrobnej analýze štruktúry siete európskych organizácií participujúcich v rámcových programoch FP1 - FP4. Títo priniesli fakticky prvú štúdiu kolaboratívnej európskej výskumnickej siete.

II. MATEMATICKÝ POPIS VYBRANÝCH VLASTNOSTÍ SOCIÁLNYCH SIETÍ

Pre reálne siete z rôznych oblastí prírody alebo spoločenského života sú typické veľmi podobné topologické vlastnosti.

A. Klasterizácia

Klasterizácia¹ (definované napr. v [5]) (alebo tiež známe pod pojmom *tranzitívnosť*) je typickou vlastnosťou, najlepšie vysvetliteľnou na príklade siete vytvorenej medzi aktérmi patriacimi do širšieho okruhu známych. V takýchto sieťach je často pozorovaným javom, že dvaja jednotlivci, ktorí majú spoločného známeho sa taktiež poznávajú medzi sebou [6]. Klasterizácia môže byť kvantifikovaná ako pomer počtu trojíc vytvárajúcich trojuholníky k počtu všetkých trojíc nachádzajúcich sa v grafe. Niekedy sa trojice nazývajú aj triádami. V literatúre sa pojem klasterizácia zvykne nazývať aj globálnym klasterizačným koeficientom.

$$T = \frac{3 \times \text{pocet trojuholnikov v grafe}}{\text{pocet spojenych trojic vrcholov v grafe}} \quad (1)$$

Číslo 3 v čitateli zaručuje, že každý trojuholník vytvorený spojením troch uzlov prispieva práve tromi spojenými trojicami (triádami), každou centrovanou v jednom z troch uzlov a zaručuje, že platí $0 \leq T \leq 1$. $T = 1$ pre K_N - kompletný graf.

B. Klasterizačný koeficient

Popri klasterizácii je možné pri popisovaní tranzitívnych vlastností použiť alternatívnu mieru tzv. **klasterizačný koeficient**² zavedený Watts-om a Strogatz-om v [7]. Povedané neformálne, klasterizačný koeficient hovorí o tom, ako dobre je prepojené okolie nejakého uzla [8]. Vysvetlené podrobnejšie, klasterizačný koeficient c_i uzla i vyjadruje pomer medzi aktuálnym počtom hrán medzi uzlami susednými k uzlu i a počtom hrán, ktoré by medzi nimi existovali, ak by tvorili kompletný graf. Klasterizačný koeficient sa potom definuje ako pomer medzi E_i (aktuálnym počtom hrán v grafe G_i , pričom G_i predstavuje podgraf susedných uzlov k uzlu i) a $\delta_i(\delta_i - 1)/2$, maximálnym počtom hrán v G_i , kde δ_i je stupeň uzla i :

$$c_i = \frac{2E_i}{\delta_i(\delta_i - 1)}. \quad (2)$$

Klasterizačný koeficient pre celý graf je potom rovný priemeru hodnôt lokálnych klasterizačných koeficientov c_i pre všetky uzly grafu G :

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in N} c_i. \quad (3)$$

C. Small-World vlastnosť

Koncom šesťdesiatych rokov uskutočnil Stanley Milgram experiment [9], ktorý viedol k objavu tzv. *problému malého sveta*³ a následne na to vyústil k dôkladnému skúmaniu štruktúry sociálnych sietí. Bol založený na jednoduchej túžbe autora zistiť aká je pravdepodobnosť toho, že ľubovoľne vybraní dvaja ľudia sa navzájom poznajú. Experiment pozostával z jednoduchej procedúry skladajúcej sa z nasledovných krokov:

- 1) Celý experiment bol založený na sledovaní reťaze odosielateľov korešpondencie z nejakého miesta v USA (išlo o mesto Omaha v Nebaske a mesto Wichita v štáte Kansas) do cieľa, ktorý sa nachádzal v Bostone. Milgram vybral niekoľko miest, ktoré boli od seba dostatočne vzdialené nielen geograficky ale aj sociálne.
- 2) Prvé, čo urobil bolo to, že poslal informačné balíčky vybraným jednotlivcom v Omaha a Wichte. Tie zahŕňali informácie o celi experimentu a základné informácie o adresátovi v nachádzajúcom sa v Bostone.
- 3) Prvá otázka, ktorá bola položená každému z participantov experimentu sa týkala priameho „poznania“ adresáta. Ak daný participant poznal adresáta priamo, bol vyzvaný poslať list adresátovi priamo.
- 4) Ak daný participant adresáta nepoznal priamo, bol inštruovaný tak, aby porozmýšľal nad priateľom alebo príbuzným, o ktorom si myslel, že môže takéhoto človeka poznať skôr ako on. Mal pripísať na list svoje meno do zoznamu a taksito poslať na Harvard korešpondenčný lístok, ktorý umožnil podrobné sledovanie aktuálneho stavu balíčka.

¹angl. clustering

²angl. clustering coefficient

³angl. small world problem

- 5) Keď balíček dorazil k adresátovi v Bostone, mohol Milgram odsledovať cestu a počet odosielateľov v reťazi. Takisto korešpondenčné lístky umožnili nájsť prípadné osoby, na ktorých sa ďalší postup zasekol.

Aj keď bol tento experiment kritizovaný za malú návratnosť balíčkov (mnohí participant experimentu odmietli preposlať balíček ďalej), priemerná vzdialenosť medzi vybraným participantom a adresátom bola 5,5 až 6 ľudí. Z toho bol vyvodený zovšeobecnený záver, že vzdialenosť medzi ľubovoľnými dvoma ľuďmi v USA je približne 6.

Takýto objav mal za dôsledok to, že odrazu malo zmysel študovať štruktúru rôznych komplexných sietí (nielen sociálnych), pretože bolo jasné, že správy, fámy, móda alebo infekčná choroba sa budú v sieti s vzdialenosťou medzi aktérmi rovnvej 6 šíriť oveľa rýchlejšie, ako keby táto vzdialenosť bola rovná desiatim alebo tisícim.

Naskytuje sa otázka, čo všetko by bolo potrebné a či by vôbec bolo možné namodelovať sieť s požadovanými vlastnosťami, aby na nej bolo možné spustiť experimenty, predikujúce trebárz šírenie hemoragickej horúčky s ohniskom niekde v Sudáne do celého sveta.

V práci [7] Watts a Strogatz podotkli, že siete je možné klasifikovať na základe dvoch nezávislých štruktúrnych vlastností. Prvou z nich je *klasterizačný koeficient* popísaný v kapitole (II-B) a druhou *stredná vzdialenosť*⁴. Ukázali, že stredná vzdialenosť uzlov v reálnych sieťach nadobúda (neočakávane) nízke hodnoty, ale ich klasterizačný koeficient je výrazne vyšší ako ten vyplývajúci z predtým získaných poznatkov o náhodných grafoch. Na základe týchto rozdielov navrhli model lepšie odrážajúci topológiu reálnych sietí.

Stredná vzdialenosť⁵ l medzi dvoma vrcholmi v neorientovanej sieti je definovaná nasledovne:

$$l = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i \geq j} d_{ij}, \quad (4)$$

kde d_{ij} je **vzdialenosť**⁶ z uzla i do uzla j [11]. Hodnota strednej vzdialenosti l môže byť pre sieť s N uzlami a E hranami odmeraná napr. algoritmom *prehľadávania do šírky*⁷ podľa [12] pracujúcom v čase $O(NE)$.

Ak sa sieť, ktorú skúmame skladá z viacerých komponentov (to znamená, že v nej existuje dvojica vrcholov, medzi ktorými neexistuje cesta), potom sa definícia (4) strednej vzdialenosti tak ako je uvedená v texte hore stáva problematickou. Na vyriešenie tohto problému sa v literatúre objavujú dva prístupy. Prvý je ten, že sa hodnota strednej vzdialenosti počíta len zo vzdialeností uzlov, medzi ktorými existuje cesta [11]. Dvojice uzlov, ktorých pôvod je z rozdielnych komponent nie sú do prímeru zahrnuté. Alternatívnou metódou výpočtu uvedenou v [13] je tzv. **stredná harmonická vzdialenosť**⁸ definovaná takto:

$$l^{-1} = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i \geq j} d_{ij}^{-1}. \quad (5)$$

Pri takomto zadefinovaní, nekonečné vzdialenosti d_{ij} uzlov, medzi ktorými neexistuje cesta prispievajú k celkovej sume nulovým prírastkom.

Nech priemerný stupeň ľubovoľného uzla je z_1 . Potom môžeme povedať, že približne z_1^l uzlov je k tomuto uzlu vo vzdialenosti l alebo bližšie. Ak N je počet všetkých vrcholov v sieti tak platí, že $N \sim z_1^l$. Z toho vyplýva, že $l \sim \ln N / \ln z_1$. Tento efekt veľmi malého nárastu strednej vzdialenosti s pridaním aj vysokého počtu vrcholov sa nazýva **Small-World efekt** [7][14].

D. Distribúcie stupňov vrcholov

Ak uvažujeme siete, modelujúce reálne vzťahy, rôzne uzly, ktoré predstavujú konkrétnych aktérov sú rôzneho stupňa. Situáciu si je jednoduché predstaviť, rôzni ľudia, organizácie, atď. majú napr. rôzny okruh známych, rôzny počet transakcií, ktoré realizujú s rôznym počtom parterov. Rozloženie stupňov vrcholov je v takejto sieti dané distribučnou funkciou pravdepodobnosti $P(k)$, ktorá vracia pravdepodobnosť p_k , že vybraný uzol bude incidovať s práve k hranami [1]. Graf funkcie $P(k)$ pre ľubovoľnú sieť môžeme zostrojiť ako histogram stupňov uzlov. Takto vytvorený histogram potom predstavuje distribúciu stupňov uzlov danej siete [11].

⁴Niekedy je tiež použitý pojem priemerná vzdialenosť. Iní autori [10] nazývajú túto veličinu „priemer siete (grafu)“. Rozlišujú pri tom dva rôzne prípady. Buď priemer definujú ako priemernú hodnotu vzdialeností medzi dvojicami uzlov, alebo najväčšiu z vyskytujúcich sa vzdialeností zadefinovanú v kapitole nižšie.

⁵angl. mean geodesic

⁶angl. geodesic

⁷tiež známy pod názvom *Burning algorithm*

⁸angl. mean harmonic geodesic

Historicky sa na objavovanie štatistických vlastností komplexných sietí používal model náhodného grafu. V ňom sú hrany medzi jednotlivými uzlami umiestňované náhodne s konštantnou pravdepodobnosťou a teda z toho vyplýva dôsledok, že väčšina uzlov bude mať približne rovnaký stupeň, veľmi blízky priemernej hodnote $\langle k \rangle$. Distribúcia stupňov vrcholov náhodného grafu zodpovedá Poissonovej distribúcii s maximom v bode $P(\langle k \rangle)$. To, čo sa ale ukázalo ako jeden z najzaujímavejších poznatkov o topológii reálnych sietí bol objav, že distribúcia stupňov vrcholov sietí s veľkým počtom vrcholov je úplne iná, ako tá pozorovaná pri náhodných grafoch. Konkrétne výskumy boli realizované na rôznych dátach z rôznych oblastí ľudského záujmu vrátane WWW [15], Internetu [16] alebo metabolických sietí [17]. Tieto skúmania prišli s poznatkom, že mnohé reálne sieťové štruktúry vykazujú spoločnú štatistickú vlastnosť - ich pravdepodobnostná distribúcia stupňov vrcholov sleduje tzv. *mocninový zákon*⁹ [18]. Inými slovami môžeme napísať:

$$P(k) \propto k^{-\gamma}, \quad (6)$$

kde γ predstavuje **škálovací exponent** distribúcie stupňov vrcholov.

Táto univerzálnosť (aj keď je iba kvalitatívna) znamená, že tvar funkcie distribúcie stupňov vrcholov je podobný pre všetky skúmané prípady sietí spomenutých v predchádzajúcom odstavci. Z toho vyplýva poznatok, že na systémy s podobnými vlastnosťami sa na určitej úrovni abstrakcie dá pozeráť ako keby boli rovnaké. Siete s takýmito vlastnosťami sa nazývajú tzv. **bezškálové (angl. Scale-Free) siete** [19]. Bezškálovosť je v tomto prípade vlastnosť, hovoriaca o absencii škály – t.j. typického stupňa uzla.

III. SOCIÁLNA SIEŤ EURÓPSKÝCH VÝSKUMNÝCH ORGANIZÁCIÍ

V nasledujúcom texte sú uvedené priebežné výsledky analýzy, ktorá je súčasťou výskumu [20] sociálnej siete európskych organizácií. Analýza ďalších vlastností, podľa toho ako sú uvedené v texte, bude predmetom ďalšej práce. Bude však nevyhnutné vysporiadať sa s problémami vyplývajúcimi z nekonzistentnosti dát.

A. Popis dát

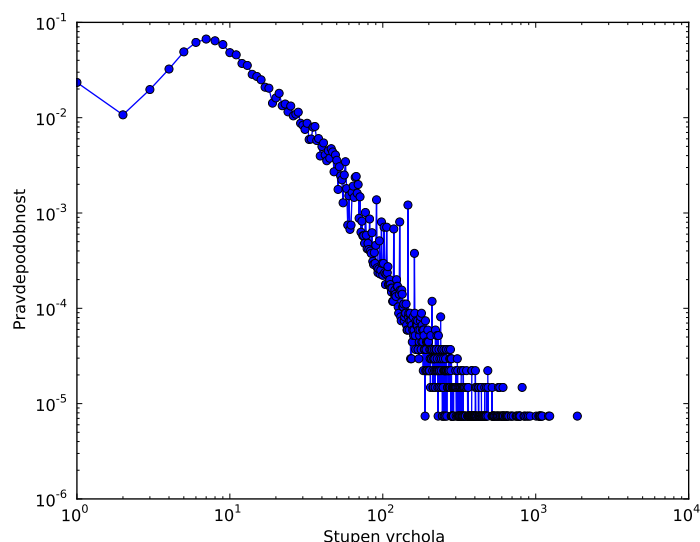
Dáta, z ktorých bola vyextrahovaná sociálna sieť boli poskytnuté organizáciou CORDIS [21] spravujúcou všetky údaje o projektoch a organizáciách participujúcich na projektoch spolufinancovaných Európskou komisiou. Poskytnuté dáta obsahujú údaje o projektoch realizovaných v rámci programov FP5 - FP7, pričom export údajov bol realizovaný k 30. máju 2010. Keďže koniec siedmeho rámcového programu je plánovaný na rok 2013, kvôli aktuálnosti bolo vhodné doplniť údaje tak, ako postupne neustále pribúdajú. Pre tento účel bol zrealizovaný automatický *crawler*, ktorý automaticky sleduje lokalitu CORDIS-u a aktualizuje údaje.

Databáza obsahuje celkovo 35263 projektov. Z piateho rámcového programu sa v nej nachádza 17628 projektov, zo šiesteho 10660 a zo siedmeho 6975. Z údajov o projektoch boli vyextrahované identifikátory participujúcich organizácií, ktoré boli neskôr použité pri vytváraní sociálnej siete. Spôsob, akým sú jednotlivé organizácie v databáze identifikované je nejednoznačný, z čoho vyplýva, že aj výsledky prezentované v tejto štúdii sú skreslené. Znamená to možnosť, že tá istá organizácia môže byť v databáze uvedená pod rôznymi identifikátormi, a teda vo výslednej sieti sa objaví reprezentovaná dvoma alebo viacerými uzlami. Automatickému zjednoteniu rôznych identifikátorov jednej organizácie bránia dve veci. Tou prvou je fakt, že nedokážeme odhadnúť, koľko rovnakých organizácií s rôznymi identifikátormi sa v skutočnosti v databáze nachádza. A tá druhá je to, že ani na základe názvu organizácie a jej adresy nie je možné organizáciu jednoznačne identifikovať. Ide o situácie kedy je napr. v prípade univerzít v jednom projekte uvedený názov fakulty spolu s menom univerzity a s príslušnou adresou a v inom projekte zase len názov univerzity. Ak by sme chceli byť presní, tak na oboch projektoch participovali riešitelia práve len danej fakulty, ale z údajov o druhom projekte to nie je možné zistiť. Na druhej strane máme možnosť znížiť granularitu a identifikovať aktérov na základe všeobecnejšieho mena. Tým ale stratíme možnosť, že v dátach objavíme zaujímavé vzťahy napr. na úrovni komunít.

B. Sociálna sieť organizácií

Naša sociálna sieť bola vyextrahovaná z databázy projektov a organizácií, ktoré sa podieľali na riešení projektov. Formálne je naša sieť reprezentovaná jednoduchým neorientovaným grafom, ktorého vrcholy predstavujú identifikátory jednotlivých organizácií. Hrana medzi dvoma vrcholmi existuje práve vtedy, ak dané organizácie medzi sebou niekedy v histórii spolupracovali. V jednoduchom grafe medzi dvoma rôznymi vrcholmi neexistujú viacnásobné hrany. Teda, ak dve

⁹angl. power-law



Obr. 1 Distribúcia stupňov vrcholov v sociálnej sieti európskych výskumných organizácií.

organizácie spolupracovali na riešení viacerých projektov, bude tento vzťah obsiahnutý v jedinej hrane.

V ďalšom texte uvádzame prehľad výsledkov základnej analýzy sociálnej siete výskumných organizácií. Zistené charakteristiky poslúžia pri smerovaní ďalšieho výskumu ako aj pre porovnanie našej siete s inými reálnymi komplexnými sieťami.

V nepredspracovanej verzii siete (teda pri základnej analýze sme zatiaľ nepoužili žiadnu z metód na zjednotenie identifikátorov rôznych organizácií) sa nachádza celkovo 135236 vrcholov a 1381977 hrán. Už zo samotného počtu vrcholov a hrán je vidno, že ide o sieť obrovských rozmerov. Pri takejto veľkosti by iste bolo užitočné vidieť vizualizáciu celkovej topológie avšak akékoľvek snahy byť v tomto smere úspešný stroskotali. Aj keď sa na webe dajú nájsť rôzne kvalitné vizualizácie sietí s rádovo stotisícami uzlov, nám sa s dostupnými open-source nástrojmi použitými na bežných súčasných počítačoch nepodarilo uspieť. Každý z pokusov skončil na nedostatku pamäťových zdrojov.

Distribúcia stupňov vrcholov vykreslená na diagrame s osami v logaritmických mierkach jasne sleduje mocninný zákon. Detaily je možné vidieť na obrázku 1.

Sieť obsahuje celkovo 4101 súvislých komponentov. Najväčší súvislý komponent obsahuje 107331 vrcholov a 1238251 hrán. Jeho krivka funkcie distribúcie stupňov vrcholov má takmer totožný tvar ako distribúcia stupňov vrcholov celkovej siete, preto ju už v tomto príspevku neuvádzame. Druhý súvislý komponent v poradí má len 157 vrcholov. Veľkosť ostatných komponentov prudko klesá (v sieti sa nachádzajú komponenty pozostávajúce z dvojíc alebo trojíc rádovo v tisícoch). Predbežne môžeme povedať, že táto situácia nastala kvôli nejednoznačnosti identifikácie organizácií.

IV. PRÁCA V BUDÚCNOSTI A ZÁVER

V príspevku sme predstavili náhľad na štruktúru sociálnej siete európskych výskumných organizácií. Uviedli sme základné štatistické informácie, ktoré bližšie popisujú jej topológiu.

Keďže sa už v minulosti niekoľko autorov venovalo analýze na rovnakej doméne, naša práca ich nemá v úmysle reprodukovat'. Namiesto toho plánujeme doplniť ich výsledky o pozorovania na novších dátach (rámcové programy FP5 - FP7). Svoju pozornosť chceme zamerať na niekoľko oblastí: Prvá sa týka predspracovania, v ktorom bude potrebné zjednotiť identifikáciu rovnakých organizácií jednotnými označeniami. Aj keď takmer určite nebudeme vedieť zaručiť úplné vyriešenie tohto problému, veríme, že určitý posun v presnosti sa nám podarí dosiahnuť. V tomto kroku bude potrebné použiť metódy dolovania z textov. Druhou oblasťou je predikcia rastu siete počas toho ako sú vyhlasované nové výzvy a schvaľované nové projekty. Takéto modely by mohli byť nápomocné grantovým agentúram pri alokácii finančných zdrojov. Tretím rozšírením bude identifikácia vedeckých komunít pôsobiacej v určitých oblastiach výskumu. Zaujímá nás predovšetkým emergencia a prepojenie komunít na základe rôznych sémantických anotácií ich členov.

POĎAKOVANIE

Táto práca bola vytvorená realizáciou projektu Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (kód ITMS projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.

LITERATÚRA

- [1] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 6 2001.
- [2] P. Erdős and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.
- [3] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.
- [4] M. J. Barber, A. Krueger, T. Krueger, and T. Roediger-Schluga, "Network of european union-funded collaborative research and development projects," *Phys. Rev. E*, vol. 73, no. 3, p. 036132, 2006.
- [5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, 2006.
- [6] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, 1st ed. Cambridge University Press, 1994.
- [7] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [8] M. Markošová, *Dynamika komplexných sietí*. Slovenská technická univerzita v Bratislave, 2010.
- [9] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, p. 61, 1967.
- [10] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187, June 2002.
- [11] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [12] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, February 1993.
- [13] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Phys. Rev. Lett.*, vol. 87, no. 19, p. 198701, Oct 2001.
- [14] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness (Princeton Studies in Complexity)*, illustrated edition ed. Princeton University Press, November 2003.
- [15] R. Albert, H. Jeong, and A.-L. Barabasi, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, September 1999.
- [16] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. New York, NY, USA: ACM, 1999, pp. 251–262.
- [17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, October 2000.
- [18] G. Caldarelli and A. Vespignani, *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science (Complex Systems and Interdisciplinary Science)*. World Scientific Publishing Company, June 2007.
- [19] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science (New York, N.Y.)*, vol. 286, no. 5439, pp. 509–512, October 1999.
- [20] G. Lukáč and M. Marián, "Sociálna sieť európskych výskumných organizácií," in *WIKT 2010: Proceedings of 5th Workshop on Intelligent and Knowledge Oriented Technologies*, November 2010, pp. 137–140.
- [21] P. Office. (2010) Eu-funded r&d projects website. [Online]. Available: <http://cordis.europa.eu/>

Matlab-Based Tools for Analysis and Control of Inverted Pendula Systems

Slávka JADLOVSKÁ, Ján SARNOVSKÝ

Dept. of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

sjadlovska@gmail.com, jan.sarnovsky@tuke.sk

Abstract — The aim of this paper is to provide an original approach to the analysis and control of inverted pendula systems which employs custom blocks and demo schemes of a Simulink block library designed by the authors. The capabilities of the library are enhanced by software tools which provide a user-friendly graphical interface to modeling and linearization. The tools are next shown in action as classical double and rotary single pendulum systems are analyzed, modeled and successfully stabilized in the unstable inverted position of the pendula.

Keywords — system of inverted pendula, state-space control, MATLAB/Simulink block library

I. INTRODUCTION

Inverted pendula systems represent a significant group of mechanical systems used in control education with a number of practical applications. It is evident that the stabilization of a walking human or robot, a launching missile or the vertical movement of a shoulder or arm can all be modeled by some kind of an inverted pendulum system. The diversity of modeled systems is reflected in the variety of available inverted pendula models. These may differ by:

- *the type of actuating mechanism* – the system base is moving either in a single axis (*classical pendulum system*) or in a plane (*rotary pendulum system*)
- *the number of pendulum links attached to the mechanism* – single and double pendulum systems are common control plants; triple and quadruple pendulum are rare but controllable
- *the distribution of mass within the pendulum rod* – the pendulum links are either homogenous rods with the mass concentrated in the center of gravity; or the rod is considered massless and the mass is concentrated in the load at its end.

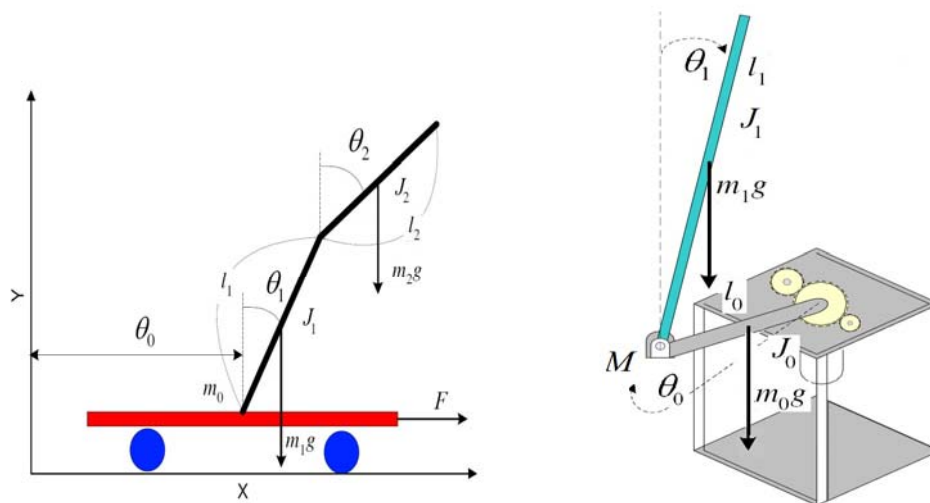


Fig. 1 a) Classical double inverted pendulum system - scheme,
b) Rotary single inverted pendulum system – scheme

The most standard representative of the family of inverted pendula systems, the classical single inverted pendulum system, was thoroughly analyzed in [1][2] together with suitable

control algorithms and therefore will not be included in this paper. This paper will instead focus on the analysis and control of the more challenging systems, i.e.:

- double (two-link) classical inverted pendulum system (Fig. 1a)
- single (one-link) rotary inverted pendulum system (Fig. 1b)

II. A LIBRARY OF MODELS

The nonlinear *mechanic system of n inverted pendula* is basically a set of $n > 1$ pendulum links attached to a stable base which may be moveable in one axis (cart) or in a horizontal plane (rotary arm). It is a typical example of an underactuated system since the number of actuators is lower than the number of system links: the only input (the force acting upon the cart/ the momentum applied on the rotary arm) is used to control the $n+1$ outputs of the system: cart position [m] or arm angle [rad], and pendula angles [rad].

Since 2009, a thematic Simulink block library *Inverted Pendula Modeling and Control (IPMaC)* has been developed. The purpose of the block library is to provide software support for analysis, simulation and control of inverted pendula systems using custom-designed blocks and „demo“ simulation schemes which illustrate the way the blocks may be interconnected to solve various analysis- and control-related problems. The installation process and sublibrary structure of the *IPMaC* was described in detail in [1].

A. Derivation of Motion Equations

An integral part of the *IPMaC*, the *Inverted Pendula Model Equation Derivator*, is a MATLAB GUI tool which generates the motion equations for a user-chosen type of inverted pendula system. Such automatic approach has a number of advantages: it yields a particularly precise approximation of the real system's dynamics and eliminates any factual or numeric errors which could arise during manual mathematical modeling. Fig. 2 shows a preview of the *Derivator* output for the single rotary inverted pendulum system.

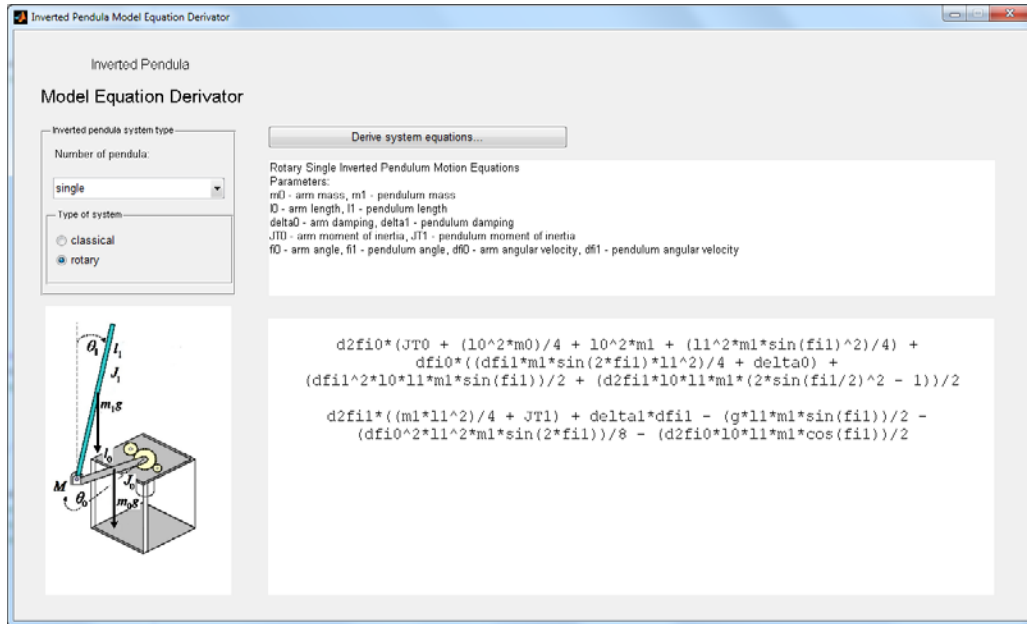


Fig. 2 *Inverted Pendula Model Equation Derivator* GUI tool in use:
 rotary pendulum model derivation

The core of the *Derivator* tool is represented by MATLAB functions that use the *Symbolic Math Toolbox* to implement general procedures that derive the motion equations for a classical or rotary inverted pendula system. If we represent the system's outputs as a vector of generalized coordinates [2]:

$$\theta(t) = (\theta_0(t) \quad \theta_1(t) \quad \dots \quad \theta_n(t))^T \quad (1)$$

then the system can be mathematically described by the Euler-Lagrange equations:

$$\frac{d}{dt} \left(\frac{\partial L(t)}{\partial \dot{\theta}(t)} \right) - \frac{\partial L(t)}{\partial \theta(t)} + \frac{\partial D(t)}{\partial \dot{\theta}(t)} = Q^*(t) \quad (2)$$

where $L(t)$ (*Lagrange function*) is defined as the difference between the system's kinetic and potential energy, $D(t)$ (*Rayleigh, dissipation function*) describes the viscous (friction) forces and $\mathbf{Q}^*(t)$ is the vector of *generalized external forces* acting upon the system. The process of derivation of the motion equations to describe any kind of inverted pendula system is hence transformed into the determination of kinetic, potential and dissipation energies related to the base and all pendula. Using well-known physical formulae, general relations that describe the energetic balances of the base and i -th pendulum in an inverted pendula system were derived. These can be found in [2] for the system of inverted pendula on a cart and derived analogously for the rotary pendulum system. The whole derivation process can be tracked in the command window (see the preview in [1]).

Using the *Derivator* tool, mathematical models of both the classical double and rotary single pendulum system were generated. The motion equations were rewritten into the standard (minimal ODE – ordinary differential equation) form:

$$\mathbf{M}(\boldsymbol{\theta}(t))\ddot{\boldsymbol{\theta}}(t) + \mathbf{N}(\boldsymbol{\theta}(t), \dot{\boldsymbol{\theta}}(t))\dot{\boldsymbol{\theta}}(t) + \mathbf{P}(\boldsymbol{\theta}(t)) = \mathbf{V}(t) \quad (3)$$

which provides the only way to express this kind of system in the nonlinear state-space form of

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), u(t), t) \\ \mathbf{y}(t) &= \mathbf{g}(\mathbf{x}(t), u(t), t) \end{aligned} \quad (4)$$

by defining the state vector as $\mathbf{x}(t) = (\boldsymbol{\theta}(t) \ \dot{\boldsymbol{\theta}}(t))^T$ and isolating the second derivative $\ddot{\boldsymbol{\theta}}(t)$ from (3).

B. Selected inverted pendula models – model analysis

The *classical double inverted pendulum system* (Fig. 1a) is composed of a pair of rigid rods which are interconnected in a joint and one of these is attached to a cart. The mathematical model of the system has the form of three rather complex second-order nonlinear differential equations which describe the dynamic behavior of the cart and both pendulum links:

$$\begin{aligned} & \begin{pmatrix} m_0 + m_1 + m_2 & \left(\frac{1}{2}m_1l_1 + m_2l_1\right)\cos\theta_1(t) & \frac{1}{2}m_2l_2\cos\theta_2(t) \\ \left(\frac{1}{2}m_1l_1 + m_2l_1\right)\cos\theta_1(t) & J_1 + m_2l_1^2 & \frac{1}{2}m_2l_1l_2\cos(\theta_1(t) - \theta_2(t)) \\ \frac{1}{2}m_2l_2\cos\theta_2(t) & \frac{1}{2}m_2l_1l_2\cos(\theta_1(t) - \theta_2(t)) & J_2 \end{pmatrix} \begin{pmatrix} \ddot{\theta}_0(t) \\ \ddot{\theta}_1(t) \\ \ddot{\theta}_2(t) \end{pmatrix} + \\ & + \begin{pmatrix} \delta_0 & -\left(\frac{1}{2}m_1l_1 + m_2l_1\right)\dot{\theta}_1(t)\sin\theta_1(t) & \frac{1}{2}m_2l_2\cos\theta_2(t) \\ \frac{1}{2}m_1l_1\cos\theta_1(t) & \delta_1 + \delta_2 & -\delta_2 - \frac{1}{2}m_2l_1l_2\dot{\theta}_2(t)\sin(\theta_1(t) - \theta_2(t)) \\ \frac{1}{2}m_2l_2\cos\theta_2(t) & -\delta_2 - \frac{1}{2}m_2l_1l_2\dot{\theta}_2(t)\sin(\theta_1(t) - \theta_2(t)) & \delta_2 \end{pmatrix} \begin{pmatrix} \dot{\theta}_0(t) \\ \dot{\theta}_1(t) \\ \dot{\theta}_2(t) \end{pmatrix} + \\ & + \begin{pmatrix} 0 \\ -\left(\frac{1}{2}m_1 + m_2\right)gl_1\sin\theta_1(t) \\ -\frac{1}{2}m_2gl_2\sin\theta_2(t) \end{pmatrix} = \begin{pmatrix} F(t) \\ 0 \\ 0 \end{pmatrix} \end{aligned} \quad (5)$$

where m_0 is the cart mass, m_1, m_2 are the pendula masses, l_1, l_2 are the pendula lengths, δ_0 is the friction coefficient of the cart, δ_1, δ_2 are the damping constants in the joints of the pendula, $J_1 = \frac{1}{3}m_1l_1^2, J_2 = \frac{1}{3}m_2l_2^2$ are the moment of inertia of the pendula with respect to the pivot points and $F(t)$ is the force induced on the cart.

The *rotary single inverted pendulum system* (Fig. 1b) consists of a pendulum rod attached to an arm rotating in a horizontal plane. Once again, the mathematical model of the system, as it was generated by the *Derivator*, is composed of two second-order nonlinear differential equations which respectively correspond to the rotary arm and the pendulum:

$$\begin{aligned}
 & \begin{pmatrix} J_0 + m_1 l_0^2 + \frac{1}{4} m_1 l_1^2 \sin^2 \theta_1(t) & -\frac{1}{2} m_1 l_0 l_1 \cos \theta_1(t) \\ -\frac{1}{2} m_1 l_0 l_1 \cos \theta_1(t) & J_1 \end{pmatrix} \begin{pmatrix} \ddot{\theta}_0(t) \\ \ddot{\theta}_1(t) \end{pmatrix} + \\
 & + \begin{pmatrix} \delta_0 & -\frac{1}{2} m_1 l_0 l_1 \dot{\theta}_1(t) \sin \theta_1(t) \\ -\frac{1}{8} m_1 l_1^2 \dot{\theta}_0(t) \sin 2\theta_1(t) & \delta_1 \end{pmatrix} \begin{pmatrix} \dot{\theta}_0(t) \\ \dot{\theta}_1(t) \end{pmatrix} + \\
 & + \begin{pmatrix} 0 \\ -\frac{1}{2} m_1 g l_1 \sin \theta_1(t) \end{pmatrix} = \begin{pmatrix} M(t) \\ 0 \end{pmatrix} \quad (6)
 \end{aligned}$$

where m_0 , m_1 stand for the masses of the arm and the pendulum, l_0 , l_1 are their respective lengths, δ_0 , δ_1 are the damping constants in the joints of the arm and pendulum, $J_0 = \frac{1}{3} m_0 l_0^2$ and $J_1 = \frac{1}{3} m_1 l_1^2$ are the moments of inertia of the arm and the pendulum with respect to their pivot points and $M(t)$ is the input momentum applied on the arm.

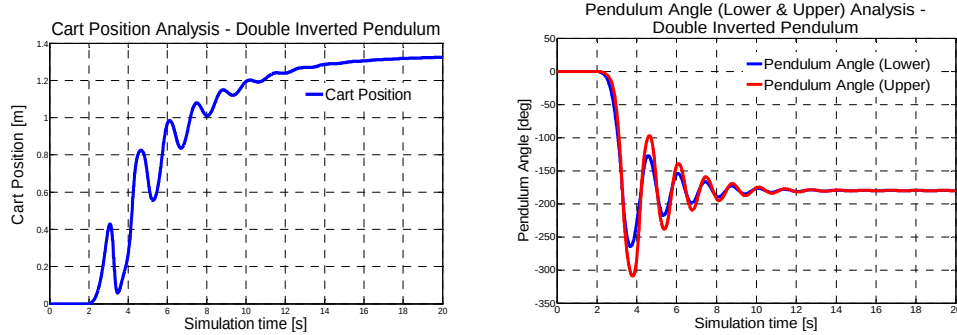


Fig. 3 Classical double inverted pendulum system time behavior – cart position and pendula angles

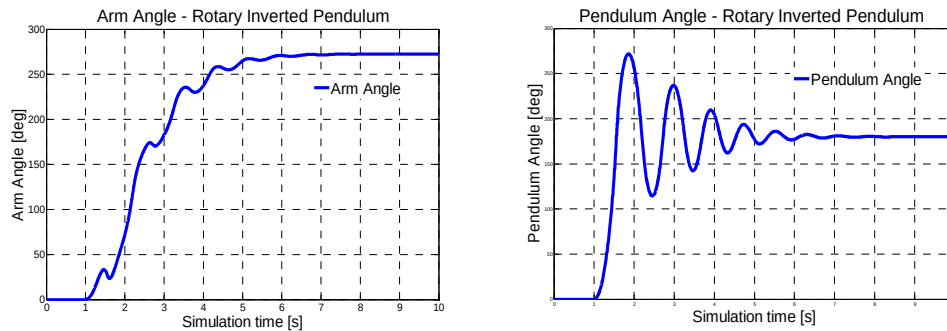


Fig. 4 Rotary single inverted pendulum system time behavior – arm and pendulum angles

Both models were included in the *IPMaC* block library in form of atomic Simulink blocks with a dynamic mask which supports useful features such as editable parameter constants and initial conditions as well as an adjustable number of input and output ports [1]. The obvious complexity of model equations is the price paid for a particularly accurate simulation model. To determine the behavior of the models in response to an impulse signal, simulation experiments were performed, to satisfactory results (Fig. 3, Fig. 4). Long-observed empirical findings about pendula behavior are confirmed: each pendulum of the system passes through oscillatory transient state until the system reaches the stable equilibrium point with all pendula pointing downward. The backward impact of the pendulum/pendula on the base (cart/arm), which increases with the weight of the load, is also visible. The models can therefore be considered accurate enough to serve as a reliable testbed for control algorithms.

III. CONTROL ALGORITHM BLOCKS

Controllability properties of inverted pendula systems were verified using both the double and rotary inverted pendulum system [3][4][5]. In order to meet the standard control objective in pendula systems, i.e. to stabilize all pendulum links in the upright, unstable position, *linear* methods of synthesis were used. As a result, linear approximation of the originally nonlinear inverted pendulum systems is required. The process can be considerably sped up with help of another GUI tool from the *IPMaC*: the *Inverted Pendula Model Linearizator & Discretizer*. In case the system type, model parameters and equilibrium point have been provided by the user, the tool generates the numeric state-space matrices of a linearized system using the standard (Taylor) series expansion around a given equilibrium point (Fig. 5)[3]. The discretized state-space matrices of the system are also returned if the sample time constant has been provided.

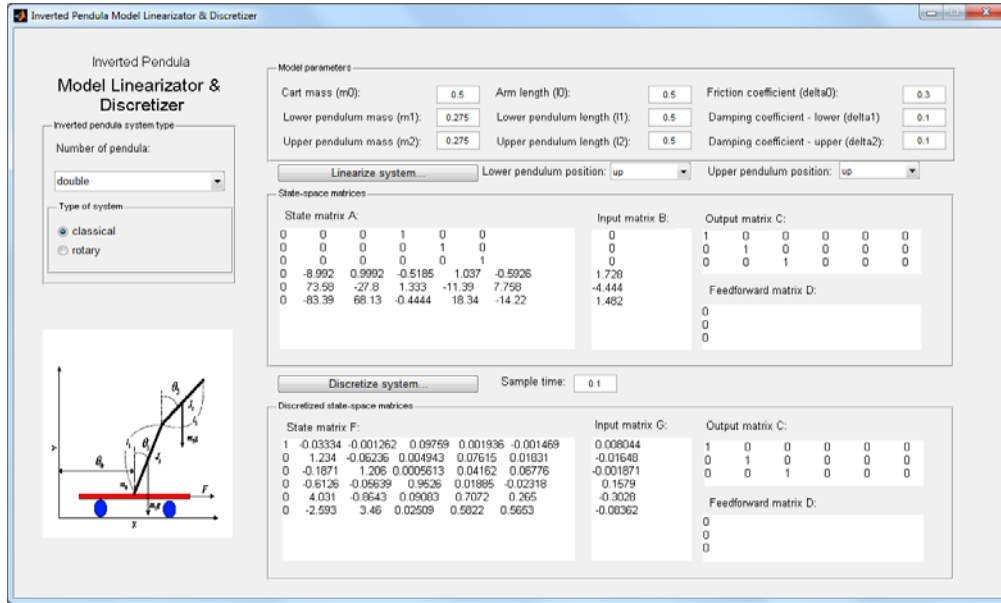


Fig. 5 *Inverted Pendula Model Linearizator & Discretizer* GUI tool in use:
 linearized and discretized state-space matrices of the double inverted pendulum system

All inverted pendula systems included in the *IPMaC* were modeled in a way which defines the “all upright” equilibrium as $\mathbf{x}(t) = \mathbf{x}_s = \mathbf{0}^T$. If $u(t) = u_s = 0$, the state-space description of the continuous linearized system is given as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{d}u(t) \end{aligned} \quad (7)$$

in case $\mathbf{A}$, $\mathbf{b}$, $\mathbf{C}$, $\mathbf{d}$ are the numeric continuous state-space matrices generated by the *Linearizator & Discretizer* tool.

To provide program support for inverted pendula control, linear state-space algorithms were implemented into Simulink and encapsulated into dynamic-masked control blocks. Most importantly, the *State Space Controller* block evaluates the relation

$$u(t) = u_f(t) + u_v(t) + d_u(t) = -\mathbf{k}\mathbf{x}(t) + \mathbf{k}_v\mathbf{w}(t) + d_u(t), \quad (8)$$

where $\mathbf{k}$ is the *feedback gain* which brings the system's state vector to the origin of the state space [3][4], $\mathbf{k}_v$ represents the *setpoint gain* which needs to be applied if a nonzero required value is specified and $d_u(t)$ is the unmeasured disturbance. To match an additional control objective (initial deflection, compensation of disturbance signal, tracking a reference position of the cart or a combination of the three), the block's appearance may be adjusted by optional enabling or disabling of the nonzero setpoint input $\mathbf{w}(t)$ and the disturbance input $d_u(t)$.

The *State Estimator* block implements the Luenberger state estimator:

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}\hat{\mathbf{x}}(t) + \mathbf{b}u(t) + \mathbf{L}(\mathbf{y}(t) - \mathbf{C}\hat{\mathbf{x}}(t)) \quad (9)$$

where $\mathbf{L}$ is the estimator gain matrix and $\hat{\mathbf{x}}(t)$ is the estimated state vector ([2][6]).

The *Demo Simulations* section of the *IPMaC* documents several simulation experiments which were performed to prove that both the LQR-designed controller and standard pole-

placement are able to meet the required control objectives for both systems in question. Fig. 6 and Fig. 7 show the simulation results in case the control objective was to maintain the desired cart position/arm angle while keeping the pendulum/pendula upright. Measurement limitations were simulated and an estimator block was included in both schemes to provide the controller block with a complete state-space vector (see [2] for the structure of control loop).

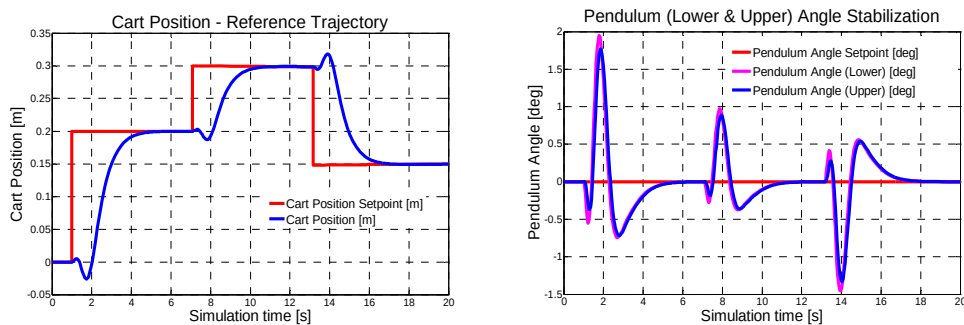


Fig. 6 Classical double inverted pendulum: simulation results for pole-placement control, estimator included

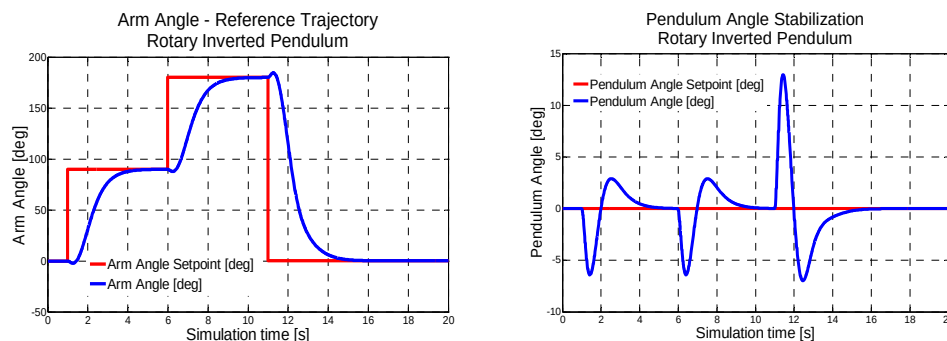


Fig. 7 Rotary single inverted pendulum: simulation results for LQR control, estimator included; note that the arm is supposed to rotate for exactly a half-circle before returning to its initial position

IV. CONCLUSION

The purpose of this paper was to propose an original conception of solving the task of modeling and control of the inverted pendula dynamical systems. Using the custom-designed Simulink block library *Inverted Pendula Modeling and Control* and custom program tools with graphical user interface, classical double and rotary single pendulum systems were analyzed and modeled. Algorithms of linear state-space control that stabilize the pendula in the inverted position were incorporated into the demo simulation archive of the library.

ACKNOWLEDGMENT

This work is the result of the project implementation: Development of the *Center of Information and Communication Technologies for Knowledge Systems* (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

REFERENCES

- [1] S. Jadlovská – A. Jadlovská, A Simulink library for inverted pendula modeling and simulation. In: 17th Annual Conference Proceedings of the International Scientific Conference - Technical Computing Prague 2009, Congress Center of Czech Technical University in Prague, November 19, 2009, ISBN 978-80-7080-733-0.
- [2] S. Jadlovská – A. Jadlovská, Inverted pendula simulation and modeling – a generalized approach. In: Proceedings of the 9th International Scientific – Technical Conference on Process Control '10, Říp 2010, June 7-10, University of Pardubice, Czech Republic, ISBN 978-80-7399-951-3.
- [3] M. Demirci, Design of feedback controllers for a linear system with applications to control of a double inverted pendulum. In: International Journal of Computer Cognition, Vol. 2, No. 1, 2004.
- [4] J. Sarnovský – A. Jadlovská – P. Kica, Optimal and adaptive systems theory [in Slovak]. ELFA s.r.o., Košice 2005, ISBN 80-8086-020-3.
- [5] P. Ernest, P. Horáček, Algorithms for control of a rotating pendulum. In: MED'03, National Technical University in Athens, Greece, 2003.
- [6] J. Mertl, J. Sobota, M. Schlegel, P. Balda, Swing-up and stabilization of a rotary inverted pendulum. In: 15th International Conference on Process Control 2005, Štrbské Pleso, Slovakia 2005.

Extrakcia informácií pre podporu rozhodovania v oblasti obchodného práva

Peter SMATANA

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

psmatana@gmail.com

Abstrakt – Článok pojednáva o aplikovanom výskume v rámci projektu ITLIS, ktorého cieľom je vyvinúť vyhľadávací a vizualizačný nástroj nad dátami integrovanými z rozličných zdrojov. Zdrojom dát sú verejne prístupné registre obsahujúce informácie hlavne o ľuďoch, firmách a iných organizáciach. Jedným zo spracovávaných zdrojov je aj obchodný vestník, v ktorom sú niektoré sekcie prezentované vo forme neštruktúrovaného textu, z ktorého je potrebné extrahovať dôležité informácie. V článku je prezentovaný návrh systému slúžiaci na extrakciu týchto informácií a experimentálne overenie existencie vzorov, ktoré budú používané pri vytváraní extrakčných šablón.

Kľúčové slová – extrakcia informácií, rozpoznávanie názvoslovných entít, spracovanie prirodzeného jazyka, obchodné právo

I. ÚVOD

Obchodné právo ako jeden z pilierov legislatívneho systému právnych štátov upravuje právne postavenie podnikateľov a vzťahy, do ktorých podnikatelia navzájom vstupujú pri uskutočňovaní podnikateľskej činnosti. Na kontrole dodržiavania vzťahov medzi podnikateľmi sa podieľa štát ale aj podnikatelia navzájom. Informovanosť ohľadom týchto vzťahov môže znamenať konkurenčnú výhodu, či uchránenie si svojich investícií. K dispozícii je viacero zdrojov ako si overiť svojich obchodných partnerov:

- obchodný vestník¹
- obchodný register Slovenskej Republiky²
- živnostenský register Slovenskej Republiky³
- sociálna sieť firiem v Slovenskej republike foaf.sk (<http://www.foaf.sk>)
- vizuálny obchodný register Slovenskej republiky (<http://www.vorsr.sk>)
- univerzálny register Slovenskej republiky (<http://www.ur.sk>)

Každý systém má svoje pre a proti, svoje obmedzenia a limity. Prvé tri zdroje sú spravované ministerstvami spravodlivosti a vnútra Slovenskej Republiky a slúžia ako hlavné zdroje združujúce informácie o jednotlivých podnikateľských subjektoch, no ich použitie je dosť neprehľadné a z toho dôvodu vznikli aplikácie, ktoré poskytujú jednotný pohľad nad konkrétny subjekt a taktiež vizualizáciu jeho histórie a prepojenia na iné podnikateľské subjekty. Avšak tieto aplikácie čerpajú hlavne zo štruktúrovaných dát prezentovaných v dostupných registroch a tie zaujímavejšie vzťahy, ktoré je možné objaviť v publikovaných neštruktúrovaných dokumentoch (napr.: rozhodnutia súdov vo veciach konkurzov a reštrukturalizácii, oznámenia o začatí konania o zrušení spoločnosti bez likvidácie, atď.) ostávajú nespracované a systémy ich zvyčajne len full-textovo zaindexujú. Tým sa strácajú dôležité informácie s ktorými pracuje každý právnik venujúci sa obchodnému právu ale aj organizácie, ktoré sa snažia uplatniť si nárok na vyrovnanie dlhu dlžníkom, ktorý je v konkurznom konaní.

¹online verzia: <http://www.justice.gov.sk/yyyy.aspx>

²online verzia: <http://www.orst.sk>

³online verzia: <http://www.zrsr.sk>

II. ANALÝZA PROBLÉMU

Analýza neštrukturovaných dokumentov v prirodzenom jazyku je vo všeobecnosti veľmi zložitá a použiteľné výsledky sa dosahujú v prípadoch ak extrahujeme informácie v dokumentoch zo špecifickej domény (lekárske záznamy pacientov[1], správy z poistných udalostí a iné) alebo ak z dokumentov z rôznych domén extrahujeme špecifické informácie (citácie literatúry, identifikácia stretnutí v emailovej komunikácii pre automatické pridávanie udalostí do organizéra, a iné). Momentálne neexistujú systémy, ktoré by boli schopné ľubovoľný textový dokument analyzovať a transformovať ho do reprezentácie, ktorá by bola použiteľná pre zodpovedanie otázok, hľadanie konkrétnych faktov či iných úloh, ktoré človek robí automaticky pri práci s textovými dokumentami.

Problémy súvisiace so spracovaním prirodzeného jazyka (viac nájdete v [2]) sú pre nami zvolenú doménu minimalizované pretože dokumenty aj keď sú v neštrukturovanej forme, je možné na základe špecifickosti jazyka právnikov a pevných formulácií objavovať štandardné vzory. Tieto vzory sú použiteľné pre extrakciu požadovaných informácií, ktoré môžu slúžiť po ďalšej analýze ako pomocný prvok pri rozhodovaní vo sfére obchodu.

Sekcia "Konkurzy a reštrukturalizácie" v obchodnom vestníku obsahuje neštrukturovaný text (viď. Obr. 1), ktorý je potrebné transformovať do konkrétného scenára. Pre vzorový dokument bol ručne vytvorený scenár zobrazený na Obr. 2. Tieto scenáre sú navrhované v spolupráci s expertom v danej doméne. V našom prípade právnik venujúci sa obchodnému právu, pretože on je hlavnou cieľovou skupinou používateľov a on definuje požadovanú funkčnosť systému.

K000645
sp.zn.: 1K/56/2010

Okresný súd Banská Bystrica v právnej veci navrhovateľa - dlžníka **Vladimír Šáli**, nar. 20.10.1969, bytom Majerský rad 59, 963 01 Krupina, prechodne bytom Školská 10, 963 01 Krupina, o návrhu na vyhlásenie konkurzu na jeho majetok, takto rozhodol

Vyhlásuje konkurz na majetok dlžníka Vladimír Šáli, nar. 20.10.1969, bytom Majerský rad 59, 963 01 Krupina. Otvára malý konkurz.

Ustanovuje do funkcie správcu konkurznej podstaty JUDr. Andrea Balážiková, so sídlom kancelárie Partizánska cesta 3, 974 01 Banská Bystrica.

Súd vyzýva veriteľov, aby prihlásili svoje pohľadávky v lehote 45 dní od vyhlásenia konkurzu. Súd ukladá správcovi povinnosť bez zbytočného odkladu informovať o vyhlásení konkurzu známym veriteľom dlžníka, ktorí majú trvalé bydlisko alebo registrované sídlo v iných členských štátoch Európskej únie ako v Slovenskej republike v súlade s č. 40 Nariadenia Rady (ES) č. 1346/2000 z 29. mája 2000 o konkurznom konaní.

Poučenie:

1. Proti uzneseniu o vyhlásení konkurzu môže podať odvolanie dlžník, ak nie je navrhovateľom a to v lehote 15 dní odo dňa jeho doručenia (odo dňa nasledujúceho po dni zverejnenia tohto uznesenia v Obchodnom vestníku) prostredníctvom tunajšieho súdu ku Krajskému súdu v Banskej Bystrici. V odvolaní sa má podľa všeobecných

ktoré nespina námietky zaujatosti, sa neprijímajú; v tomto prípade sa vec nadriadenému sudu neprekladá (§ 15a ods. 3 Občianskeho súdneho poriadku).

17. Dlžník-fyzická osoba má právo sa po zrušení konkurzu za podmienok ustanovených týmto zákonom domáhať na súde zbavenia svojich dlhov (§ 166 ods. 1 ZKR veta prvá). Dlžník nemá právo domáhať sa zbavenia svojich dlhov, ak bol konkurz zrušený preto, že majetok dlžníka nepostačoval ani na úhradu pohľadávok proti podstate. Návrh na oddĺženie je oprávnený podať dlžník, ktorý je fyzickou osobou, spolu s návrhom na vyhlásenie konkurzu, prípadne počas konkurzného konania až do zrušenia konkurzu. Návrh na povolenie oddĺženia okrem všeobecných náležitostí návrhu musí obsahovať aj odôvodnenie, ktoré vyjadruje poctivý zámer dlžníka vynaložiť primerané úsilie na uspokojenie svojich veriteľov (167 ods. 1 ZKR).

Okresný súd Banská Bystrica dňa 17. 1. 2011
JUDr. Eva Miháľechová, samosudca

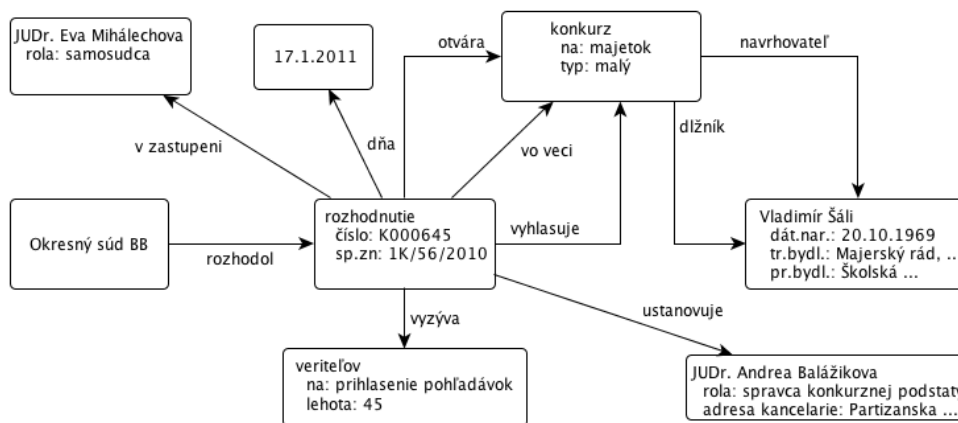
Obr. 1 Vzor uznesenia z obchodného vestníka

III. NÁVRH SYSTÉMU

Základná architektúra systému je prezentovaná na Obr. 3, kde celá extrakčná časť obchodného vestníka (OV) ukladá dáta do už pripraveného dátového úložiska. Toto centrálné úložisko integruje dáta z viacerých dátových zdrojov a poskytuje ich používateľovi pomocou webového rozhrania. Jednotný pohľad na entity, je možný vďaka agregátoru entít, ktorý sa pokúša identifikovať entity referujúce na tú istú osobu či organizáciu. Identifikácia zhodnosti entít je na základe vlastností extrahovaných entít osôb (meno, priezvisko, tituly, adresa, rola, trvanie danej role) a vlastností extrahovaných entít organizácii (názov, adresa, IČO).

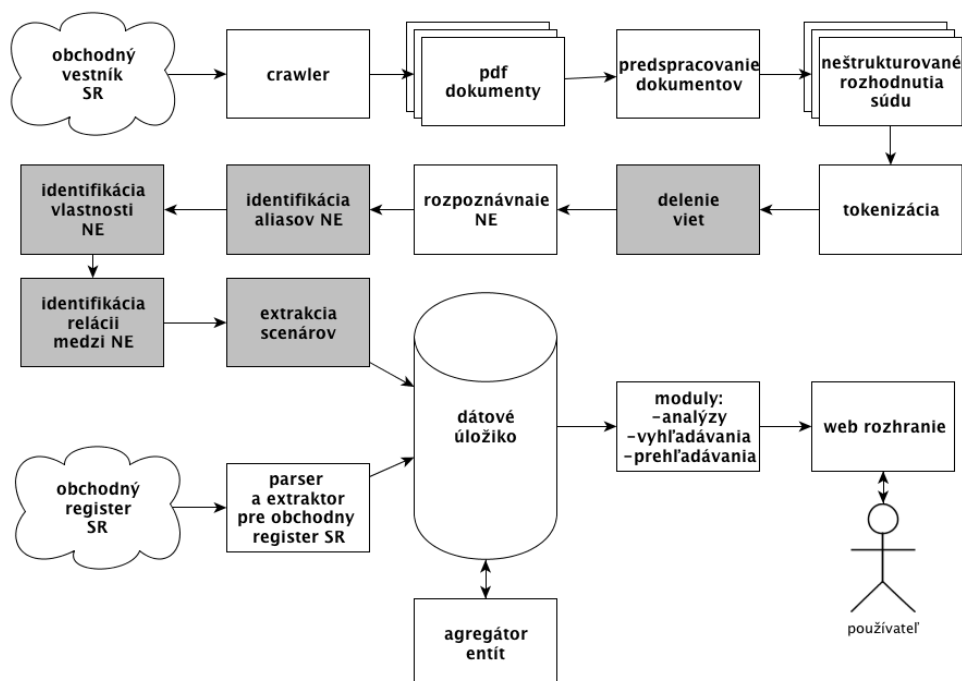
Dokumenty z OV sú k dispozícii vo forme pdf dokumentov, ktoré je potrebné previesť do formy textových dokumentov s UTF-8 kódovaním a následne rozdeliť na jednotlivé rozhodnutia súdu. Úlohou tokenizácie je rozdeliť text na jeho arbitrárne jednotky tokeny[3], ktoré nasledujúci modul spojí do viet (členenie textu do viet je dôležité hlavne pre modul identifikácie aliasov NE).

Ďalšie moduly systému vychádzajú z piatich typov úloh v extrakcie informácií, ktoré boli definované MUC (Message Understanding Conference) konferenciami[4][5]:



Obr. 2 Vzor extrahovaného scenára

- Rozpoznávanie pomenovaných entít - nájdenie a identifikácia pomenovaných objektov (NE) v texte (dátumy, osoby, adresy, organizácie, sumy peňazí).
- Identifikácia aliasov NE - zahŕňa hlavne identifikáciu relácie identity medzi NE a entitami v texte (napr.: okresný súd Bratislava I je neskôr v texte uvádzaný ako súd a pritom je to tá istá entita, identifikácia zamlčaného podmetu vo vete).
- Šablónové prvky(TE) – priradenie jednoducho identifikovateľných vlastností NE (v okolí názvu firmy sa vyskytuje jej IČO, v okolí človeka sa nachádza jeho dátum narodenia)
- Šablónové relácie(TR) – nájdenie a vyjadrenie relácií medzi NE (konkrétne rozhodnutie ustanovuje konkrétneho správcu)
- Scenárové šablóny – mapujú TE a TR do špecifických scenárov (viď. Obr. 2)



Obr. 3 Architektúra systému (moduly zobrazené tmavšou farbou sú vo fáze vývoja alebo implementácie)

IV. EXPERIMENT

Cieľom experimentu je experimentálne overenie existencie vzorov, ktoré budú používané pri vytváraní extrakčných šablón. Na vzorke 4664 dokumentov zo sekcie "Konkurzy a reštrukturalizácie" obchodného vestníka z roku 2009 sme spustili analyzátor spoluvýskytov slov. Táto analýza hľadala spoluvýskyt až do úrovne desiatich slov.

Tabuľka 1
VÝSKYTY NAJZAUJÍMAVEJŠÍCH FRÁZ V 4664 DOKUMENTOCH OV

fráza	výskyt v dokumentoch	poznámka
v právnej veci navrhovateľa-dlžníka	1119	konkrétny súd vydal rozhodnutie v určitej veci kde sú definované roly pre entitu nasledujúcu za frázou entita je navrhovateľ a zároveň aj dlžník
v právnej veci navrhovateľa-veriteľa	559	entita nasledujúca za frázou je navrhovateľ a zároveň veriteľ
v právnej veci navrhovateľa	374	entita nasledujúca za frázou je navrhovateľ
v právnej veci vyhláseného konkurzu	1000	vec v rámci ktorej vydal súd rozhodnutie
o návrhu na vyhlásenie konkurzu na majetok	784	fráza definujúca v akej právnej veci je dané rozhodnutie
v lehote 15 dní	782	súdom definovaná lehota na uplatnenie si svojich nárokov
v lehote 45 dní	353	súdom definovaná lehota na uplatnenie si svojich nárokov
so sídlom kancelárie:	2609	najčastejšie sa táto fráza spája so sídlom správcu konkurznej podstaty
konkurzné konanie voči dlžníkovi	651	fráza definujúca zahájenie alebo ukončenie konkurzného konania voči dlžníkovi
pre nedostatok majetku	428	dôvod zastavenia konkurzného konania

Jedinečných fráz z frekvenciou výskytu nad 250 je 4990, no mnohé z nich nie sú pre analýzu dôležité. Väčšina z nich je zo sekcie "poučenie" (viď.: Obr. 1), ktoré má štandardnú formu no z hľadiska informačnej hodnoty pre používateľa nie je dôležitá. Najzaujímavejšie frázy sú prezentované v Tabuľke 1.

V. ZÁVER

I keď analýza neštrukturovaných dokumentov v prirodzenom jazyku je veľmi náročná na kvalitne spracovanie pomocou lingvistických metód (morfologická, syntaktická, sémantická, kontextová a pragmatická analýza), tak na základe striktného jazyka právnikov a experimentálnym objavením viacerých vzorov očakávame vysokú kvalitu extrahovaných dát.

Na základe tohoto experimentu a deklarovanej pridanej hodnoty pre cieľovú skupinu používateľov sme sa rozhodli implementovať tento systém. Momentálne je navrhovaný systém modulov na extrakciu dát z obchodného vestníka do prezentovaného modelu vo výskumno implementačnej fáze.

LITERATÚRA

- [1] J. Paralič and P. Smatana, "Transformation of free-text electronic health records for efficient information retrieval and support of knowledge discovery," in *Proc. of the 16th International Conference on Information and Intelligent Systems (IIS 2005)*, Varaždin, Croatia, September 2005, 2005, pp. 139–144.
- [2] E. Páleš, *Sapfo – Parafrazovač slovenčiny.*, 1st ed. Bratislava: VEDA, 1994, no. ISBN 80-224-0109-9.
- [3] R. Garabík, L. Gianitsová, A. Horák, and M. Šimková, "Tokenizácia, lematizácia a morfológická anotácia slovenského národného korpusu." 2004, interný materiál.
- [4] H. Cunningham, "Information extraction, automatic," *Encyclopedia of Language and Linguistics.*, 2005.
- [5] J. Paralič, K. F. Furdík, G. Tutoky, P. Bednár, M. Sarnovský, P. Butka, and F. Babič, *Dolovanie znalosti z textov.* Equilibria, s.r.o., Košice, 2010, no. ISBN 978-80-89284-62-7.

Solder Paste Dispensers and Their Parameters

Alena PIETRIKOVÁ, Michal KRAVČÍK, Juraj BANSKÝ

Dep. of Technologies in Electronics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

alena.pietrikova@tuke.sk, michal.kravcik@tuke.sk, juraj.bansky@tuke.sk

Abstract — Solder pastes are key materials in surface mount technology (SMT) for assembly of printed circuit boards (PCBs). As the trend towards miniaturisation of electronic products continues, there is an increasing demand for better understanding of the flow and deformation that is, the rheological behaviour of solder paste formulations. Wall slip effect that is connected with requirement of very accurate (shape and volume) solder paste deposition plays an important role in characterising the flow behaviour of solder paste materials. The aim of this study is to investigate the influence of the solder paste formulation on wall-slip formation and its effect on the printability of these pastes material. Paper refers to dynamics of printing parameters for one of the most common methods of solder paste deposition – dispenser and its parameters.

Keywords — dispensing, solder paste, viscosity of solder paste

I. INTRODUCTION

Over the last 30 years, the SMT assembly process has become increasingly more sophisticated. There are two primary methods of applying solder paste to a PCB: Stencil printing and dispensing. Which method will be used for deposition depend on the size and the type of assembled components, availability of technology, massiveness of production and etc. While each method has its advantages and disadvantages, this study focuses on dynamics and different aspects of dispenser deposition on the solder paste print deposition quality. There is a wide opinion in the industry that the paste printing process accounts for the majority of assembly defects. Experience with this process has shown that typically over 60% of all soldering defects are due to the problems associated with deposition process of solder paste. Those defects faults are a major source of board failure. Unfortunately, understanding the causes of those defects can be challenging because there are numerous factors to consider. There is several factors influence to printing quality of solder paste, which has equivalent relevance:

- **Environment:** temperature of solder paste during storage and during printing, room temperature, humidity, solder paste storage time, solder paste shelf time, etc.
- **Solder paste formulation and material:** alloy type, powder size, manufacture date, water soluble, viscosity, thixotropy, relaxation time, flow curve, metal content, flux type, tack time and force, etc.
- **Equipment:** type of dispenser, needle size, printer alignment, optical inspection system.
- **Human factor:** employee skills and knowledge, part of day, operator errors and training frequency.

The paste dispensers themselves are not very accurate. Most are reciprocating bulk dispensers that apply air pressure to the back of the stopper in the cartridge. The time it takes to pressurize the cartridge varies significantly through its usage, depending on how much paste remains in the tube. As a result, the actual amount of solder paste dispensed varies significantly from the first to the last dispense cycle. Lastly, when dispensing a large amount of solder paste with little paste remaining in the tube, the full amount required may not be dispensed, adding further error to the total amount of solder left on the stencil.

Controlling these important parameters is not an easy task with many of today's printers. However, as we continue to push the envelope into smaller and finer pitch devices, it becomes even more critical to fully understand and minimize the effects of even the smallest variation.

II. DEPOSITION OF SOLDER PASTE BY DISPENSER

Deposition solder paste by dispenser is semi or full automatic process, where solder paste is pushed out from cavity needle. Fully automatic dispenser is move along PCB controlled by computer. Fully automated dispensers have very accurate position of deposition. This type of dispensers is use in industry for fine pitch applications of for adhesive applications.

Dispenser advancement:

- fast, easy and cheap change of motive,
- possible of deposition different volumes of solder paste,
- very accurate deposition of small volumes with ball shape,
- proper method for development or small series production,
- can by use for partly assembled PCBs or for component replacement.

Dispenser disadvantage:

- PCB speed deposition depends on the number of solder pads,
- necessary thixotropy preparation of solder paste (sacrifice solder paste on deposition start).

There are used generally two types of dispensers with different principle dispensing solder paste (Fig. 1). Dispenser with Archimedes screw dosing solder paste by revolution screw. Advancement of this type is very accurate deposition of small volumes with high repeatable [6], [7]. The second type of dispenser use principle piston pressure on container [6], [7]. Advancement of this type is low price.

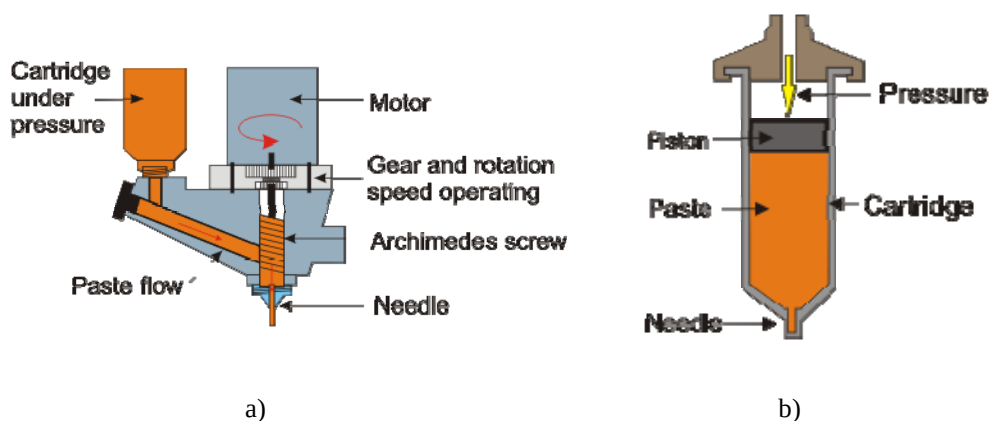


Fig. 1 a) Principle of deposition solder paste with Archimedes screw dispenser,
b) principle of deposition solder paste with piston pressure dispenser

Solder pastes use in dispenser have lower viscosity (100 – 150 Pa.s), less powder size (5- 45 μm) [6], [7], less metal content (75 – 85%) [6], [7] and higher thixotropy too in compare to pastes recommended for stencil printing or screen printing. Correct setup of dispenser by parameters of solder paste is especially critical for solder paste deposition in ball volumes for ultra fine pitch components (μBGA). The most important task in deposition process is high accurate volume deposition on every single soldering pad. Only correct deposition of solder paste can guaranty correct joint function, for example BGA components [4].

Quality of deposition process can by evaluate by this parameters:

- speed of deposition,
- volume consistence of solder paste,
- shape stability of deposited solder paste.

Speed of deposition means number of correct deposited solder paste balls per minute, which during wrong dispenser setup markedly decrease.

Volume consistence of solder paste is connected with viscosity and means volume divergence from required volume. When maximum allowed divergence is reach, dispensing process should be stop and setup correct printing parameters.

Shape stability of solder paste is evaluated by measuring solder paste ball dimensions and shape. When allowed limits of dimension are reach, should by dispensing process stop and setup correct printing parameters (Table 1.)

Effect of solder paste parameters and dispenser setup parameters on those three parameters

(speed of deposition, volume consistence of solder paste, shape stability of deposited solder paste) is shown in Table 1. For more precision solder paste deposition is approve dispenser system with Archimedes screw than piston pressure dispenser system. Archimedes screw system is able to use thinner needle and also can deposit smaller solder paste volumes, which is important for ultra fine dot application [8].

Table 1
 COMPARISON OF SOLDER PASTE PARAMETERS ON QUALITY AND STABILITY OF DEPOSITION RESULT [8]

Parameters	Speed of deposition	Volume consistence	Shape stability
Viscosity	For higher speed is required higher viscosity.	Unwatched effect.	Low viscosity has tendency to separate flux from solder paste. This can cause shape stability problem.
Powder size	Small powder size allows fast deposition with high pressure.	Unwatched effect.	Small size powder has tendency to lower shape stability because wetting of high surface problem.
Metal content	Less metal content suited for fast deposition with high pressure	Low metal content can affect low consistence. Higher filling pressure it can improve.	Higher metal content can affect compactness of solder paste there through also shape stability.
Thixotropy preparation	For higher speed deposition is necessary higher thixotropy preparation.	For longer thixotropy preparation except good paste consistence.	Better stability with longer thixotropy preparation.
Flux temperature activation	Unwatched effect.	Unwatched effect.	Shape stability can by low with low flux temperature activation during cyclic deposition with high pressure (evaporation of flux components can change viscosity).
Needle size	Coarser needles are proper for high speed deposition. Thinner needles are proper for fine dot application.	Low volume consistence with thinner needles.	Thinner needles inclined to choke.
Filling pressure	High pressure for high speed.	High consistence with high pressure when Archimedes screw system used.	High pressure with piston pressure dispenser system can affect shape stability.
Type of dispenser	Archimedes screw system can by use for higher speed deposition.	Piston pressure dispenser can have lower consistence of solder paste volume deposition.	Piston pressure dispenser can have lower shape stability.

Shape stability, thickness and homogeneity of deposited solder paste is periodically optical evaluate. In particle are used two different principles for evaluate solder paste deposition quality. First one principle measure shape and volume by laser raster scan of solder paste surface, when output is digital 3D image (Fig. 2a). The software equipment of these devices can calculate statistic information about height, area, deposited volume and their divergence. The second principle of evaluate deposition quality is based on optic microscopy (Fig. 2b). The software equipment of these devices after calibration can measure solder paste deposited areas, but measurement of height and volume is only informative.

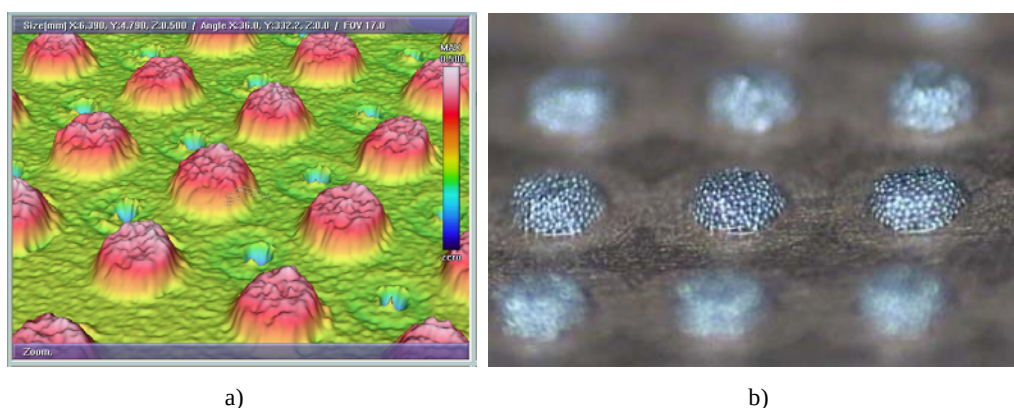


Fig. 2 a) 3D output from SPI (Solder Paste Inspection) device. Colour of surface means height level of deposited solder paste. b) Detail of deposited solder paste for BGA component taken by optical device Optilia.

III. CONCLUSION

Robust dispensing process requires that we control as many sources of variation as possible. Of course, not all of these variables are easy to control, but it is helpful to understand which ones are the largest sources of variation and target these. Controlling these important parameters is not an easy task with many of today's dispensers. However, as we continue to push the envelope into smaller and finer pitch devices, it becomes even more critical to fully understand and minimize the effects of even the smallest variation.

Ambition of most electronic application producers is their minimalization. Application smaller components and then smaller PCBs is possible reduce quantity of material required for article production. Minimalization in this surface mounts technology press on producers to use still more optimal and more accuracy manufacture technology. Requirement build up quality electronic joint was and still is mainly impulse for development in solder paste deposition technology. Ambition is build up good quality joint with the longest life time. Properties of solder pastes like a viscosity, quantity and type of flux, quality of deposition, metal content and metal alloy quality, solder paste properties during reflow process are most common parameters, which assist extend status of knowledge in solder paste domain.

Archimedes pumps rely on material flow characteristics such as viscosity and shear behaviour to force material down the screw and out the needle. Material viscosity affects the volume of material dispensed, and is itself affected by the dispensing temperature and the rate of shear due to the rotation of the auger screw. Any changes in material flow characteristics affect the volume of material dispensed. Therefore, when using Archimedes pumps, it is important to maintain stable dispensing cabinet and material temperatures to maintain consistent material viscosity. Archimedes pumps have a definite advantage over time and pressure dispensing—dispensing consistency. The advantage Archimedes pumps have over piston pumps is a little more sublime, however. The major advantage is the flexibility of unlimited dot sizes. With Archimedes pumps, dot size is simply adjusted through the software, and is proportional to the length of time the screw is allowed to rotate. Throughput is compromised with Archimedes pumps when there is a need for dispensing larger components. This is because the screw has to turn longer to get the necessary volume of material. The limitations for screw pumps in getting smaller dot sizes relate to the electrical and mechanical abilities of the clutch or motors to turn on and off. Actuators of 10 ms or less are approaching the mechanical limitations, and repeatability can be a problem.

ACKNOWLEDGEMENT



This paper was developed with support of the project "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií" v oblasti automobilovej elektroniky" (Centre of Excellence of the Integrated Research & Exploitation the Advanced Materials and Technologies in the Automotive Electronics)", ITMS 26220120055, that is co-financed from Structural Funds EU ERDF within Operational programme Research and Development OPVaV-2009/2.1/03-SORO and preferred axis 2 Support of Research and Development.

REFERENCES

- [1] SZENDIUCH I.: Základy technológie mikroelektronických obvodov a systémov, ISBN 80-214-3292-6, Brno 2006
- [2] BIOCCA, P.: Lead-free SMT Soldering Defects How to Prevent Them, Kester, October 2004
- [3] HUI, K., RAPLH, B.: Voiding and printing characteristics of solder paste at different degrees of exposure, Proc. Instn. Mech. Engrs. Vol. 214 Part B, IMechE 2000
- [4] JIANBIAO P., TONKAY G., STORER R. H., SALLADE R. M., LEANDRI D. J.: Critical Variables of Solder Paste Stencil Printing for Micro-BGA and Fine-Pitch QFP, IEEE TRANSACTIONS ON ELECTRONICS PACKAGING MANUFACTURING, VOL. 27, NO. 2, APRIL 2004
- [5] Cheng Lee, N.: Reflow soldering processes and troubleshooting: SMT, BGA, CSP, and flip chip technologies. ISBN 0-7506-7218-8
- [6] Koki Solder Paste at Manual, <http://www.christopherweb.com/datasheet/Koki>
- [7] Practical Component Handbook, http://fl.hw.cz/docs/smt/pci_cat.pdf
- [8] Ludwing, R.D., Lee, N.CH.: Achieving Ultra-Fine Dot Solder Paste Dispensing

Vplyv vlhkosti na funkčnosť súčiastok

Eubomír LIVOVSÝ, Juraj BANSKÝ, Alena PIETRIKOVÁ

Katedra technológií v elektronike Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

lubomir.livovsky@tuke.sk, juraj.bansky@tuke.sk, alena.pietrikova@tuke.sk

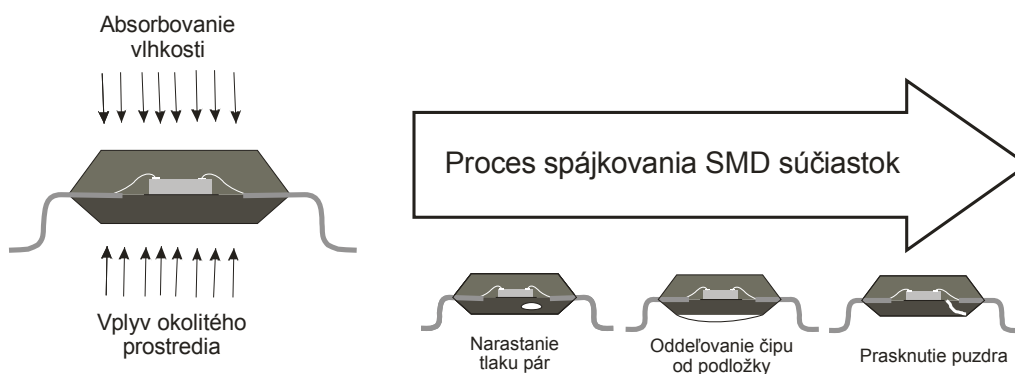
Abstrakt — Článok opisuje vplyv pôsobenia vlhkosti na plastové puzdra súčiastok určených pre povrchovú montáž. Vlhkosť, ktorá sa môže dostať do plastového puzdra súčiastok počas ich skladovania resp. dopravy v dôsledku nevhodnej manipulácie, sa môže nepriaznivo prejavíť vytvorením prasklín, alebo trhlín vo vnútri puzdra súčiastok. Poškodenie súčiastok vzniká počas procesu spájkovania.

Kľúčové slová — citlivosť, skladovanie, spájkovanie, sušenie, vlhkosť, trhlina

I. ÚVOD

Prvé puzdra súčiastok pre povrchovú montáž SMD (Surface Mount Devices) boli dostupné v keramických puzdách, ktoré boli hermetické a preto neboli citlivé na vlhkosť. Požiadavka zmenšovania veľkosti súčiastok a ich cenovej dostupnosti si vyžiadala použitie plastových puzdiel integrovaných obvodov ako napríklad QFP (Quad Flat Pack), BGA (Ball Grid Arrays), SOP (Small Outline Package). Materiál, z ktorého sú tieto puzdra vyrábané je pórovitý, označovaný ako nehermetický. Plastové puzdra sú preto náchylné na absorbovanie vlhkosti z okolitého prostredia do vnútra puzdra. Absorbovanie vlhkosti z prostredia je možné charakterizovať parametrom *citlivosťou súčiastok na vlhkosť*. Tento parametre je možné klasifikovať v niekoľkých stupňoch – MSL (Moisture Sensitive Level). Vplyv absorbovanej vlhkosti v puzdre súčiastok sa nepriaznivo prejavuje dvoma spôsobmi:

- Vlhkosť obsiahnutá v puzdre SMD súčiastky sa aktivuje v procese spájkovania. Pri spájkovaní v pretavovacej peci sa za relatívne krátku dobu ohreje súčiastka na vysokú teplotu (max. 260°C), čo má za následok premenu vlhkosti v puzdre súčiastky na tlak pár, ktorý spôsobuje prasknutie alebo roztrhnutie puzdra, oddeľovanie (delamináciu) plastovej časti puzdra od kovových vývodov, alebo roztrhnutím kontaktov medzi čipom a vývodmi súčiastok (Obr. 1).
- Druhý spôsob vplyvu vlhkosti je vznik korózií kovových štruktúr vo vnútri puzdra súčiastky.



Obr. 1 Degradácie zmeny v puzdre SMD súčiastky počas spájkovacieho procesu

II. KLASIFIKÁCIA SÚČIASTOK CITLIVÝCH NA VLNKOSŤ

Súčiastky pre povrchovú montáž, ktoré sú zapuzdrené v nehermetických puzdách, sa klasifikujú podľa IPC/JEDEC štandardu J-STD-20 „Moisture/Reflow Sensitivity Classification for Nonhermetic Solid State Surface Mount Devices“ stupňom citlivosti na vlhkosť, ktorú

vykonáva výrobca SMD súčiastok, výrobcovia informujú predajcov, obchodníkov (spôsob a podmienky skladovania) a výrobcov elektronických zariadení (proces spájkovania), ako sa má správne so súčiastkami manipulovať, aby sa predišlo ich zničeniu pri ich pretavovaní. Na základe doporučení štandardu J-STD-20 výrobca zatriedi súčiastky do stupňa citlivosti na vlhkosť podľa tabuľky 1. V tabuľke sú uvedené podmienky klasifikácie pre jednotlivé stupne citlivosti, podľa ktorých vykonáva klasifikáciu výrobca a tiež podmienky skladovania súčiastok po otvorení ochranného obalu.

Tabuľka 1
 KLASIFIKAČNÉ STUPNE CITLIVOSTI SÚČIASTOK NA VLNKOSŤ

Stupeň citlivosti na vlhkosť ¹	Skladovanie súčiastok		Štandardné podmienky klasifikácie	
	Čas ¹	Parameter	Čas [hod]	Parameter
1	Neobmedzene	≤ 30°C/85% RH	168	85°C/85% RH
2	1 rok	≤ 30°C/60% RH	168	85°C/85% RH
2a	4 týždne	≤ 30°C/60% RH	696	30°C/60% RH
3	168 hodín	≤ 30°C/60% RH	192	30°C/60% RH
4	72 hodín	≤ 30°C/60% RH	96	30°C/60% RH
5	48 hodín	≤ 30°C/60% RH	72	30°C/60% RH
5a	24 hodín	≤ 30°C/60% RH	48	30°C/60% RH
6	DUO ²	≤ 30°C/60% RH	DUO ²	30°C/60% RH

Poznámky: 1 - čas skladovania sa meria od momentu odstránenie ochranného obalu.

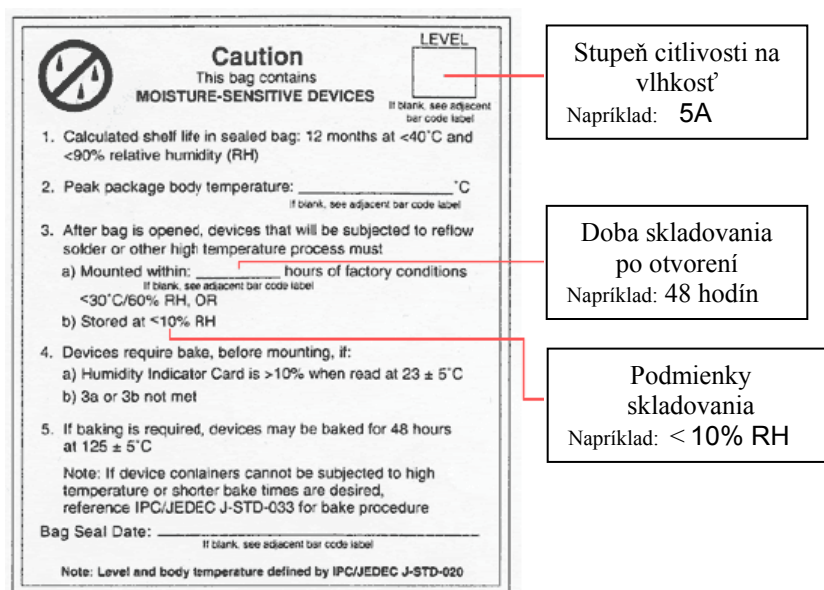
2 - doba musí byť uvedená na ochrannom obale súčiastky.

3 - RH relatívna vlhkosť.

Ak sú SMD súčiastky po klasifikácii výrobcom zaradené do 1. stupňa znamená to, že nie sú citlivé na vlhkosť a tieto súčiastky nie je potrebné skladovať za špeciálnych podmienok.

Ak sú súčiastky klasifikované stupňom väčším ako 1, potom sú považované za súčiastky citlivé na vlhkosť. Na ich skladovanie a prepravu je potrebné použiť obaly, ktoré neprepúšťajú vlhkosť a do vnútra obalov je potrebné vkladať materiál pohlcujúci vlhkosť (silikagel, siran vápenatý) vlhkosti podľa stupňa citlivosti.

Ak sú súčiastky klasifikované stupňom 6, znamená to, že sú extrémne citlivé na vlhkosť. Tieto súčiastky sa dodávajú s informáciou na štítku obalu o dobe, počas ktorej môže byť súčiastka mimo obalu, resp. čase a teplote sušenia súčiastky pred jej spájkovaním, aby sa zabezpečilo zbavenie vlhkosti puzdra súčiastky. Všeobecne platí pravidlo okamžitého použitia súčiastok, po vybratí z obalu.



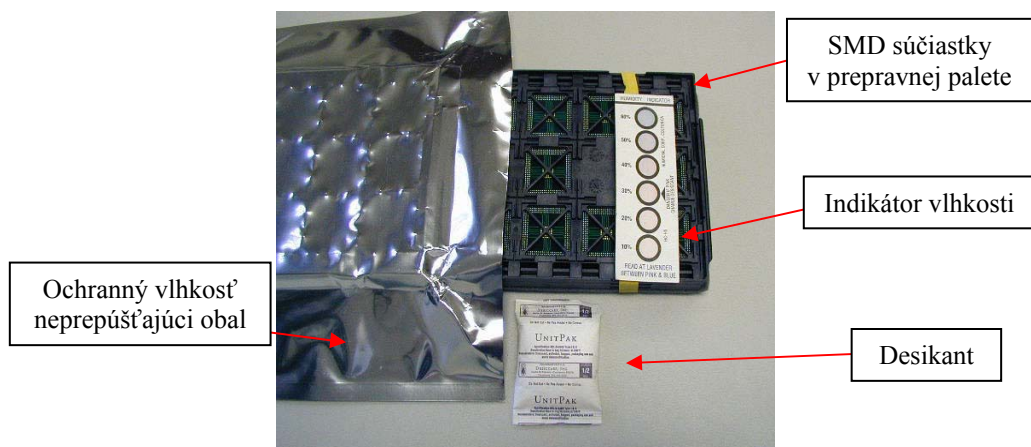
Obr. 2 Sprievodný štítok obalu súčiastok s informáciami o stupni citlivosti na vlhkosť, dobe skladovania a podmienkach skladovania

Na obrázku Obr. 2 je znázornený štítok od výrobcu o výsledku klasifikácie citlivosti na vlhkosť.

III. ODSTRAŇOVANIE VLHKOSTI Z PUZDIER SÚČIASTOK

Elektronický priemysel sa musí stále prispôsobovať prechodu od olovnatej technológie spájkovania k bezolovnatej technológii spájkovania na základe direktívy Európskej únie. Prechod k „zelenej“ technológii výroby elektronických zariadení sa nepriaznivo prejavuje v dodatočných finančných nákladoch na zabezpečenie spoľahlivosti spájkovania bezolovnatou technológiou, ktorá ešte stále nedosahuje úroveň kvality a spoľahlivosti olovnatej technológii. Podstatná zmena je v teplote tavenia bezolovnatej spájky na báze zliatiny Sn-Ag-Cu, ktorá je asi o 30°C vyššia ako je teplota tavenia (183°C) tradičných olovnatých spájk. Zväčšenie teploty tavenia bezolovnatej zliatiny má za následok aj zvýšenie teploty pretavenia spájkovacej pasty, ktorá môže dosahovať až hodnotu 260°C. Vysoké teploty bezolovnatého spájkovacieho procesu sú dôvodom na zvýšenie nárokov nielen na technológiu púzdrenia SMD súčiastok, ale aj na plasty používané pri výrobe pasívnych SMD súčiastok ako sú konektory, cievky a elektrolytické kondenzátory.

Postupy, ktorými je možné predísť zničeniu súčiastok sú opísané v IPC/EDEC štandarde s označením J-STD-033 „*Handling, Packing, Shipping and Use of Moisture/Reflow Sensitive Surface Mount Devices*“. Obsahom dokumentu je poskytnúť SMD výrobcovi a užívateľovi štandardizované postupy a metódy ako manipulovať, baliť, prepravovať a používať SMD súčiastky, ktoré sú podľa J-STD-020 klasifikované ako súčiastky citlivé na vlhkosť. Používaním týchto procedúr je možné predĺžiť dobu skladovania súčiastok až na 12 mesiacov od dátumu ich výroby. Hlavným kritériom pre zvýšenie doby skladovateľnosti a tým aj doby používania súčiastok je správne (tzv. suché) balenie súčiastok citlivých na vlhkosť. Suché balenie pozostáva z materiálu pohlcujúceho vlhkosť (*dessicant*) a karty indikátora vlhkosti, ktoré sú spolu s SMD súčiastkami vložené a uzatvorené do špeciálneho obalu, ktorý neprepúšťa vlhkosť. Ochranný obal taktiež vyhovuje podmienkam na ochranu pred elektrostatickým nábojom. Na obrázku Obr.2 je znázornené správne balenie súčiastok podľa doporučení J-STD-33.



Obr. 3 Balenie súčiastok podľa doporučení štandardu J-STD-033

Jediný spôsob, ktorým je možné zbaviť puzdro súčiastok vlhkosti je ich sušenie. Tento proces musí byť v súlade s odporúčením J-STD-033, pričom je potrebné kontrolovať dobu sušenia, teplotu a vlhkosť prostredia. Treba pripomenúť, že proces sušenia SMD súčiastok prináša so sebou aj riziko oxidácie vývodov, čo môže neskôr spôsobovať problémy pri spájkovacom procese. Ak nie je uvedené iné, potom teplota sušenia by mala byť v rozsahu 90°C až 125°C a čas sušenia by nemal prekročiť 96 hodín. Ak teplota sušenia neprekročí hodnotu 90°C, potom časové ohraničenie nie je. Teploty sušenia vyššie ako 125°C sa neodporúčajú, resp. je potrebné tieto postupy konzultovať s dodávateľom súčiastok. Doba sušenia a podmienky sušenia pre jednotlivé stupne citlivosti sú uvedené v tabuľke Tabuľka 2. Hodnoty platia pre súčiastky, ktoré neboli vystavené väčšej relatívnej vlhkosti ako 60% RH. Doba sušenia súčiastok je závislá od výška puzdra súčiastky.

V Tabuľke 2 nie je uvedený stupeň citlivosti 6, pretože na tento sa vzťahujú špeciálne podmienky sušenia, ktoré musia byť uvedené na obale, v ktorom sa SMD súčiastky dodávajú. Problémom so sušením súčiastok sa môžeme vyhnúť, ak sa rozbalené súčiastky zo „suchého“ obalu kompletne použijú, teda prejdú spájkovacím procesom. Ak sa nepodari všetky súčiastky použiť, je potrebné ich skladovať v suchom prostredí. Problém s vlhkosťou môže nastať aj pri opakovanom spájkovacom procese už spájkovaných súčiastok, ak boli spolu s doskou plošných spojov vystavené nekontrolovateľnému pôsobeniu vlhkosti. V takomto prípade je potrebné pred

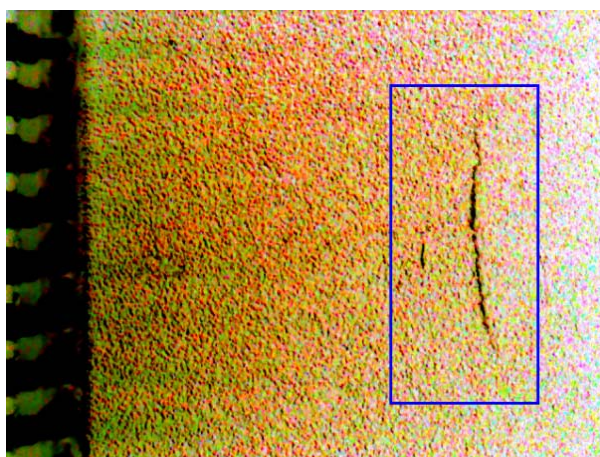
nasledujúcim spájkovacím procesom vysušiť už spájkované súčiastky aj s doskou plošných spojov. Zdroj možných problémov je aj používanie čistiacich kúpeľov po spájkovacom procese, ktoré zvyšujú vlhkosť súčiastok po spájkovacom procese. Treba pripomenúť, že je nežiaduce vystavovať SMD súčiastky viac ako trom spájkovacím procesom.

Tabuľka 2
 PODMIENKY SUŠENIA SMD SÚČIASTOK PODĽA STUPŇA CITLIVOSTI NA VLHKOSŤ

Puzdro SMD súčiastky	Stupeň citlivosti na vlhkosť	Podmienky sušenia súčiastok		
		125°C	90°C ≤ 5% RH	40°C ≤ 5% RH
		Čas sušenia	Čas sušenia	Čas sušenia
Výška puzdra ≤ 1.4mm	2	5 hod.	17 hod.	8 dní
	2a	7 hod.	23 hod.	9 dní
	3	9 hod.	33 hod.	13 dní
	4	11 hod.	37 hod.	15 dní
	5	12 hod.	41 hod.	17 dní
Výška puzdra > 1.4mm ≤ 2.0mm	5a	16 hod.	54 hod.	22 dní
	2	18 hod.	63 hod.	25 dní
	2a	21 hod.	3 dni	29 dní
	3	27 hod.	4 dni	37 dní
	4	34 hod.	5 dni	47 dní
Výška puzdra > 2.0mm ≤ 4.5mm	5	40 hod.	6 dni	57 dní
	5a	48 hod.	8 dní	79 dní
	2	48 hod.	10 dní	79 dní
	2a	48 hod.	10 dní	79 dní
	3	48 hod.	10 dní	79 dní
	4	48 hod.	10 dní	79 dní
	5	48 hod.	10 dní	79 dní
	5a	48 hod.	10 dní	79 dní

IV. EXPERIMENT S ABSORBOVANÍM VLHKOSTI

Experiment poukazuje na dôsledok nesprávnej manipulácie s SMD súčiastkami. Súčiastky boli vystavené extrémnym podmienkam, vlhkosť 100% RH po dobu 48 hodín. Následne boli súčiastky vystavené pôsobeniu teploty počas štandardného teplotného profilu pre bezolovnaté spájky. Na puzdre súčiastka MC68322 v puzdre PQFP160 s rozstupom vývodov 0,5mm bolo možné spozorovať puklinu na spodnej časti puzdra súčiastky, ktorá vznikla v dôsledku veľkého tlaku pár. Poškodenie je znázornené na Obr.1. Na ostatných súčiastkách nebola viditeľná zmena poškodenia puzdra. Nakoľko sa na súčiastkách nerealizoval elektrický (funkčný) test (súčiastky neboli spájkované na dosku plošných spojov), nie je možné tvrdiť, že nedošlo k vnútornému poškodeniu.



Obr. 4 Vznik pukliny v puzdre integrovaného obvodu (MC68322) počas spájkovania v pretavovacej peci pri štandardnom teplotnom profile

V. ODPORUČENIA PRE ZAOBCHÁDZANIE SO SÚČIASTKAMI

Úlohou tohto príspevku bolo poukázať na vplyv vlhkosti na súčiastky pre povrchovú montáž a ich znehodnotenie pri spájkovacom procese. Vplyvy vlhkosti na súčiastky a možnosti eliminovania týchto vplyvov sú spracované v IPC/JEDEC dokumentoch J-STD-020, J-STD-

033. Tento článok si kládol za úlohu priblížiť tieto skutočnosti odbornej verejnosti a vybrať zásadné odporúčania pre manipuláciu s elektronickými súčiastkami.

Na záver je potrebné konštatovať, že súčiastka od svojho vzniku (výroby) môže niekoľkokrát zmeniť svoj obal, až sa dostane ku konečnému užívateľovi. Konečný užívateľ často nepozná cestu súčiastok medzi skladmi veľko-dodávateľov a malo-dodávateľov, dobu skladovania a taktiež podmienky za akých sú súčiastky na regáloch skladov uložené. Existuje pravdepodobnosť, že súčiastka mohla byť na svojej ceste k užívateľovi vystavená pôsobeniu vlhkosti. Z týchto dôvodov je dobré, ak sa preventívne SMD súčiastky pred ich spájkovaním v pretavovacej peci sušia v prostredí 125°C po dobu 8 hodín. Pokiaľ je to možné, treba okamžite použiť rozbalené množstvo. Ten materiál, ktorý sa nepoužil, treba skladovať v priestore s teplotou menšou ako 30°C a relatívna vlhkosť by nemala prekročiť 60%.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku. Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

Tento článok vznikol aj ako súčasť riešeného projektu KEGA 3/7117/09.

LITERATÚRA

- [1] IPC/JEDEC J-STD-20D „Moisture/Reflow Sensitivity Classification for Nonhermetic Solid State Surface Mount Devices“, June 2007
- [2] IPC/JEDEC J-STD-33B „Handling, Packing, Shipping and Use of Moisture/Reflow Sensitive Surface Mount Devices“ October 2005
- [3] T.Y.Lin, Michael G. Pecht, „The Influence of Substrate Enhancement on Moisture Sensitivity Level (MSL) Performance for Green PBGA Packages“, IEEE Transactions on Components and Packaging Technologies, Vol.29, No.3, September 2006
- [4] A publication of the National Electronics Manufacturing Center of Excellence. May 2001. www.aciusa.org
- [5] R. Rowland, „Moisture-sensitive Components“. Surface Mount Technology, October 2000, PennWell Corporation, www.penwell.com
- [6] National Semiconductor „Plastic Package Moisture-Induced Cracking“, MS011822, 2006, www.national.com

Ultrazvukovo kontaktované spoje pre elektrotechniku a ich kvalita

Igor VEHEC, Juraj BANSKÝ, Alena PIETRIKOVÁ

Katedra technológií v elektronike, Fakulta elektrotechniky a informatiky, Technická
univerzita v Košiciach, Slovenská republika

i.vehec@tuke.sk, juraj.bansky@tuke.sk, alena.pietrikova@tuke.sk

Abstrakt — Kontaktované spoje sa v súčasnosti používajú v mnohých aplikáciách kde je kvalita spoja kľúčová. Vytvorenie kvalitného kontaktovaného spoja si vyžaduje zohľadniť množstvo faktorov, ktoré na kvalitu vplyvajú (použitie materiály, spôsob kontaktovania, parametre kontaktovania a pod.). Článok stručne pojednáva o kontaktovaných spojoch a základných faktoroch, ktoré vplyvajú na kontaktovaný spoj, pričom sa stručne venuje aj základným metódam posudzovania kvality kontaktovaných spojov.

Kľúčové slová — kontaktované spoje, kvalita kontaktovaných spojov, optimalizácia procesu kontaktovania, skúška ťahom

I. ÚVOD

Veľké množstvo náročných aplikácií pozostáva okrem iných komponentov aj z mikročipov, ktoré majú prepojenia vytvorené pomocou kontaktovaných spojov. Kontaktovanie spojov je jedným zo spôsobov prepájania polovodičového čipu s vonkajším okolím a predstavuje jeden z kritických krokov pri kompletizácii komponentu. G. Harman uvádza, že v roku 2008 bolo na celej planéte kontaktovaných viac ako 8 až 9 miliárd spojov [1]. Okrem zvyšovania produkcie IO (integrovaných obvodov) a s tým spojeným nárastom kontaktovaných spojov je ďalším trendom zvyšovanie integrácie a s tým spojené znižovanie kontaktovaných plôch a rozstupov medzi nimi. Vytvorenie kvalitných kontaktovaných spojov je podmienkou pre dostatočnú spoľahlivosť elektronických komponentov, obzvlášť v kritických a náročných podmienkach.

II. KONTAKTOVANÉ SPOJE

Kontaktované spoje môžeme v zásade rozdeliť buď z hľadiska fyzikálneho princípu na akom sa kontaktovaný spoj vytvára, alebo z hľadiska tvaru kontaktovaného spoja. Z hľadiska fyzikálneho princípu rozoznávame kontaktovanie termokompresné (T/C), ultrazvukové (U/S) alebo termosonické (T/S) (Tab.1). Z hľadiska tvaru spoja rozoznávame kontaktovanie „na guľôčku“ (ball bonding) alebo kontaktovanie „na hranu“ (wedge bonding) (Obr. 1 a Tab. 2) [1][3].

Najstaršia technológia kontaktovania je T/C kontaktovanie. Vysoké teploty a relatívne vysoký tlak počas kontaktovania však prinášajú značné problémy pri kontaktovaní tenších čipov. Ako vhodnejšie technológie pre kontaktovanie čipových súčiastok bolo preto vyvinuté U/S a T/S kontaktovanie, ktoré ponúkajú viacero výhod [8].

Typická frekvencia používaná pri U/S kontaktovaní sa pohybuje v rozsahu 20 kHz až 300 kHz, pričom staršie zariadenia pre kontaktovanie pracovali s frekvenciou 60 kHz a väčšina moderných zariadení pracuje s frekvenciou 120 až 140 kHz. Prevodník elektrickej energie na mechanickú tvorí väčšinou piezoelektrický element. Amplitúda generovaných ultrazvukových vln je zosilnená v prvku, ktorý zužuje svoj priemer a na konci ktorého je kontaktačný hrot [1][7].

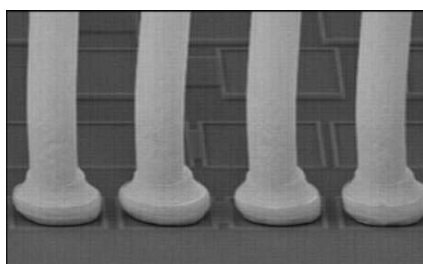
Mechanizmus vzniku spoja je zložitý proces, ktorý nie je stále plne pochopený. Počas kontaktovania dochádza k znáčeniu materiálu drôtu aj kontaktovanej plošky, čo má za následok spolu s aplikovaným tlakom hrotu deformáciu. Na začiatku procesu kontaktovania drôt prechádza cez krehký povrchový film oxidov, odtláča ich nabok čím čistí povrch pričom dochádza k miernemu nárastu teploty v dôsledku trenia. Typická teplota vzniknutá trením neprekročí 10 až 100°C. O niekoľko milisekúnd neskôr sa začínajú formovať malé mikrozvary

na obvode materiálových povrchov a v tomto bode pohyb drôt/plôška prestáva. Vytváranie zvaru pri kontaktovaní teda začína na okrajoch spoja odkiaľ smeruje smerom ku stredu. Po vytvorení zvaru na okraji spoja je ultrazvuková energia absorbovaná celou zváranou plochou (drôt, rozhranie a plôška) pričom ultrazvuková energia tvaruje spoj. Aktivačná energia na formovanie spoja, ale aj vznik prípadných intermetalických zlúčenín, je zrejme prenášaná vo forme interakcie fonón-mriežka.

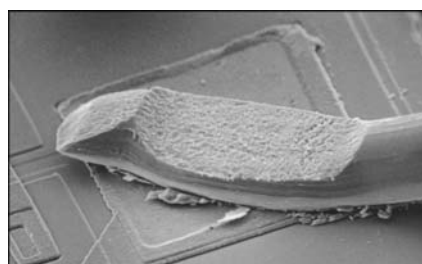
Tabuľka 1
 ROZDELENIE KONTAKTOVANIA Z HEADISKA PRINCÍPU [3][9]

	Typ kontaktovania		
	Termokompresné (T/C)	Ultrazvukové (U/S)	Termosonické (T/S)
Tlak	vysoký	nízky	nízky
Teplota	300 – 500°C	25°C	100 – 150°C
Ultrazvuková energia	nie	áno	áno
Materiál drôtu	Au	Au, Al	Au
Materiál plôšky	Al, Au	Al, Au	Al, Au

Treba poznamenať dve skutočnosti. Prvou skutočnosťou je, že rovnakú deformáciu drôtika môžeme vyvolať aj aplikovaním väčšej sily na drôtik bez ultrazvukovej energie a tepla, výsledkom ale nie je vznik zvaru. Druhou skutočnosťou je, že či už ultrazvuk alebo teplo môžu nezávisle od seba spôsobiť ekvivalentnú deformáciu pri danom namáhaní, jednako existuje významný rozdiel medzi oboma typmi excitácií. Hlavným rozdielom je to, že hustota ultrazvukovej energie potrebná pre deformáciu Al je asi $\sim 10^6$ krát menšia ako pri ekvivalentnej deformácii vyvolanej tepelnou energiou. Pri Al drôtoch bolo pozorované aj celkové stvrdnutie materiálu vplyvom ultrazvukovej energie po jej odobratí [1].



(a)



(b)

Obr. 1 Kontaktované spoje „na guľôčku“ (a) a „na hranu“ (b) [8]

Ako už bolo spomenuté v úvode, kontaktovanie môžeme rozdeliť podľa tvaru kontaktovaného spoja. Kontaktovanie „na hranu“ (Obr.1b) patrí medzi najstaršie technológie kontaktovania (T/C od roku 1957, od roku 1960 U/S a od 1970 aj T/S). Kontaktovanie „na guľôčku“ (Obr. 1a) je rýchlejšia a robustnejšia metóda kontaktovania napriek tomu v prípadoch kde sa vyžadujú prepojenia s nízkym profilom, dominuje kontaktovanie „na hranu“. Kontaktovanie „na hranu“ dominuje aj v prípadoch, v ktorých sa vyžaduje neprerušený steh alebo kontaktovanie čip/čip [5][8].

Tabuľka 2
 ROZDELENIE KONTAKTOVANIA Z HEADISKA TVARU KONTAKTOVANÉHO SPOJA [3][8][9]

	Typ spoja	
	„na guľôčku“	„na hranu“
Typ kontaktovania	T/C, T/S	T/S, U/S
Materiál drôtu	Au	Au, Al
Materiál kontaktovacieho hrotu	karbid volfrámu keramika	karbid volfrámu – pre Al karbid titánu – pre Au
Rýchlosť vytvárania	10/sec (T/S)	4/sec

Základnými parametrami nastaviteľnými pri kontaktovaní sú tlak hrotu na spoj počas vytvárania kontaktu, ultrazvuková energia a čas pôsobenia ultrazvukovej energie. Presné nastavenie parametrov je závislé napr. od použitých materiálov, tvaru hrotu a pod., a výrazne ovplyvňuje kvalitu vytvoreného spoja. Tvar a kvalita kontaktovaného spoja závisí aj od tvaru a parametrov použitého kontaktovacieho hrotu. Samozrejme naopak, typ a tvar kontaktovacieho hrotu je závislý na špecifických požiadavkách daných aplikáciou. Napríklad sklon pod ktorým kontaktovací drôt vchádza do kontaktovacieho hrotu pri kontaktovaní „na hranu“ môže byť rôzny. Menší uhol (30°) je vhodnejší pre kontakty s vyššou slučkou, drôt je ale vystavený väčšiemu namáhaniu ako pri väčších uhloch (60°), ktoré sú vhodné pre vytváranie nižších slučiek. Pritom z hľadiska zachovania tvaru slučky sú vhodnejšie vyššie slučky. Na kvalitu spoja z hľadiska pevnosti v ťahu výrazne vplýva aj dĺžka konca kontaktovacieho hrotu a jeho polomer zaoblenia na prednej a zadnej strane [9].

III. MATERIÁLY POUŽÍVANÉ PRI KONTAKTOVANÍ

Kontaktovaním vytvárame vodivé prepojenie pomocou kontaktovacieho drôtu medzi kontaktovaciu plôškou umiestnenou na čipe a kontaktovaciu plôškou buď na inom čipe alebo na substráte na ktorom je čip umiestnený. V celom procese teda vystupujú materiál kontaktovacieho drôtu a materiál kontaktovanej plôšky. Použité materiály, ale aj kompatibilita medzi nimi, sú ďalšími faktormi výrazne vplyvujúcimi na kvalitu kontaktovaného spoja.

A. Materiály pre kontaktovacie drôty

Najčastejšie používané materiály pre kontaktovacie drôty sú na báze Au, Al a Cu. Kvalita spoja je daná celým radom vlastností materiálu od mechanických až po elektrické. Každý z týchto materiálov má svoje špecifické vlastnosti z ktorých vyplývajú ich výhody a nevýhody z hľadiska použiteľnosti pri vytváraní kontaktovaných spojov.

Au kontaktovacie drôty sú široko používané pre T/C a T/S kontaktovanie. Čisté Au poskytuje primeranú pevnosť v ťahu a vhodné pomerné predĺženie pri namáhaní. Ultračisté Au je ale príliš mäkké a preto sa tu pridávajú prímеси ako sú Be (5 – 10 ppm) alebo Cu (30 – 100 ppm). Ag dotované Be má oproti drôtom dotovaným Cu o 10 – 20% vyššiu pevnosť. Pri kontaktovaní Au drôtov je potrebná zvýšená teplota typicky 150°C [3][8].

Al kontaktovacie drôty sú v svojej čistej podobe príliš mäkké, preto sa používajú najčastejšie v podobe zliatin s prídavkom 1% Si alebo Mg. Al-1%Mg drôty majú oproti Al-1%Si vyššiu odolnosť voči poruche spôsobenej únavou a voči degradácii medze pevnosti po vystavení zvýšenej teplote. Pri kontaktovaní Al drôtov nie je potrebný dodatočný ohrev a postačuje teplota okolia [3][8].

Cu kontaktovacie drôty sa javia ako vhodná náhrada Au drôtov a to hlavne z ekonomických dôvodov. Z hľadiska elektrických vlastností ponúka Cu významne lepšiu konduktivitu a vyšší tavný prúd. Nevýhodou je ale ľahšia oxidácia Cu a jeho horšia kontaktovateľnosť, nakoľko na zmäkčenie Cu je potrebný väčší výkon, ktorý spôsobuje aj nadmerné zmäknutie kontaktovanej plôšky. Tvrdšia Cu môže navyše spôsobiť vytváranie „kráterov“ [3][6].

Hrúbka drôtov sa pohybuje od $\sim 10\ \mu\text{m}$ do rádovo $100\ \mu\text{m}$ a je závislá od požiadaviek na príslušnú aplikáciu. Pre jemné aplikácie sú vhodné tenšie drôty. Hrubšie kontaktovacie drôty sú vhodné hlavne z hľadiska ich vyššej pevnosti a lepšej odolnosti voči deformácii, ktorá môže nastať pri púzdrení a lisovaní čipov [9]. Vhodné sú aj pre aplikácie v ktorých sa vyžadujú vyššie prúdové zaťaženia. Pri vyšších prúdových zaťaženiach sa rovnako využívajú kontaktovacie drôty vo forme pásika – tzv. ribbon bonding – ktoré sa kontaktujú spôsobom „na hranu“.

B. Kompatibilita kontaktovacieho drôtika s materiálom kontaktovanej plôšky

Typické materiály pre kontaktovacie plôšky sú Au, Al, Cu, Ag a Ni. Kontakty vytvorené spájaním rovnakých materiálov (rovnaký materiál kontaktovacieho drôtu a kontaktovacej plôšky) sú najmenej náchylné na vznik intermetalických zlúčenín a z tohto dôvodu patria medzi vysoko spoľahlivé. V ďalšom preto stručne rozoberieme len kontaktovanie rozdielnych materiálov.

Kontaktovanie Au/Al patrí medzi najbežnejšiu kombináciu. Je náchylné na tvorbu rôznych intermetalických zlúčenín (Au_5Al_2 , Au_4Al , Au_2Al , AuAl , AuAl_2) a Kirkendallových pórov, ktorých formovanie môže byť ovplyvnené aj zvýšenou prevádzkovou teplotou. Intermetalická zlúčenina AuAl_2 sa môže na rozhraní styku Au/Al formovať dokonca aj počas kontaktovania pri izbovej teplote.

Kontaktovanie Au/Cu sa vyznačuje tvorbou intermetalických zlúčenín (Cu_3Au , AuCu , Au_3Cu), ktoré môžu pri zvýšených teplotách (200 - 300°C) znížiť pevnosť. Degradácia spoja závisí od nečistôt.

Kontaktovanie Au/Ag sa vyznačuje veľkou spoľahlivosťou po dlhú dobu aj pri zvýšených teplotách, nakoľko nie sú náchylné na vznik intermetalických zlúčenín. Problémy s kontaktovateľnosťou môžu byť spôsobené prítomnosťou síry.

Kontaktovanie Al/Ag je široko využívané hlavne v hrubovrstvovej technológii pri výrobe hybridných IO, kde sa využívajú hrubovrstvové pasty na báze AgPt alebo AgPd. Na rozhraní Al/Ag vzniká veľké množstvo intermetalických zlúčenín, Kirkendallove póry vznikajú zvyčajne až pri vyšších teplotách ako sú typické prevádzkové teploty. Spoje majú tendenciu degradovať a vo vlhkom prostredí oxidovať.

Kontaktovanie Al/Ni je typicky používané pri výrobe výkonových súčiastok a pre vysoko-teplotné aplikácie. Kontakty sú spoľahlivejšie ako kontakty Al/Ag a Al/Au v rôznych prevádzkových prostrediach.

Kontaktovanie Cu/Al sa vyznačuje tvorbou intermetalických zlúčenín, ktoré vedú k podobným poruchám ako pri systéme Au/Al. Nárast intermetalických zlúčenín v Cu/Al je ale pomalší ako pri Au/Al a nevedie k vzniku Kirkendallových pórov. Rýchlosť formovania intermetalických zlúčenín závisí okrem teploty aj na zložení atmosféry [3].

IV. POSUDZOVANIE KVALITY KONTAKTOVANÝCH SPOJOV

Ako bolo uvedené, optimalizácia procesu kontaktovania si vyžaduje posúdiť celý rad faktorov, ktoré vplyvajú na kvalitu kontaktovaného spoja (čip – puzdro – kontaktovací drôt – kontaktovací hrot – nastavenie parametrov kontaktovania a pod.). Pre posudzovanie kvality kontaktovaných spojov bolo navrhnutých niekoľko metód, ktoré sú buď deštruktívne alebo nedeštruktívne. Štandardnými deštruktívnymi metódami, ktoré sledujú mechanické vlastnosti spojov, sú skúšky ťahom (patrí medzi široko akceptované metódy testovania) a strihom. Medzi nedeštruktívne metódy patrí hlavne vizuálna inšpekcia, nedeštruktívne skúšky ťahom (NDPT – nondestructive pull test) a strihom (NDBS – nondestructive ball shear) [4][7]. Prehľad štandardov a noriem používaných pri posudzovaní kvality kontaktovaných spojov je zhrnutý v Tab. 3 [3][10].

Tabuľka 3
 PREHĽAD ŠTANDARDOV A NORIEM PRE POSUDZOVANIE KVALITY KONTAKTOVANÝCH SPOJOV [3][8][9]

Štandard	Popis
EIA*/JESD22-B116	Wire Bond Shear Test
ASTM** F 458	Standard Practice for Nondestructive Pull Testing of Wire Bonds
ASTM F 459	Standard Test Methods for Measuring Pull Strength of Microelectronic Wire Bonds
IPC-TM-650	2.4.42.3 - Wire Bond Pull Strength
MIL-STD-883H	Method 2010.12 - Internal Visual (Monolithic) Method 2011.8 - Bond Strength (Destructive Bond Pull Test) Method 2017.9 - Internal Visual (Hybrid) Method 2023.6 - Nondestructive Bond Pull Method 3003.1 - Delay Measurements Method 2001.3 - Constant Acceleration Method 2026 - Random Vibration Method 2002.5 - Mechanical Shock Method 1008.2 - Stabilization Bake Method 1004.7 - Moisture Resistance Method 5009.1 - Destructive Physical Analysis

* Electronics Industries Alliance

** American Society For Testing And Materials

Jedným z vhodných testov na posudzovanie kvality kontaktovaných spojov je skúška ťahom, ktorá čiastočne koreluje s vizuálnou kontrolou, ako aj so základnými nastaviteľnými parametrami kontaktovacieho procesu (výkon, čas a sila). Vyhodnotením skúšok ťahom je teda možné optimalizovať uvedené parametre.

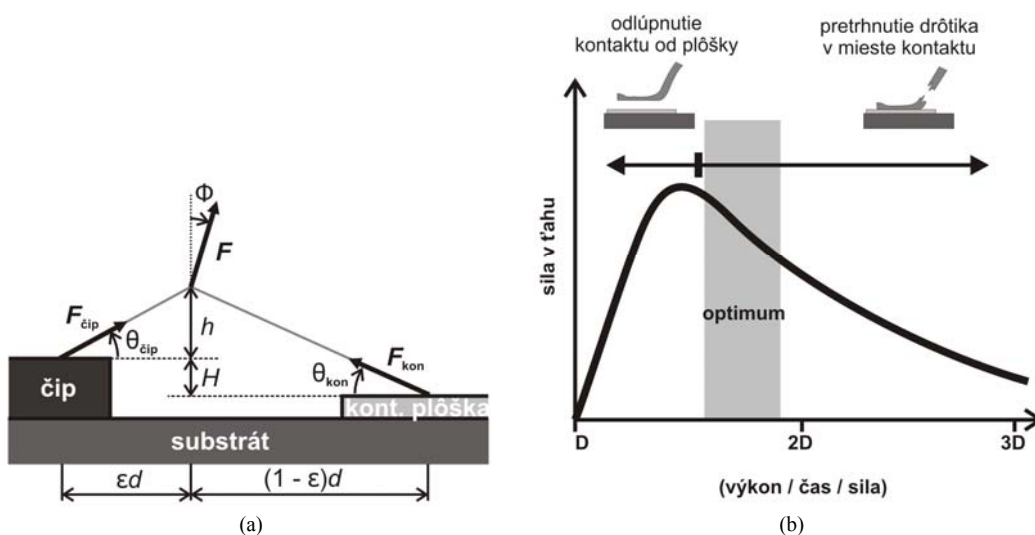
Základné usporiadanie skúšky ťahom je znázornené na Obr. 2a. Počas skúšky pôsobí ťahová sila na kontaktovací drôt, pričom sa rozkladá smerom k obojm kontaktom. Vo všeobecnom prípade je veľkosť sily pôsobiacej na kontakt na strane čipu pri usporiadaní podľa Obr. 2a daná vzťahom [1][11]:

$$F_{\check{c}ip} = F \left[\frac{\left(h^2 + \varepsilon^2 d^2 \right)^{1/2} \left((1-\varepsilon) \cos \Phi + \frac{(h+H)}{d} \sin \Phi \right)}{h + \varepsilon H} \right] = F \frac{\cos(\theta_{kon} - \Phi)}{\sin(\theta_{\check{c}ip} + \theta_{kon})} \quad (1)$$

a pre veľkosť sily pôsobiacej na kontakt na strane kontaktovacej plošky je daná vzťahom [1] [11]:

$$F_{kon} = F \left[\frac{\left(1 + \frac{(1-\varepsilon)^2 d^2}{(H+h)^2} \right)^{1/2} (h+H) \left(\varepsilon \cos \Phi - \frac{h}{d} \sin \Phi \right)}{h + \varepsilon H} \right] = F \frac{\cos(\theta_{\check{c}ip} + \Phi)}{\sin(\theta_{\check{c}ip} + \theta_{kon})} \quad (2)$$

Vzťahy nadobúdajú samozrejme jednoduchší charakter v prípade ak sú kontakty v rovnakej výške ($H = 0$ m), ťahová sila pôsobí kolmo nahor ($\Phi = 0^\circ$) a uhly $\theta_{kon} = \theta_{\check{c}ip}$. Za uvedených podmienok bude vždy platiť $F_{\check{c}ip} = F_{kon}$.



Obr. 2 Rozloženie síl pri skúške ťahom (a) a závislosť maximálnej ťahovej sily od deformácie spoja počas kontaktovania (b) [1][3][11]

Určenie optimálnych parametrov kontaktovacieho procesu sa realizuje na základe závislosti maximálnej ťahovej sily od výkonu dodávaného do spoja prostredníctvom ultrazvukovej energie, času pôsobenia ultrazvuku, alebo sily pôsobiacej na spoj počas kontaktovania. Všetky druhy závislosti vykazujú podobný charakter a korelujú so závislosťou maximálnej ťahovej sily od deformácie spoja vzniknutej počas kontaktovania (Obr. 2b). V grafe na Obr. 2b je deformácia spoja vyjadrená ako násobok priemeru kontaktovacieho drôtu D . Pri zvyšovaní ktoréhokoľvek parametra kontaktovania je spoj viac deformovaný a pevnejší. Na začiatku krivky prevláda pri skúške ťahom odlupovanie kontaktu od kontaktovacej plošky, ktoré po dosiahnutí maxima pevnosti v ťahu prechádza na porušenie spoja pretrhnutím drôтика v mieste kontaktu. Optimálne parametre kontaktovania sa nachádzajú v úzkom intervale za maximom krivky [3].

V. ZÁVER

Posudzovanie kvality kontaktovaných spojov predstavuje komplexnú problematiku, ktorá si vyžaduje široké znalosti od princípov kontaktovania, mechanických a elektrických vlastností materiálov a spojov až po znalosť relevantných noriem a štandardov. Pre vytvorenie kvalitného kontaktu je potrebné brať do úvahy nielen rozvrhnutie elektronického obvodu a ďalšie technologické operácie, ktoré bude ešte potrebné s už nakontaktovanými čipmi realizovať, ale aj pracovné podmienky pre ktoré je daná aplikácia vytváraná.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku. Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

Tento článok vznikol aj za pomoci riešeného projektu KEGA 3/7117/09.

LITERATÚRA

- [1] G. Harman, "Wire Bonding in Microelectronics", 3rd Edition, New York: McGraw-Hill, 2010, ISBN: 978-0-07-164265-1.
- [2] *Bonding Wedge Catalog* (Book style). Kulicke & Soffa, 2001.
- [3] Z. Lai and J. Liu, *Chapter A: Wire Bonding*. In: The Nordic Electronics Packaging Guideline, 2000. <http://extra.ivf.se/ngl/>
- [4] C. Wang and R. Sun, *The Quality Test of Wire Bonding*. In: Modern Applied Science, Vol. 3, No. 12, Canadian Center of Science and Education 2009, ISSN 1913-1852.
- [5] D. J. Beck and A. C. Perez, *Wire Bond Technology - The Great Debate: Ball vs. Wedge*, In: Advanced Packaging, March, 2007.
- [6] Ch. Breach, *The Great Debate: Copper vs. Gold Ball Bonding*, In: Advanced Packaging, October, 2008.
- [7] H.K. Charles, *Advanced Wire Bonding Technology: Materials, Methods, and Testing*, In: Materials for Advanced Packaging, Springer Science+Business Media, LLC 2009, ISBN 978-0-387-78219-5.
- [8] Gaiser precision bonding tools, CoorsTek, Inc. <http://www.gaisertool.com/>
- [9] Bonding Wedge Catalog, Kulicke & Soffa 2001. <http://www.kns.com/>
- [10] MIL-STD-883H - Test Method Standard Microcircuits.
- [11] Wire Bonding Quality Assurance and Testing Methods, http://ssd-rd.web.cern.ch/ssd-rd/bond/talks/2-04_Schneider-Ramelow.pdf

Animated Components and Their Utilization in the Education of the Electronic Materials

Slavomír KARDOS, Igor VEHEC, Juraj BANSKÝ

Department of Technologies in Electronics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

slavomir.kardos@tuke.sk, i.vehec@tuke.sk, juraj.bansky@tuke.sk

Abstract — The contribution deals with possibilities and practical utilization of animated components in educational process with the scope of materials for electronic industry. Computer animation has great potential to serve in understanding of chemical structure and physical behavior of materials. Complex nature of this subject calls for interactive elements. It can depict the things, what is able to imagine only by mind in the case of regular books. In the next, the contribution deals with multimedia support of education and with its status at the Department of technologies in electronics, Technical university of Košice. Presenting of information by multimedia form is more efficient as classic form of education, whereas enchain the perception of recipient by faster way. Interactive intermediation of information is the characteristic feature. Compared to classic forms of education it offers wider possibilities for explanation and understanding of study materials. Feeding of information in multiple formats simultaneously is the advantage for the recipients with sight and hearing handicap. With utilization of multimedia components were created a number of study modules for electronic education courses at the Department of technologies in electronics. The custom internet server and specialized software were acquainted for these purposes.

Keywords — animated components, education, electronic materials

I. INTRODUCTION

The multimedia types of information are minimally the two component combination of text, sound, graphics, animation and video item, what simultaneously effect on multiple senses. They are mainly intended for education, information and amusement. From the hardware side, the multimedia can be defined as electronic media for holding and interpretation of multimedia contents. One of the basic feature is their interactivity, i.e. the user controls behavior of multimedia application. Web presentations, intelligent information boards, electronic encyclopedias, catalogues, computer games and others belong into the multimedia radius. Television, music, dance, song and projections are not multimedia even though they can effect on multiple senses. We meet with multimedia show or various presentations in masmedia, galleries and podiums.

We distinguish two basic types of multimedia components – the first ones are time independent (text, picture), the second ones are dynamic (sound, video, animations). The information in graphic form is synoptical and interesting and increases psychological effect of the information (for example, jokeful pictures can positively tune the percipient). Sound has multiple functions – positive in tuning, under voice along with pictures or texts, sound effects, messages for user, what notices about changes in action, etc.

Animation (the word is derived from Latin term anima – spirit, revival) introduces in motion something motionless, whereby creates the illusion of motion. Computer animations we can dissect into animations of two-dimensional objects (of planar models) and of three-dimensional objects animations (of 3D stereo metric models). Video sequences can multiple the synoptictness and resultant sense from multimedia application.

ISO (International Standard Organization) defines the multimedia as creation, modification and/or assembly of objects, what consist of whatever combination of media contents. The medium is the resource, by which the information is sensed, expressed or transmitted. So the multimedia we understand as a integration of text, graphics, sound, image sequences (animations) and videosequences for the purpose of information intermediation with the

possibility interactively participate at multimedia application behaviour by the participant [1]. A number of technical and software facilities, i.e. multimedia system, typical by simultaneous processing and interactivity of multimedia components, are needed for creation and use of multimedia application.

E-learning or CBT (Computer-Based Training) or IBT (Internet-Based Training), whether WBT (Web-Based Training) represents the form of multimedia support of education, what is specialized for individual education of participants, while the contact with tutor or consultant is mostly maintained by electronic form. Information and communication resources can have on-line or off-line form. So e-learning allows by the medium of personal computer, memory medias and internet the transfer of knowledge from the creators of study materials to their percipients as well as allows the feedback in the form of messages, discussions, self-evaluation and exam tests [2].

Such style of study needs the usage of certain platform, what includes the mentioned CBT, WBT, virtual classroom and various forms of electronic communication. Such communication is mediated by internet or intranet. The forms of multimedia support can be various audio/visual media, satellite television, etc., what contain study texts, pictures, photos, animations, sound and video records. Virtual classroom allows the evidence and management of on-line education, on-line conferences, etc., in which the direct contact of participants by the medium of video and audio is used. The respective software, what allows synchronous audiovisual data transfer is necessary for virtual classroom operation. Virtual classroom brings the possibilities to join the students and pedagogues every time and everywhere in the world by the medium of fully interactive audio/videoconference tools hosted by educational institutions. Such solution essentially allows schools and universities to open their academic premises to whole world [3].

In the interest of educational institutions is to increase the quality, attractivity and flexibility of academic programs and to motivate the high-quality pedagogues and students. In the fight with concurrency there are needs for high standard of study programs, flexible timetable and original supply of educational and working opportunities.

During the relative short time of experience with electronic form of education, there were elaborated various studies [4]. According these studies, we can achieve better results of education process as in classic attendance form of education, when we use suitable facilities and form. Clearly, this form of education has its valid position and in some events it is the only real way, e.g. when the student is limited by employment or geographic position.

II. DEVELOPMENT AND UTILIZATION OF MULTIMEDIA COMPONENTS IN STUDY PRACTICE

In the frame of e-learning support of education were created a number of study modules at The Department of Technologies in Electronics with the support of a few projects. In the last year, there was provided and installed own server for these purposes. Current endeavor of department's workers was expanded by following activities:

- creation of new study materials and actualization of existing study materials for electronic support of education,
- implementation and configuration of open-source e-learning application Moodle according to the study materials and course requirements,
- creation of animated components under the Adobe Flash CS4 platform and their implementation to the study materials.

A variety of software products were provided for creation of electronic study materials. For technical support of creation and presentation of multimedia contents, there was also provided further equipment as the camcorder, personal workstations, wide screen television set at mobile pedestal, electronically controlled canvas, wiping tables, etc. The modern workstation tables with models for handicapped students are planned.

Several study modules were put into practice under the Moodle platform for the application in the bachelor and master of science study subjects (Fig. 1).

Creation and actualization of contents is in progress with utilization of automated web-design studio. Illustrations are created mostly by vector software – Corel X4, whose objects are also used as base components for animations under Adobe Flash CS4 [6]. The last mentioned software was selected behalf various reasons. In particular it provides dynamic aspect, what allows animation of objects as a presentation of real behaviors and processes. In the following it is the possibility of increased interactivity with the user, what is scarcely possible to achieve with utilization of classic scripting tools. In the Fig. 2 is static view to one of created Flash CS4 animations, the diamante crystallography structure.



Fig. 1 Sight into menu of educational modules in the open-source e-learning environment of Moodle system [5]



Fig. 2 View to the Adobe Flash CS4 animated component embedded in e-learning module

III. CONCLUSION

Aim of the contribution was to shortly resume the possibilities of multimedia components utilization in the education of electronics materials against background of pedagogic activities at the Department of technologies in electronics. The problematic is described in close context with area of electronic education and creation of electronic study materials for these purposes. After previous technical support, created study materials are designed as individual modules for use in courses oriented at study of electronic technologies.

ACKNOWLEDGMENT

The contribution was elaborated in the frame of the KEGA project No.: 3/7117/09.

REFERENCES

- [1] Vladimír Dráb, Multimédiá, 2004, <http://www.infovek.sk/predmety/inform/materialy/namety/multimedia/html/mm1.html>
- [2] e-Learning portál, Technická univerzita v Košiciach, <http://ep.tuke.sk/home/index.php?Lang=sk&Theme=Default&Mode=Portal&Portal=1>.
- [3] Cisco virtuálna učebňa pre stredné a vysoké školy, 2006, <http://servis.inet.sk/clanok/3757-cisco-virtualna-ucebna-pre-stredne-a-vysoke-skoly>
- [4] Ready to Teach Grant Program, United States Department of Education, <http://www2.ed.gov/programs/readyteach/2005awards.html>
- [5] Moodle Open Source Course Management System, <http://moodle.org/about/>
- [6] Adobe Flash Professional, <http://www.adobe.com/products/flash/whatisflash/>

Melting and Solidification of Solder Alloys

Alena PIETRIKOVÁ, Juraj ĎURIŠIN

Dept. of Technologies in Electronics, Faculty of Electrical Engineering and
Informatics, Technical University of Košice, Slovak Republic

alena.pietrikova@tuke.sk

Abstract — In-situ X-ray diffraction experiments using synchrotron radiation were employed to analyze microstructure evolution of 95.5Sn3.8Ag0.7Cu (wt. %), 99Sn1Cu (wt. %) and 96Sn4Ag (wt. %) lead-free solder alloys during heating (30-250°C), isothermal annealing (240°C) and cooling (250-30°C). The special emphasis was placed at the study of melting and solidification process, explaining formation, distribution and the order of crystallization of the crystal phases (β -Sn, intermetallic compounds) in the solder alloys.

Keywords — X-ray diffraction, microstructure, lead-free solders, crystal phase

I. INTRODUCTION

Lead-free solders based on SnAgCu, SnAg and SnCu alloys (with melting temperature under 250°C) are extensively used as low temperature solder materials in the electrotechnical industry. Therefore solder joints play important role from view of quality and reliability of electronic devices. Significant factor defining properties of solder joints is their microstructure. Microstructure of solder joints of final electronic devices is generally influenced by applied solder alloy, process of soldering, composition of soldered materials and utility of the final device. All four factors have significant impact on phase transformations in solder joints consequently influencing their properties. Microstructure of lead-free solders consists of large β -Sn dendritic crystals with typical length of 10^1 - 10^2 μm . Another very important component of the solders are intermetallic compounds. Solders based on Sn with additions like Ag or Cu contain dominantly fine Ag_3Sn precipitates (≈ 1 - 10^1 μm) or Cu_6Sn_5 precipitates (≈ 1 μm). Distribution of intermetallic compounds in solder alloys is based on weight ratio of chemical elements in the solders [1].

Realized in-situ measurements of 95.5Sn3.8Ag0.7Cu, 99Sn1Cu and 96Sn4Ag solder alloys illustrate behavior of the alloys during temperature increase (30-250°C) and during temperature decrease (250-30°C), thus simulating the soldering process. The special emphasis was placed at the observation of the phase evolution mainly during the solidification process.

II. MEASUREMENT

As a source of X-ray radiation was used for the diffraction measurements B2 bending magnet beamline of the DORIS III positron storage ring at HASYLAB/DESY (Hamburg, Germany). The beamline allows performing X-ray diffraction measurements in Debye-Scherrer geometry (Fig. 1). For our measurements the wavelength was set to $\lambda = 0.5384$ Å. The use of synchrotron radiation has numerous advantages in comparison with conventional laboratory X-ray lamps since it offers high brilliance photon beams covering relatively wide energy spectrum. To speed up the measuring, diffracted X-rays were detected on position-sensitive image plate detector OBI, covering the angle range $2\theta_{\text{max}} = 110^\circ$.

The samples of the alloys in the form of powder were placed into glass capillaries having the inner diameter of 0.28 mm. The capillaries containing samples were consequently (for single alloys) mounted into the STOE furnace, thus simulating the soldering process (temperature increase and decrease). The temperature in the sample vicinity was measured by thermocouple. To eliminate the texture effect and to ensure correct intensities of the Bragg peaks, the capillary was rotating with frequency of 2 Hz (ω) around its horizontal axis, perpendicular to the incoming monochromatic beam (Fig. 1). The sample was illuminated with the monochromatic

beam having the width and height of 12 and 5 mm, respectively. Since the intensity of the incoming beam decays due to time decay of the positron current in the storage ring, exposure was defined by achieving the preset number of photons detected by an incoming beam monitor. Such procedure yields data normalized to the incoming beam intensity and thus simplifies further analysis. Measurement at every temperature took between 12-14 minutes. The temperature step of 10 °C was used during heating (30-210°C) and cooling parts (190-30°C), whereas the 2°C step was applied for more detailed inspection of the melting and the solidification processes occurring in the temperature ranges (210-250°C) and (250-190°C), respectively.

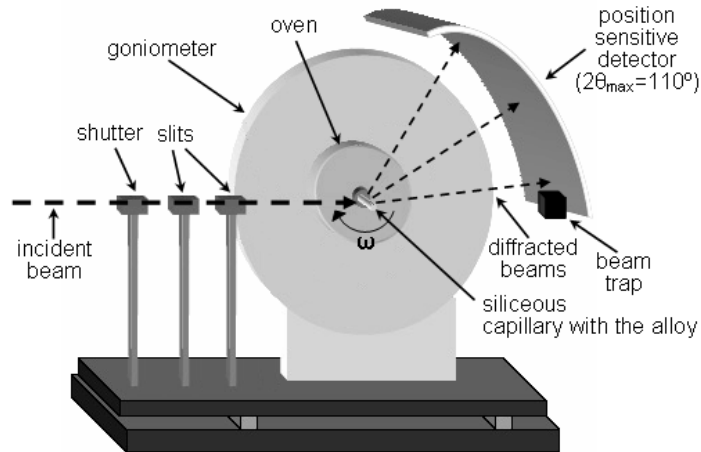


Fig. 1 Configuration of diffractometer in Debye-Scherrer geometry

III. RESULTS AND DISCUSSION

X-ray diffraction patterns of the fresh samples taken at 30°C revealed the presence of the sharp and intense Bragg peaks stemming from β -Sn phase (PDF Nr. 4-673). Furthermore slight traces of Ag_3Sn crystal phase (PDF Nr. 44-1300) are clearly visible for 95.5Sn3.8Ag0.7Cu and 96Sn4Ag alloys and also slight traces of Cu_6Sn_5 crystal phase (PDF Nr. 45-1488) for 95.5Sn3.8Ag0.7Cu a 99Sn1Cu alloys [3]. Because our measurements were mainly aimed at evaluation the phase changes, next diffraction patterns depict behavior of the analyzed solder alloys only during the melting process (temperature increase) and at selected temperatures during the temperature decrease, where crystallization of crystal phases takes its course.

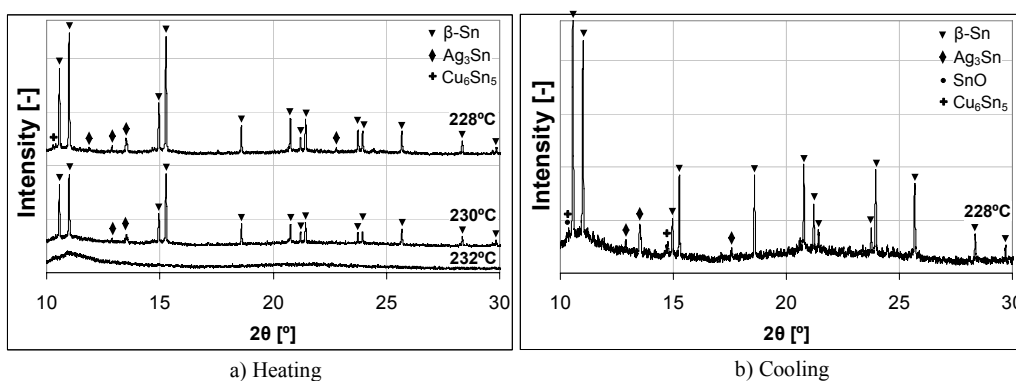


Fig. 2 Diffraction patterns of the 95.5Sn3.8Ag0.7Cu alloy at selected temperatures

Heating the alloys from 30°C up to melting temperatures of single alloys reveals no phase changes (phases formations). After reaching the temperature of 228°C (95.5Sn3.8Ag0.7Cu), 238°C (99Sn1Cu) and 232°C (96Sn4Ag) diffraction peaks belonging to major constituent phase β -Sn are rapidly diminishing, what is a sign of reaching the melting point of the solders. After increasing the temperature of 2°C it was possible to observe next depression of diffraction peaks intensity together with significant relative rising of the background. X-ray diffraction patterns obtained at 232°C (95.Sn3.Ag0.Cu), 244°C (99Sn1Cu) and 236°C (96Sn4Ag) exhibit no Bragg peaks at all and thus confirm completely molten state of the alloys (Fig. 2a, Fig. 3a, Fig. 4a).

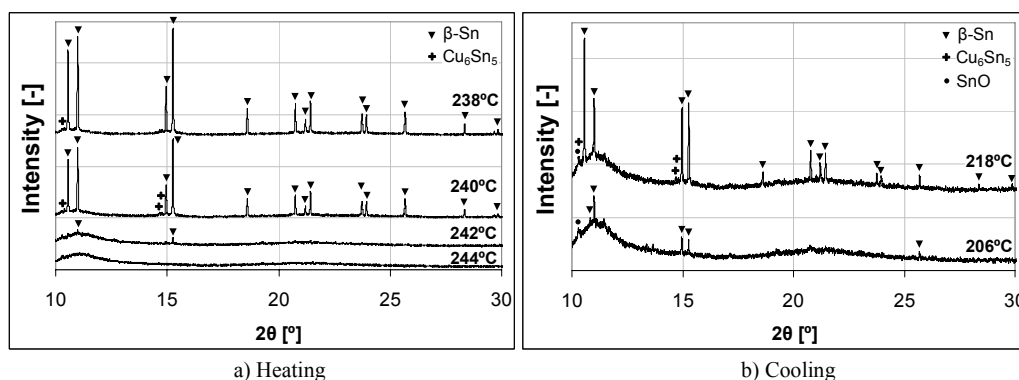


Fig. 3 Diffraction patterns of the 99Sn1Cu alloy at selected temperatures

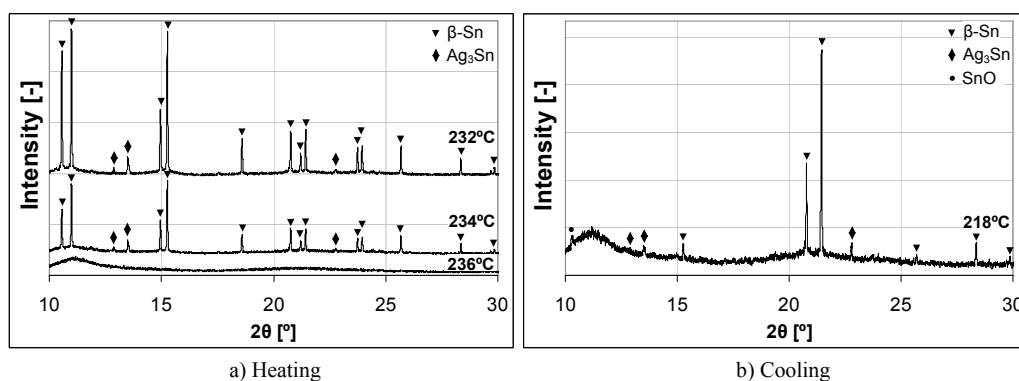


Fig. 4 Diffraction patterns of the 96Sn4Ag alloy at selected temperatures

It should be noted here that according to values published in database [2] and in binary diagrams [4], the liquidus temperatures of the alloys were measured within the tolerance of $+(8-12)^{\circ}\text{C}$. These differences are probably consequence of temperature gradient inside the furnace combined with the temperature step (2°C).

Further isothermal annealing at 250°C for two hours revealed no phase changes and diffuse character of the X-ray diffraction (XRD) patterns remained unaffected.

In the next part we focused on the solidification process, or more precisely on crystallization of the solder alloys. XRD patterns reveal diffuse character specific for amorphous materials till temperature of 228°C ($95.5\text{Sn}3.8\text{Ag}0.7\text{Cu}$), 218°C ($99\text{Sn}1\text{Cu}$) and 218°C ($96\text{Sn}4\text{Ag}$) - Fig. 2b, Fig. 3b, Fig. 4b, when the first Bragg peaks corresponding to the major $\beta\text{-Sn}$ phase (or also to intermetallic compounds) appear. Similar diffraction patterns (if only $\beta\text{-Sn}$ reflections are taken into consideration) were recorded during temperature increase (melting process) at the temperature of 230°C ($95.5\text{Sn}3.8\text{Ag}0.7\text{Cu}$), 240°C ($99\text{Sn}1\text{Cu}$) and 234°C ($96\text{Sn}4\text{Ag}$). One may conclude here that the temperature difference between the melting liquidus and the solidification liquidus of the single alloys indicates kind of thermal hysteresis.

Next table (Tab. 1) summarizes data gained from the measurements, or more precisely from XRD patterns: liquidus temperatures (for temperature increase and decrease) and also temperatures of formation of crystal phases during cooling.

Table 1
 TEMPERATURES OF LIQUIDUS AND CRYSTALLIZATION OF CRYSTAL PHASES

	Temperature increase	Temperature decrease			
	liquidus	liquidus	$\beta\text{-Sn}$	Ag_3Sn	Cu_6Sn_5
$95.5\text{Sn}3.8\text{Ag}0.7\text{Cu}$	232°C	228°C	228°C	228°C	228°C
$99\text{Sn}1\text{Cu}$	244°C	218°C	218°C	-	206°C
$96\text{Sn}4\text{Ag}$	236°C	218°C	218°C	218°C	-

IV. CONCLUSION

Diffraction measurements allowed detail analysis of the melting and solidification process of the solder alloys. Significant result of the measurements is definition of liquidus of the solders (for temperature increase and decrease) and also definition of crystallization order of crystal phases in the solder alloys, which explains distribution of the phases in lead-free solders.

ACKNOWLEDGEMENT



This paper was developed with support of the project „Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy“, ITMS 26220220029, that is co-financed from Structural Funds EU ERDF within Operational programme Research and Development V-2008/2.2/01-SORO and preferred axis 2 Support of Research and Development.

REFERENCES

- [1] B. A. Cook, I. E. Anderson, J. L. Harringa, S. K. Kang: Isothermal Aging of Near-Eutectic Sn-Ag-Cu Solder Alloys and Its Effect on Electrical Resistivity, *Journal of Electronic Materials*, Vol. 32, No. 12, 2003.
- [2] Cooperation on Science and Technology – European Lead-free Soldering Network: Database for Properties of Lead-free Solder Alloys, February 2006.
- [3] The International Centre for Diffraction Data®: PDF-2, 1997.
- [4] ASM International's Binary Alloy Phase Diagrams, ASM International, Materials Park.

Implementácia kamerového systému do výučby na Katedre technológií v elektronike

Dominik DEMETER, Juraj BANSKÝ, Ľubomír LIVOVSÝ

Katedra technológií v elektronike, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

dominik.demeter@tuke.sk, juraj.bansky@tuke.sk, lubomir.livovsky@tuke.sk

Abstrakt — Táto publikácia popisuje implementáciu kamerového systému ako súčasť e-Learningového systému pre podporu výučby na Katedre technológií v elektronike. V článku sú rozoberané očakávané prínosy kamerového systému, samotných audio a videozáznamov pre zefektívnenie výučby ako aj predpokladané technické problémy pri implementácii celého systému.

Kľúčové slová — e-Learning, kamerový systém, video prenos, virtuálne laboratórium

I. ÚVOD

Súčasný rýchly vývoj informačných technológií zasahuje každú oblasť našej spoločnosti. Výnimkou nie je ani oblasť vzdelávania. Čoraz viac škôl resp. pracovísk pociťuje potrebu nasadenia virtuálneho laboratória resp. e-Learningového systému.

Samotný výraz e-Learning je možné aplikovať na rôzne formy podpory výučby pomocou informačných technológií. Môžu to byť rôzne dokumenty a skriptá v elektronickej forme, prezentácie, multimediálne nahrávky, interaktívne aplikácie pre podporu výučby, animácie, elektronické testy, dokonca aj on-line prístup ku zariadeniam v laboratóriách s možnosťou riadenia z diaľky.

Na Katedre technológií v elektronike bola donedávna podpora výučby pomocou e-Learningových nástrojov pomerne nízka. V súčasnosti sú postupne vypracované rôzne elektronické prezentácie, ktoré sú autorizovaným študentom sprístupnené pre potreby štúdia. Pri spracovaní materiálov je kladený dôraz na interaktivitu študijných materiálov. Cieľom je, aby študent musel študovať materiály a aby sa mohol postupne preklikať ku všetkým informáciám, zároveň však zabezpečiť tieto materiály pred nepovoleným kopírovaním, ako aj proti zneužitiu. Tieto kritériá najviac spĺňajú flash animácie resp. rôzne prezentácie, ktoré neumožňujú zobrazit' celý svoj obsah naraz a nemajú podporu pre vytlačenie obsahu. Do tohto konceptu zapadá aj využitie on-line prenosov z prednášok, z laboratórnych cvičení ako aj vytvorenie audio a video-archívu, prípadne zverejňovanie prezentácií vo forme videa.

II. NÁVRH SYSTÉMU

A. Súčasný stav

Katedra technológií v elektronike má stiesnené priestorové možnosti. Štyri laboratória z ôsmich majú rozlohu do 20m², pričom ani jedno z nich nie je širšie ako 3m. Na tak malej ploche sú umiestnené minimálne 3, v jednom laboratóriu dokonca až 6 technologických zariadení. Pri týchto rozmeroch nie je možné zabezpečiť, aby každý študent videl na ukážkových cvičeniach každý krok práce s jednotlivými zariadeniami, následkom čoho je nepripravenosť študentov na samostatnú prácu so zariadeniami. Dôsledkom daného stavu je, že študenti síce majú teoretické vedomosti z literatúry, avšak nemajú reálnu predstavu, ako fungujú konkrétne zariadenia a ako prebiehajú jednotlivé výrobné procesy, teda nie sú dostatočne pripravení ani na laboratórne cvičenia.

Cieľom zavedenia kamerového systému je zlepšiť tento stav a zabezpečiť všetkým študentom rovnaké podmienky na štúdium. Kamerový systém sme navrhli tak, aby umožňoval zachytiť prácu na každom zariadení, ktoré sa nachádza v laboratóriách. Prihliadali sme pritom na rôzne

povahy týchto zariadení, z čoho vyplývajú aj rôzne požiadavky (hlavne šírka, výška, vzdialenosť, rozlíšenie) na zábery z kamier. Pri niektorých z nich stačí zachytiť samotné zariadenie, pri iných je dôležitejšie sledovať samotný spôsob práce a jednotlivé postupy.

Samotný návrh počíta s možnosťou umiestniť 2 kamery do každého laboratória, pričom kamery bude možné jednoducho natočiť do potrebného smeru pre sledovanie konkrétneho zariadenia. Z finančných dôvodov, v prvej fáze nasadenia systému, nebudú k dispozícii všetky kamery, ale iba konzoly pre uchytenie, na ktoré bude možné jednoducho namontovať kameru do miestnosti, v ktorej práve prebieha výučba. Každá kamera bude vybavená objektívom so zoomom, pomocou ktorého bude možné presne nastaviť a priblížiť oblasť, ktorá sa má nahrávať.

V prípade potreby vytvorenia prenosu resp. nahrávky z jedného procesu z viacerých uhlov záberu resp. z viacerých pohľadov, bude k dispozícii prenosný statív, na ktorý bude možné umiestniť ďalšiu IP kameru, alebo videokameru, ktorej obraz bude možné on-line vysielat pomocou televíznej karty a notebooku.

B. Návrh rozmiestnenia kamier a prípojných bodov štruktúrovanej kabeľáže

Katedra sa nachádza na prízemí, v dvoch protiľahlých častiach budovy. Internet je zabezpečený pomocou dvoch prípojných bodov, pričom táto sieť bola vybudovaná už dávnejšie, postupne sa rozširovala podľa aktuálnych potrieb, avšak bez jasného konceptu. Centrálny rozvádzač sa nachádza na poschodí nad katedrou. V súčasnosti sú rozvádzače (IDF) zabezpečené jednoduchými prepínačmi, ktoré sa navyše nedajú manažovať. Kabeľáž je tvorená väčšinou káblom štandardu cat.5, vedeným mimo vodiacich lúčov, zakončená na oboch koncoch konektorom RJ45. Takáto kabeľáž je náchylná na poruchy. Na to, aby do nej bolo možné pripojiť kamerový systém založený na IP technológii, od ktorej sa vyžaduje spoľahlivá prevádzka, je potrebné túto sieť rozšíriť o ďalšie prvky. Zároveň, rozširovanie siete poskytuje ideálnu príležitosť urobiť štruktúrovanú kabeľáž podľa aktuálnych štandardov, čo zabezpečí zvýšenie bezpečnosti a spoľahlivosti.

Pre obidve prípojky sa vytvorí IDF, ktorý bude umiestnený do uzamykateľnej skrinky určenej pre takéto účely – tzv. rack skriňa resp. dátový rozvádzač. Prípojka bude zabezpečená linkou rýchlou 1Gbps. Rozvody budú ukončené v rozvádzači v tienených patch-paneloch, na druhom konci budú terminované tienenými ethernetovými zásuvkami. Do každého rozvádzača sa nainštaluje prepínač, ktorý podporuje štandard IEEE 802.1Q (podpora virtuálnych LAN). Týmto bude možné v rámci jednej siete od seba logicky oddeliť kancelárske počítače od počítačov študentov, ako aj vytvoriť samostatnú LAN pre kamerový systém, čo zvýši bezpečnosť komunikácie po sieti a zabráni možnému odpočúvaniu komunikácie po sieti.

Katedra sa nachádza v staršej budove, ktorá má široké vnútorné steny (hrúbka až 70 cm). Z toho dôvodu je ťažké ťahať cez tieto steny zväzky káblov, je rozumnejšie v prípade potreby pripojenia viacerých zariadení vytvoriť ďalší IDF. Preto pre laboratória sa vytvorí samostatný IDF. Tento rozvádzač bude zabezpečovať samotné pripojenie kamier, prípadne aj ostatných zariadení v laboratóriách. V prípade tohto rozvádzača bude potrebné zabezpečiť napájanie zariadení podporujúcich štandard IEEE 802.3af (napájanie cez Ethernet). Pripojenie medzi jednotlivými rozvádzačmi bude zabezpečené linkou poskytujúcou rýchlosť 1Gbps.

C. Technické pozadie riešenia

Záznam bude zabezpečený 1,3Mpix IP kamerami, ktoré sú vybavené 1/3"-ým CMOS čipom. Na kamery bude nasadený objektív s rozsahom 2,8 – 12 mm a s automaticky riadenou clonou (alternatívne bude možné vymeniť objektív za iný, s rozsahom 5 - 50mm). Tento rozsah zabezpečí možnosť nahrávať detailný záznam, aj celkový pohľad na človeka pracujúceho so zariadením. V tabuľke Tab.1 a Tab.2 je prehľad vybraných ohniskových vzdialeností s vypočítaným uhlom záberu a s vypočítanou plochou záberu.

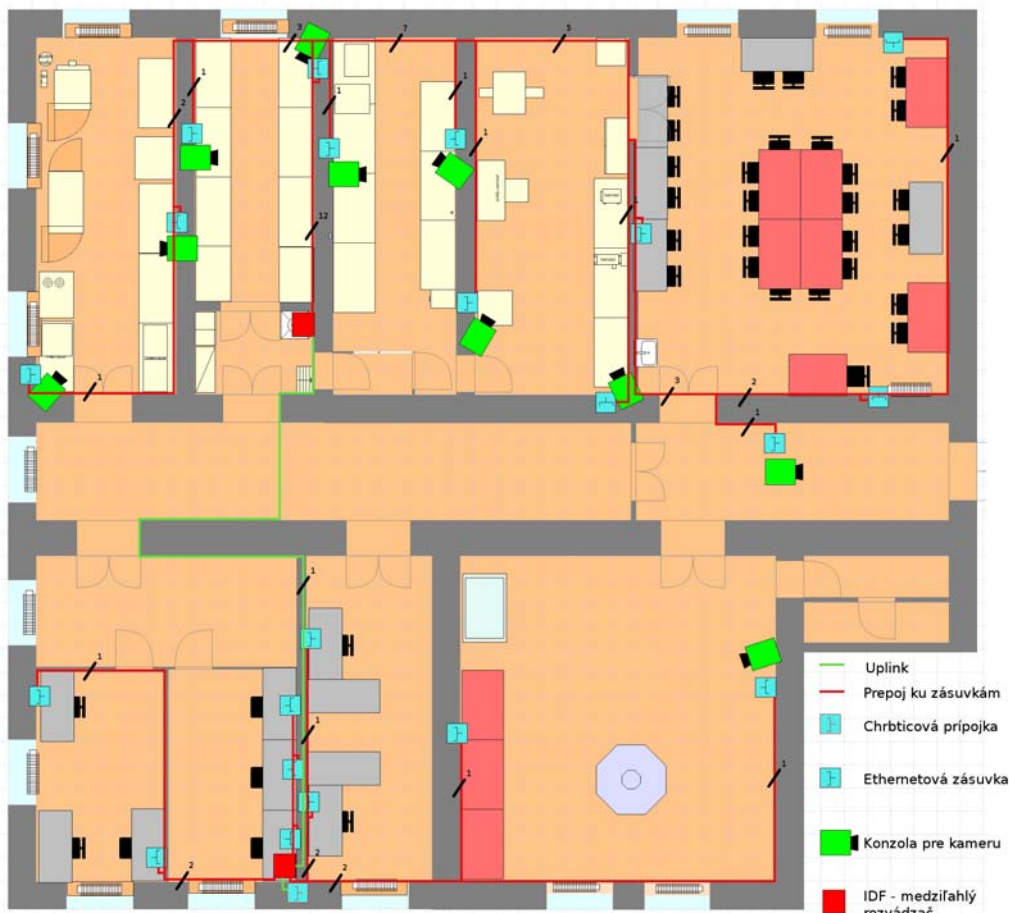
Tieto kamery sú schopné nahrávať HD video s frekvenciou 15fps, ktoré je dostatočné pre scény, pri ktorých je potrebné vyššie rozlíšenie, pričom na scéne sa nevyskytujú rýchle pohyby (väčšina zariadení na katedre). V prípade, že je potrebné zaznamenať nejakú scénu plynulo, kamery sú schopné nahrávať frekvenciou 30fps v rozlíšení VGA.

V prípade potreby detailného záberu z iného uhla, než ktoré umožňujú namontované držiaky pre kamery, bude k dispozícii prenosný statív, na ktorý bude možné umiestniť IP kameru, alebo aj klasickú digitálnu videokameru.

Všetky video streamy budú v rámci LAN vysielané multicastom, čím je zabezpečené, že v rámci LAN budú prenosy prístupné pre každého. Na to však, aby video prenos bolo možné sledovať aj z miesta mimo LAN, je potrebné zabezpečiť unicastové vysielanie. Z toho dôvodu bude implementovaný video server, ktorý bude zbierať jednotlivé multicastové vysielania a

bude ich vysielat' do internetu formou unicastového prenosu. Toto riešenie je výhodné jednak z hľadiska, že samotné kamery nebudú preťažované viacnásobným prístupom a tiež tým, že jediným zariadením, ktoré bude prístupné z internetu, bude video server.

Video server bude zároveň poskytovať aj audio a video archív. Bude tiež zabezpečovať nahrávanie každého on-line záznamu pre prípad, že by ich bolo potrebné dodatočne zaradiť do archívu.



Obr. 1 Rozmiestnenie konzol pre kamery a jednotlivých častí štruktúrovanej kabeláže

Súčasťou celého systému budú aj už existujúce počítače, ktoré sa nachádzajú v laboratóriách. Tieto počítače sú väčšinou pripojené na zariadenia v laboratóriách, a sú používané buď na priame riadenie zariadení, alebo na získavanie a spracovanie údajov z niektorých technologických zariadení. V rámci e-Learningového systému budú sprístupnené aj obrazovky týchto počítačov vo forme on-line videa, s možnosťou komentovať tieto videá priamo počas nahrávania. Toto streamovanie pracovných plôch bude zabezpečené pomocou programu VLC, ktorý dokáže streamovať obraz pracovnej plochy počítača a pomocou ktorého je možné pripojiť aj externý mikrofón pre prenos komentárov ku videu.

Podobným spôsobom sa budú vysielat' aj prednášky. Jednak celá prednáška bude vysielaná pomocou IP kamery, ale taktiež sa bude nahrávať pracovná plocha počítača s prezentáciou, ktorá bude doplnená o zvuk z mikrofónu prednášajúceho. Týmto zabezpečíme, že študenti si budú môcť pozrieť identickú prednášku kedykoľvek počas prípravy na skúšku a taktiež budú výukové materiály chránené proti neautorizovanému kopírovaniu, vytlačeniu a pod.

Celý systém bude mať jednotné rozhranie vo forme internetovej stránky, ktorá bude mať implementovanú autorizáciu, autentifikáciu a správu používateľov ako aj používateľských skupín. Jednotlivé videá budú prístupné iba autorizovaným osobám. Osoby s príslušnými právami budú môcť na diaľku zapínať počítače pomocou funkcie WOL (wake on lan) a taktiež spustiť streamovanie pracovnej plochy. Na to posluží jednoduchá klientska aplikácia, ktorá bude počúvať na inštrukcie vysielané videoseverom. Tento portál bude poskytovať rozhranie na výber, ktoré on-line streamy sa majú uchovať a zaradiť do video archívu.

Multimédiá v archíve budú kategorizované, bude možné osobitne prehliadať videá jedného prednášajúceho, jedného predmetu, jednej oblasti a pod.

D. Výpočet rozmerov snímanej plochy a určenie potrebných objektívov v závislosti na vzdialenosti medzi kamerou a snímaným objektom

Použité kamery majú snímací čip formátu 1/3", teda ich rozmery sú 4,8x3,6mm (pomer strán 4:3). Vzdialenosť, pre ktorú je potrebné zabezpečiť snímání v laboratóriách, sa pohybuje od 1,5m až po 6m.

Potrebnú ohniskovú vzdialenosť je možné vypočítať pomocou nasledujúceho vzorca:

$$\text{ohnisková vzdialenosť} = \frac{\text{vzdialenosť snímaného objektu}}{\text{šírka snímanej plochy}} \times \text{šírka snímacieho čipu}$$

Z uvedeného vzorca vieme vyjadriť šírku záberu nasledovným spôsobom:

$$\text{šírka snímanej plochy} = \frac{\text{vzdialenosť snímaného objektu}}{\text{ohnisková vzdialenosť}} \times \text{šírka snímacieho čipu}$$

Tabuľka 1
 ROZMERY SNÍMANEJ SCÉNY PRI RÔZNYCH VZDIALENOSTIACH PRE ZÁKLADNÝ OBJEKTÍV

Vzdialenosť scény	Oblasť záberu pri ohniskovej vzdialenosti 2,8mm		Oblasť záberu pri ohniskovej vzdialenosti 12mm	
	Šírka	Výška	Šírka	Výška
1,5m	2,5m	1,86m	0,6m	0,45m
3m	5,14m	3,86m	1,2m	0,9m
4,5m	7,71m	5,79m	1,8m	0,45m
6m	10,29m	7,71m	2,4m	1,8m

Tabuľka 2
 ROZMERY SNÍMANEJ SCÉNY PRI RÔZNYCH VZDIALENOSTIACH PRE ALTERNATÍVNY OBJEKTÍV

Vzdialenosť scény	Oblasť záberu pri ohniskovej vzdialenosti 5mm		Oblasť záberu pri ohniskovej vzdialenosti 50mm	
	Šírka	Výška	Šírka	Výška
1,5m	1,44m	1,08m	0,14m	0,11m
3m	2,88m	2,16m	0,28m	0,21m
4,5m	4,32m	3,24m	0,43m	0,32m
6m	5,76m	4,32m	0,58m	0,44m

III. ZÁVER

Navrhnutý systém pomôže zefektívniť výučbu predmetov, ktorých výučba prebieha v priestoroch Katedry technológií v elektronike. Zároveň rozšíri bázu študijných materiálov o video ukážky a poskytne tak študentom väčší priestor na osvojenie si jednotlivých krokov niektorých výrobných postupov. Zabezpečí tiež väčšiu teoretickú a z časti aj praktickú pripravenosť na samostatnú prácu v rámci laboratórnych cvičení, prípadne tak semestrálnych ako aj záverečných prác. V neposlednom rade bude pre študentov motiváciou pre priebežné štúdium a tým prispeje k zvýšeniu kvality vzdelávania.

POĎAKOVANIE

Tento článok vznikol ako súčasť riešeného projektu KEGA 3/7117/09.

LITERATÚRA

- [1] Sidney F. Ray, „Applied photographic optics: lenses and optical systems for photography“, Focal Press, 2002, ISBN: 0240515404
- [2] David Austerberry, „The technology of Video and Audio Streaming – Second Edition“, Focal Press, 2005, ISBN: 0240805801
- [3] Tom Sheldon, „McGraw-Hill's Encyclopedia of Networking and Telecommunications“, McGraw-Hill Professional, 2001, ISBN: 0072120053
- [4] IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture 802-2001, The IEEE, Inc, ISBN: 0738129410
- [5] Lieberman Software Corporation, „White Paper: Wake on LAN Technology“, <http://liebesoft.com>, 2006
- [6] Herman Kruegle, „CCTV Surveillance, 2nd Edition: Video Practices and Technology, Butterworth-Heinemann, 2006, ISBN: 9780750677684

Harmonické potenciálové pole pri orientácii robotov v priestore

Michal KALAVSKÝ

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

michal.kalavsky@tuke.sk

Abstrakt — Harmonické potenciálové polia je jedna z bežne používaných metód pre plánovanie cesty robotov v priestore s prekážkami. Zväčša sa používa pre statické prostredie, čo znamená, že vopred poznáme veľkosť prostredia, polohu cieľa a polohu prekážok. Metóda harmonických potenciálových polí zaradíme medzi metódy plánovania na mriežke.

Kľúčové slová — potenciál, Dirichletov problém

I. ÚVOD

V súčasnosti poznáme viacero metód, ktoré sa používajú pri určovaní smeru robotov v priestore, čiže pri plánovaní ich cesty. Medzi tieto metódy patria napríklad Bug algoritmy, exaktné plánovanie, plánovanie na mriežke, či neurónové siete. Použitie jednotlivých metód závisí od kritérií, ktoré musí daná metóda spĺňať pri riešení úlohy. V tomto článku sa budeme venovať metóde, ktorá patrí do kategórie plánovania na mriežke.

II. METÓDA HARMONICKÝCH POTENCIÁLOVÝCH POLÍ

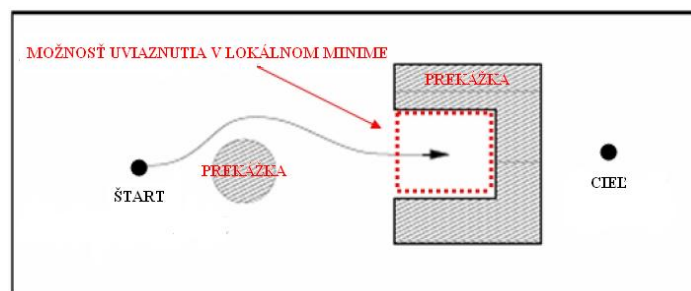
Jednou z najpoužívanejších metód pre plánovanie cesty robotov v priestore je metóda potenciálových polí. Ako už bolo vyššie spomenuté, táto metóda patrí do kategórie plánovania na mriežke. Čo vlastne znamená plánovať na mriežke? Predstavme si, že máme prostredie, v ktorom sa má robot pohybovať. Dané prostredie rozdelíme na bunky, pričom každá bunka bude predstavovať určitú časť daného priestoru a každá bunka bude mať určité ohodnotenie. Takýmto spôsobom dosiahneme, že prostredie sa rozdelí na určitý počet buniek resp. oblastí, čo nám vlastne predstavuje mriežku s určitým počtom buniek.

A. Potenciálové polia

Potenciálové polia sú založené na konštrukcii virtuálnych máp, ktoré simulujú reálne prostredie. To znamená, že pomocou potenciálových hladín vytvárame virtuálnu mapu reálneho prostredia, ktorú má robot k dispozícii pri určovaní smeru. Pohybuje sa smerom na vrchol, čiže v smere najvyššieho nárastu potenciálu.

Ako už z názvu môžeme vidieť, pri metóde potenciálových polí sa pri určovaní smeru využíva potenciál. Pomocou potenciálu vznikajú potenciálové hladiny, pomocou ktorých sa robot vie rozhodnúť, ktorým smerom sa má vydať, aby sa bezpečne dostal do cieľa. Bunke, ktorá bude reprezentovať počiatočnú pozíciu, priradíme najnižšiu hodnotu a bunke, ktorá bude reprezentovať cieľovú pozíciu, priradíme najvyššiu hodnotu. Týmto ohodnotením dosiahneme, že robot sa bude pohybovať smerom na vrchol. V priestore sa môžu vyskytovať aj prekážky, čiže robot musí byť schopný sa im vyhýbať. Tu vzniká priestor pre základný predpoklad pri metóde potenciálových polí, ktorým je, že všetky prekážky musia byť konvexné. Ak tento predpoklad nie je splnený, môže sa stať, že robot uviazne v bode, ktorý nie je cieľový a zacyklí sa. Potenciálová funkcia vytvára v tomto bode lokálne minimum, čo je hlavným nedostatkom metódy potenciálových polí. Tvorbu lokálneho minima môžeme vidieť na Obr. 1 (viď nižšie). Tento problém sa dá vyriešiť použitím harmonických potenciálových polí, kde existuje len jedno minimum, a to globálne minimum reprezentované cieľovou pozíciou.

Keďže metóda potenciálových polí je založená na potenciáli, potenciálových hladinách. Je potrebné si definovať, čo je potenciál. Budeme uvažovať elektrický potenciál. Pomocou analógie s elektrickým potenciálom môžeme definovať metódu potenciálových polí.



Obr. 1 Možnosť tvorby lokálnych miním pri metóde potenciálových poli

Elektrický potenciál je skalárna fyzikálna veličina, ktorá popisuje potenciálnu energiu jednotkového elektrického náboja v nemennom elektrickom poli. Z toho môžeme definovať potenciál medzi dvoma bodmi ako množstvo práce potrebné na prenesenie tohto jednotkového elektrického náboja z bodu vzťažného, kde je potenciál nulový, do daného miesta. Za vzťažný bod (miesto s nulovým potenciálom) považujeme nekonečne vzdialený bod alebo povrch zeme.

Podľa [9] je z definície potenciálu zrejmé, že

$$d\varphi = -\vec{E} \cdot d\vec{r} \quad (1)$$

kde $\vec{E}$ je vektor intenzity elektrického poľa a $d\vec{r}$ je dráhový element.

Pre

$$\vec{E} = E_x \vec{i} + E_y \vec{j} + E_z \vec{k}$$

a pre

$$d\vec{r} = dx \vec{i} + dy \vec{j} + dz \vec{k}$$

má skalárny súčin tvar:

$$\vec{E} d\vec{r} = E_x dx + E_y dy + E_z dz \quad (2)$$

Pre $\varphi = \varphi(x, y, z)$ potom môžeme vyjadriť celkový diferenciál $d\varphi$ v tvare

$$d\varphi = \frac{\partial \varphi}{\partial x} dx + \frac{\partial \varphi}{\partial y} dy + \frac{\partial \varphi}{\partial z} dz \quad (3)$$

Rovnicu (1) môžeme prepísať pomocou rovníc (2) a (3) do tvaru

$$\frac{\partial \varphi}{\partial x} dx + \frac{\partial \varphi}{\partial y} dy + \frac{\partial \varphi}{\partial z} dz = -E_x dx - E_y dy - E_z dz \quad (4)$$

a z toho na základe nezávislosti zložiek x, y, z vyplývajú zložky vektora $\vec{E}$ v tvare:

$$E_x = -\frac{\partial \varphi}{\partial x} \quad E_y = -\frac{\partial \varphi}{\partial y} \quad E_z = -\frac{\partial \varphi}{\partial z} \quad (5)$$

Vektor intenzity elektrického poľa $\vec{E}$ je možné vyjadriť v tvare:

$$\vec{E} = E_x \vec{i} + E_y \vec{j} + E_z \vec{k} = -\frac{\partial \varphi}{\partial x} \vec{i} - \frac{\partial \varphi}{\partial y} \vec{j} - \frac{\partial \varphi}{\partial z} \vec{k} \quad (6)$$

Použitím operátora *grad* v karteziánskom súradnicovom systéme dostávame vyjadrenie

$$\text{grad } \varphi = \frac{\partial \varphi}{\partial x} \vec{i} + \frac{\partial \varphi}{\partial y} \vec{j} + \frac{\partial \varphi}{\partial z} \vec{k} \quad (7)$$

čo môžeme pomocou rovnosti z (6) zapísať ako

$$\vec{E} = -\text{grad } \varphi \quad (8)$$

kde $-\text{grad } \varphi$ je vektor v smere $\vec{E}$. Rovnica (8) umožňuje stanoviť intenzitu elektrického poľa zo známeho rozloženia elektrického potenciálu. Dosadením z rovnice (1) a (8) vzniká vzťah

$$d\varphi = \text{grad } \varphi d\vec{r} \quad (9)$$

Ak je dráhový element $d\vec{r}$ kolmý na vektor $\vec{E}$ a tým pádom je kolmý aj na $\text{grad } \varphi$, potom bude potenciál konštantný a

$$d\varphi = 0 \quad (10)$$

Grad φ určuje smer najväčšieho nárastu potenciálu φ . Absolútna hodnota grad φ sa rovná prírastku potenciálu na jednotku dĺžky v smere vektora $-\vec{E}$. Kladný elektrický náboj sa

pôsobením elektrických síl môže pohybovať smerom od vyšších hladín potenciálu k nižším, podobne ako je to u padajúceho telesa v gravitačnom poli Zeme.

Analogicky môžeme popísať umelé potenciálové pole, kde podľa [1] definujeme silové pole ako záporný gradient potenciálu, ktorý je priradený tomuto poľu. Matematicky to môžeme zapísať nasledovne:

$$\vec{F} = -\left\{\frac{\partial U}{\partial x}\vec{i} + \frac{\partial U}{\partial y}\vec{j} + \frac{\partial U}{\partial z}\vec{k}\right\} = -\nabla U \quad (11)$$

Ak v poli neprebíha žiadny proces, to znamená, že je ustálené, tak potom platí:

$$\nabla \vec{F} = 0 \quad (12)$$

Po dosadení z (11) dostávame:

$$\nabla \cdot \nabla U \equiv \nabla^2 U = 0 \quad (13)$$

alebo

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0 \quad (14)$$

Posledný vzorec (14) predstavuje Laplaceovu rovnicu popisujúcu správanie ľubovoľného ustáleného potenciálového poľa.

B. Harmonické potenciálové polia

Harmonické potenciálové pole môžeme podľa [4] definovať ako funkciu na množine Ω nasledovne:

$$\nabla^2 U(x) = 0 \quad (15)$$

pričom x patrí do množiny Ω a ∇^2 je Laplaceov operátor.

Keďže pri metóde potenciálových polí vytvárame virtuálnu mapu reálneho prostredia, je potrebné stanoviť hranice prostredia a hranice prekážok. Okraje množiny Ω definujeme ako okraje všetkých prekážok. Čo to znamená? Keďže každé ohraničené prostredie má nejaké svoje hranice, pre robota tieto hranice prostredia znamenajú v podstate prekážku. Preto musíme hranice prostredia začleniť taktiež medzi okraje všetkých prekážok. Ďalej je potrebné definovať aj cieľovú pozíciu. Z týchto poznatkov nám vznikajú podmienky, ktoré musí potenciálové pole spĺňať. Sú nimi podmienky v tvare:

1. $U|_{\Omega} = 0$
2. $U_{(cieľ)} = \text{konšt.}$

Prvá podmienka stanovuje hodnotu pre okraje všetkých prekážok prostredia, čiže aj hranice prostredia, druhá podmienka stanovuje hodnotu pre cieľovú pozíciu.

Pri metóde harmonických potenciálových polí potrebujeme je potrebné vyriešiť problém parciálnej diferenciálnej rovnice hraničných (okrajových) hodnôt, tak ako je to popísané v Dirichletovom probléme. Dirichletovom problém môžeme popísať nasledovne. Všetky fyzikálne a technické problémy, ktoré vedú k riešeniu Laplaceovej rovnice, sa vyznačujú tým, že patria do triedy okrajových problémov. Do týchto okrajových problémov patrí aj Dirichletov problém, kde je na okraji oboru predpísaná hodnota hľadaného riešenia. Pri okrajových problémoch hľadáme harmonickú funkciu. Z princípu minima a maxima vyplýva, že Dirichletov problém je jednoznačný, čo znamená, že má najviac jediné riešenie a závisí spojitě od okrajových podmienok. V tomto prípade použijeme na riešenie parciálnych diferenciálnych rovníc práve tento Dirichletov problém, keďže pri metóde harmonických potenciálových polí potrebujeme mať iba jedno minimum, a to globálne minimum. Tieto poznatky ohľadom riešenia okrajových problémov majú veľký praktický význam. Ak totiž určíme nejakú metódu pre riešenie okrajovej úlohy a vieme, že podľa uvedených viet existuje nanajvýš jedno riešenie tejto okrajovej úlohy, potom vieme, že každou inou použitou metódou musíme dospieť k rovnakému riešeniu.

Analogicky z (11) môžeme určiť vzťah pre pohyb robota v takomto potenciálovom poli, ktorý bude mať tvar:

$$\dot{x} = -\nabla U(x) \quad (16)$$

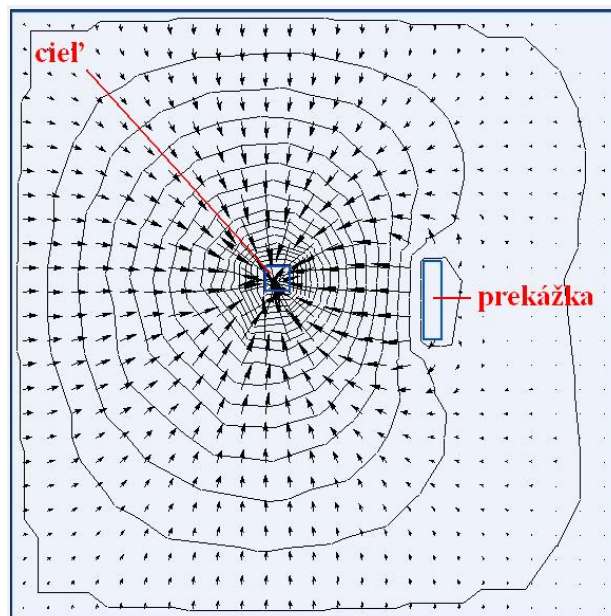
čo môžeme ďalej zapísať ako:

$$\dot{x} = -\frac{\partial U}{\partial x} \quad \dot{y} = -\frac{\partial U}{\partial y} \quad (17)$$

čo platí pre dvojrozmernú sústavu.

C. Experimentálne výsledky

Doposiaľ sa nám podarilo dosiahnuť určité experimentálne výsledky, ktoré sú znázornené na Obr. 2 (viď nižšie). Pomocou simulačného programu Students' Quick Field sa nám podarilo vytvoriť potenciálové pole, ktoré simuluje reálne prostredie. Na obrázku môžeme vidieť dvojrozmerné prostredie, v ktorom sa nachádza jedna prekážka. Keďže cieľ je jediné existujúce minimum v tomto prostredí, tak robot sa vždy dostane do cieľa, pretože jeho smer je stále vedený do tohto minima nech sa nachádza v ktoromkoľvek bode prostredia. Od prekážky a od hraníc prostredia pôsobí odpudzivá sila, k cieľu pôsobí príťažlivá sila. Na Obr. 2 sú znázornené aj potenciálové hladiny.



Obr. 2 Výsledok simulačného programu Students' Quick Field pri určovaní pohybu v potenciálovom poli

III. ZÁVER

Metóda harmonických potenciálových polí je často používaná metóda pri plánovaní cesty robotov v prostredí s prekážkami. Väčšinou sa používa pre statické prostredie, čo znamená, že mapu prostredia poznáme vopred. Tak tomu bolo aj v našom experimentálnom príklade, kde sme poznali pozíciu prekážky a pozíciu cieľa. Harmonické potenciálové polia sú použiteľné aj pre dynamické prostredie, čo by mohlo byť cieľom ďalšej štúdie.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

LITERATÚRA

- [1] Vaščák, Ján: Využitie potenciálových polí v navigácii mobilných robotov, Technická univerzita v Košiciach, 2008.
- [2] Winkler, Zbyněk: Plánování na mřížce, 2003-12-03, dostupné na internete: <<http://robotika.cz/guide/gridplan/cs>>
- [3] Siegwart, R. – Nourbakhsh, I.: Potential fields, October 16 2008.
- [4] Masoud, A. Ahmad – Masoud, A. Samer: Motion Planning in the Presence of Directional and Obstacle Avoidance Using Nonlinear, Anisotropic, Harmonic Potential Fields, Mechanical Engineering Department, Jordan University of Science & Technology, San Francisco, April 2000.
- [5] Shahab, Mohammad: Harmonic Potential Field Motion Planning, King Fahd of Petroleum and Minerals, 7 February 2009.
- [6] Hwang, K. Yong – Ahuja, Narendra: A potential Field Approach to Path Planning, IEEE Transactions on Robotics and Automation vol. 8 no. 1, February 1992.
- [7] Ge, S. S. – Cui, Y. J.: New Potential Functions for Mobile Robot Path Planning, IEEE Transactions on Robotics and Automation vol. 16 no. 5, October 2000.
- [8] RNDr. Pavel Kubeš, CSc. – doc. RNDr. Zdeněk Kyncl, DrSc.: Fyzika I, České Vysoké Učení Technické v Praze, Fakulta elektrotechnická, Ediční středisko ČVUT, Praha, 1991.
- [9] Mayer, D. – Polák, J.: Metody řešení elektrických a magnetických polí, SNTL – Nakladatelství technické literatury, Alfa, Praha 1983.

Neurónový estimátor rýchlosti v štruktúre vektorového riadenia

Peter NGUYEN, Jaroslava ŽILKOVÁ

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

peter.nguyen@tuke.sk, jaroslava.zilkova@tuke.sk

Abstrakt — Tento príspevok sa zaoberá využitím umelej neurónovej siete pre odhad uhlovej rýchlosti asynchrónneho motora s kotvou nakrátko. Pre daný asynchrónny motor je navrhnuté vektorové riadenie, z ktorého sú prostredníctvom simulácie získané učiace vzorky potrebné pre off-line učenie doprednej neurónovej siete. Navrhnutý neurónový estimátor je simulačne overený pre rôzne hodnoty rýchlosti a záťažného momentu.

Kľúčové slová — asynchrónny motor s kotvou nakrátko, bezsnímačové riadenie, neurónová sieť, vektorové riadenie

I. ÚVOD

Vektorovo riadený asynchrónny motor je v súčasnosti najčastejšie používaným typom elektrického pohonu. Je to vďaka tomu, že asynchrónny motor s kotvou nakrátko má veľmi výhodné vlastnosti oproti jednosmernému cudzobudenému motoru (JSCB), ktorý bol donedávna najčastejšie používaným typom motora kvôli jeho jednoduchšej regulácii. Asynchrónny motor má jednoduchšiu konštrukciu a z toho vyplýva aj nižšia cena motora, oveľa jednoduchšiu údržbu, vyššiu spoľahlivosť atď.

Nevýhodou pohonu s asynchrónnym motorom s kotvou nakrátko (AM) je zložitosť jeho riadenia. Samotný AM predstavuje oveľa zložitejší nelineárny systém ako JSCB motor. V 70.-tých rokoch minulého storočia publikoval Blaschke [1] vo svojej práci základné princípy oddeleného riadenia momentu a toku AM, podobne ako je to u JSCB motora. Zdokonalením riadiacich algoritmov a vývojom nových výkonových polovodičových súčiastok bola umožnená implementácia vektorového riadenia, čo spôsobilo, že sa AM začal používať aj v aplikáciách, v ktorých bolo potrebné riadiť rýchlosť, polohu, moment alebo inú veličinu.

V súčasnosti je venovaná veľká pozornosť riadeniu motora bez použitia snímača otáčok, tzv. bezsnímačového riadeniu. Hlavnými výhodami, ktoré so sebou prináša odstránenie snímača otáčok, sú menšie rozmery a nižšia hmotnosť pohonu (dôležité napr. v leteckom priemysle), vyššia spoľahlivosť, nižšia cena pohonu atď. Ďalšou veľkou oblasťou záujmu vedcov je aplikácia umelej inteligencie či už v riadení pohonov, alebo jej využitie pre pozorovanie resp. odhad veličín alebo parametrov motora.

Pre odhad rôznych veličín je možné použiť napríklad umelé neurónové siete, pretože majú výborné aproximačné vlastnosti. Tento príspevok sa venuje využitiu umelej neurónovej siete pre odhad rýchlosti motora v štruktúre vektorového riadenia.

II. VEKTOROVÉ RIADENIE

V aplikáciách, kde je požadovaná vysoká dynamika a presnosť riadenia sa používa prevažne vektorové riadenie AM. Vektorové riadenie poskytuje nezávislé riadenie momentu a toku stroja, a teda je analógiou riadenia jednosmerného cudzobudeného motora. Zložku statorového prúdu, ktorá odpovedá magnetickému toku stroja označujeme tokotvornou zložkou (i_{1x}), a zložku odpovedajúcu momentu stroja označujeme momentotvornou zložkou (i_{1y}).

Podľa toho, na ktorý magnetický tok je orientovaná tokotvorná zložka statorového prúdu, môže byť vektorové riadenie orientované na [2]:

- magnetický tok vo vzduchovej medzere
- magnetický tok statora
- magnetický tok rotora

Dôležitou časťou vektorového riadenia je určenie polohy vektora magnetického toku, pretože od presnosti jeho určenia závisí aj správna činnosť vektorového riadenia. Z polohy vektora toku sa počíta uhol potrebný pre transformáciu veličín zo stacionárnej súradnicovej sústavy (α, β) do synchronne rotujúcej sústavy (x, y) a naopak. Podľa spôsobu získavania polohy vektora magnetického toku môžeme vektorové riadenie rozdeliť na priame a nepriame.

Orientácia na rotorový tok sa využíva najčastejšie, či už v podobe priameho alebo nepriameho vektorového riadenia, pretože sa v podstate jedná o jedinú schému, pri ktorej dochádza k úplnému odstráneniu akejkoľvek väzby medzi momentotvornou a tokotvornou zložkou statorového prúdu [2].

Na účely simulácie neurónového estimátora uhlovej rýchlosti bolo najprv navrhnuté vektorové riadenie AM, ktoré sa navrhuje na základe všeobecných rovníc motora, odvodených pri známych zjednodušujúcich predpokladoch, po prepočítaní na dvoj pólový stroj a do synchronne sa točiacej súradnicovej sústavy (x, y). Ďalej po doplnení zosilnenia meniča K_t , úprave a rozpísaní na zložky dostaneme rovnice, na základe ktorých bolo navrhnuté stavové riadenie metódou určenia pólov [3]:

$$\sigma T_1 \frac{di_{1x}}{dt} + i_{1x} = \frac{K_t u_{1x}}{R_1} - (1 - \sigma) T_1 \frac{di_{2m}}{dt} + \sigma T_1 \omega_{2m} i_{1y} \quad (1)$$

$$\sigma T_1 \frac{di_{1y}}{dt} + i_{1y} = \frac{K_t u_{1y}}{R_1} - (1 - \sigma) T_1 \omega_{2m} i_{2m} - \sigma T_1 \omega_{2m} i_{1x} \quad (2)$$

$$T_2 \frac{di_{2m}}{dt} + i_{2m} = i_{1x} \quad (3)$$

$$\frac{i_{1y}}{T_2 i_{2m}} + \omega = \omega_{2m} \quad (4)$$

$$\frac{J}{p} \frac{d\omega}{dt} = \frac{3p}{2} \frac{L_h}{1 + \sigma_2} i_{2m} i_{1y} - m_z \quad (5)$$

Za stavové premenné boli vybrané nasledujúce veličiny:

$$i_{2m} = x_1, \quad i_{1x} = x_2, \quad \omega = x_3, \quad i_{1y} = x_4, \quad m_z = z, \quad u_1 = u_{1x}/K_t, \quad u_2 = u_{1y}/K_t.$$

Dosadením ω_{2m} z rovnice (4) do rovníc (1) a (2) a derivácie i_{2m} z rovnice (3) do rovnice (1) a tiež dosadením za stavové premenné, získame stavovú rovnicu v maticovom tvare:

$$\dot{x} = \begin{bmatrix} -a_1 x_1 + a_1 x_2 \\ f_2(x) \\ a_3 x_1 x_4 \\ f_4(x) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ b & 0 \\ 0 & 0 \\ 0 & b \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ -e \\ 0 \end{bmatrix} z = a(x) + Bu + ez \quad (6)$$

$$y = x_3 = c(x) \quad (7)$$

Kde jednotlivé konštanty a funkcie sú definované nasledovne:

$$a_1 = \frac{1}{T_2}, \quad a_2 = \frac{1}{\sigma T_1}, \quad a_3 = \frac{3p}{2J} \frac{L_h}{1 + \sigma_2}, \quad a_4 = \frac{1 - \sigma}{\sigma}, \quad b = \frac{K_t}{\sigma L_1}, \quad e = \frac{p}{J}$$

$$T_1 = \frac{L_1}{R_1}, \quad T_2 = \frac{L_2}{R_2}, \quad \sigma = 1 - \frac{L_h^2}{L_1 L_2}, \quad \sigma_2 = \frac{L_2 \sigma}{L_h}$$

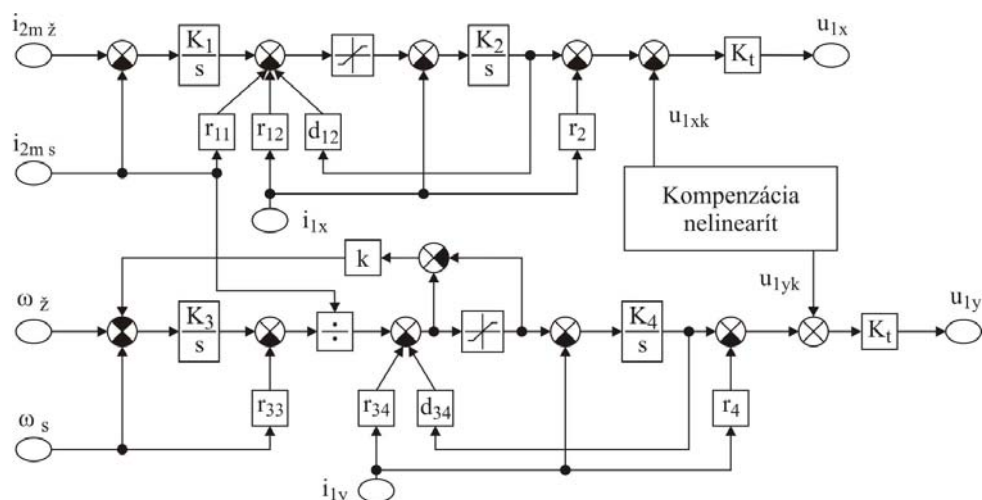
$$f_2(x) = -a_2 x_2 + a_1 a_4 (x_1 - x_2) + x_4 (x_3 + a_1 \frac{x_4}{x_1})$$

$$f_4(x) = -a_2 x_4 - (a_4 x_1 + x_2) (x_3 + a_1 \frac{x_4}{x_1})$$

Stavové riadenie bolo navrhnuté pre AM s kotvou nakrátko s týmito parametrami:

$$\begin{array}{lll} P_n = 750 \text{ W} & p = 2 & R_2' = 9,933 \, \Omega \\ U_n = 220/380 \text{ V} & s = 0.08 & L_h = 0,4494 \text{ H} \\ I_n = 3,8/2,2 \text{ A} & J = 0,0054 \text{ kgm}^2 & L_1 = L_2 = 0,4862 \text{ H} \\ n_n = 1380 \text{ ot/min} & R_1 = 9,978 \, \Omega & L_{1\sigma} = L_{2\sigma} = 0,0368 \text{ H} \end{array}$$

Podrobnejší návrh kaskádnej štruktúry stavového riadenia je uvedený napr. v [3] a [4].



Obr. 1 Regulačná schéma stavového riadenia pre AM s kotvou nakrátko

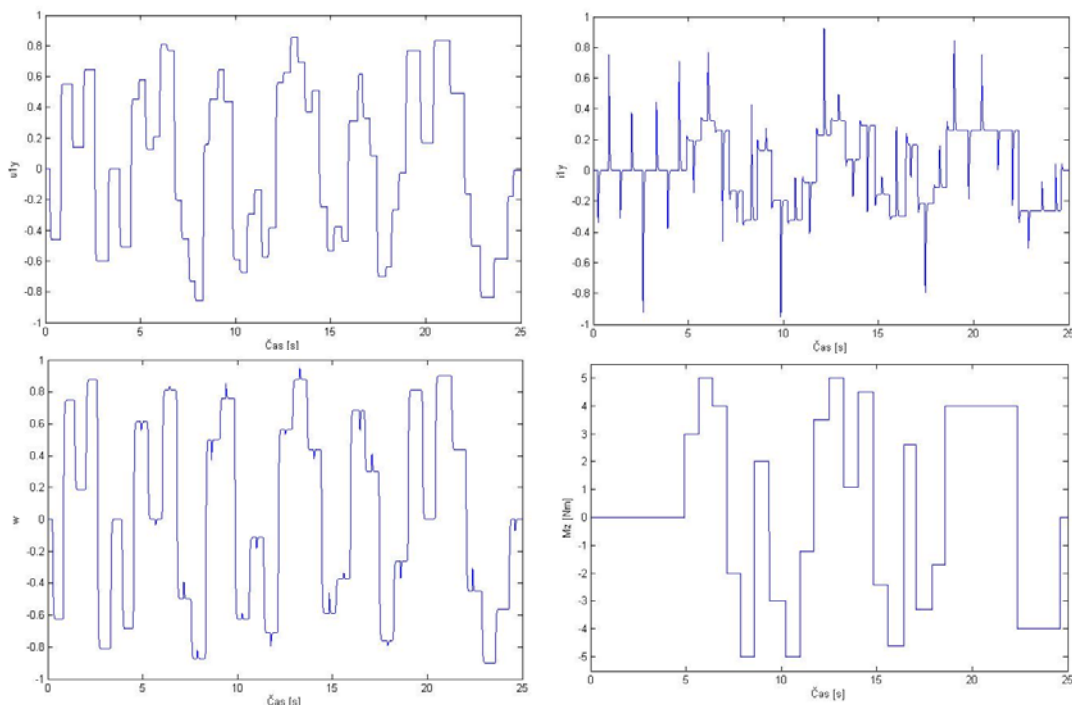
III. NEURÓNOVÝ ESTIMÁTOR RÝCHLOSTI

Vo všeobecnosti pre potreby riadenia je potrebné poznať okamžité hodnoty niektorých veličín a parametrov pohonu. Na tento účel slúžia pozorovatele resp. estimátory, ktoré môžeme rozdeliť z viacerých hľadísk.

Z hľadiska použitej metódy ich môžeme rozdeliť na:

- estimátory využívajúce konštrukčné vlastnosti motora
- estimátory a pozorovatele vychádzajúce z matematického modelu motora
- estimátory a pozorovatele využívajúce metódy umelej inteligencie

Ako už bolo spomínané v úvode, veľká časť výskumu sa sústreďuje na tzv. bezsnímačové riadenie pohonu. V tomto prípade sa informácia o uhlovej rýchlosti získava z pozorovateľa resp. estimátora rýchlosti. Nakoľko umelé neurónové siete dokážu výborne aproximovať aj zložité nelineárne funkcie, môžu sa správne navrhnuté a naučené neurónové siete použiť aj na odhad rýchlosti pohonu.



Obr. 2 Priebeh veličín učiacich vzoriek (u_{1y} , i_{1y} , ω , M_z)

Aby neurónová sieť správne pracovala, musia byť za vstupy do neurónovej siete vybrané vhodné veličiny. Ďalej je potrebné získať vhodnú učiacu množinu, ktorá by mala zohľadňovať všetky stavy riadeného systému.

Na overenie aproximačných vlastností neurónovej siete bola pre odhad rýchlosti zvolená dopredná neurónová sieť. Za vstupy do siete boli vybrané k -té, $(k-1)$ a $(k-2)$ zložky napätia u_{1y} a prúdu i_{1y} . Výstupom z neurónovej siete je odhadovaná uhlová rýchlosť ω . Pretože závislosť medzi vstupnými a výstupnou veličinou je nelineárna, pre skrytú vrstvu s 9 neurónmi, bola použitá nelineárna tansigmoidálna prenosová funkcia. Pre výstupnú vrstvu bola použitá lineárna prenosová funkcia. Neurónová sieť bola učená pomocou Levenberg-Marquardtovej metódy v režime off-line a pre učenie boli použité vzorky získané zo simulácie vektorového riadenia asynchrónneho motora s kotvou nakrátko (obr. 2).

Dopredná neurónová sieť realizuje zobrazenie popísané vzťahom:

$$\omega_{(k)} = f[i_{1y(k)}, i_{1y(k-1)}, i_{1y(k-2)}, u_{1y(k)}, u_{1y(k-1)}, u_{1y(k-2)}, \mathbf{w}] \quad (8)$$

kde f predstavuje prenosovú funkciu a $\mathbf{w}$ predstavuje vektor váhových a prahových koeficientov.

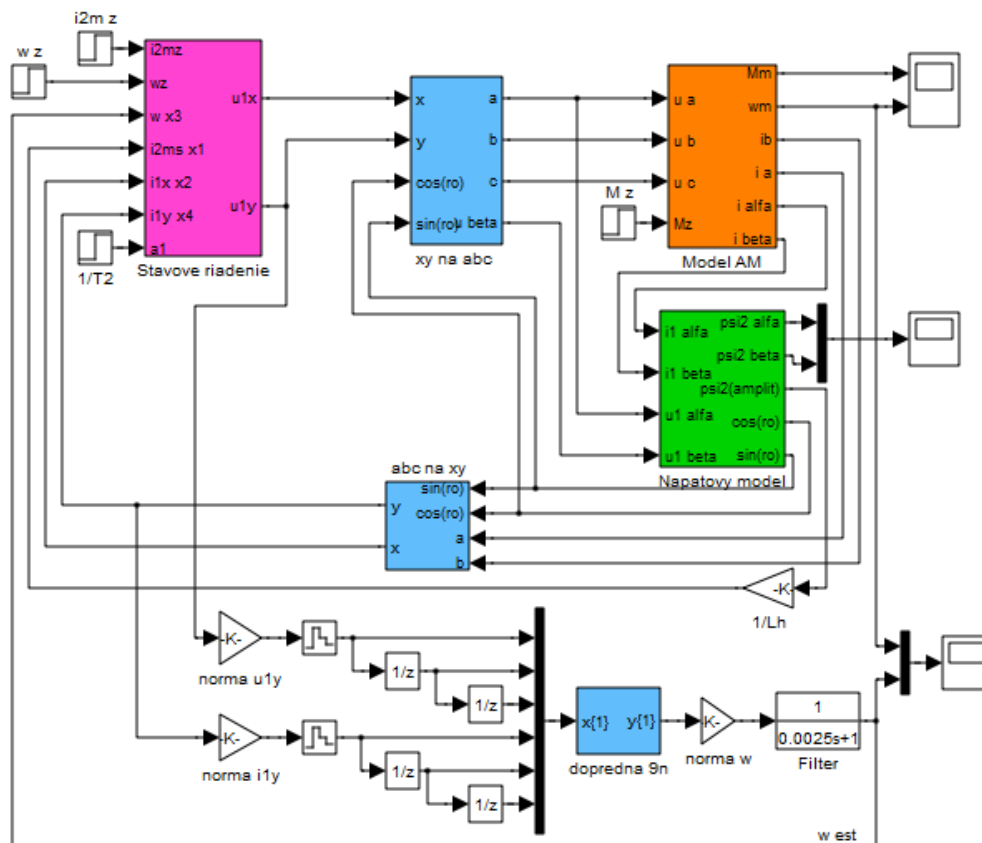
Výstupy prvej a druhej vrstvy doprednej siete sa určia podľa vzťahov [5]:

$$\mathbf{a}_{1(k)} = \text{tansig}(\mathbf{I}\mathbf{w}_i + \mathbf{b}_1), \quad \mathbf{a}_{2(k)} = \text{purelin}(\mathbf{a}_{1(k)}\mathbf{w}_j + b_2) \quad (9)$$

kde tansig a purelin sú prenosové funkcie príslušných vrstiev, $\mathbf{I}$ predstavuje vektor vstupov, $\mathbf{b}_1$ a b_2 sú biasy daných vrstiev neurónovej siete.

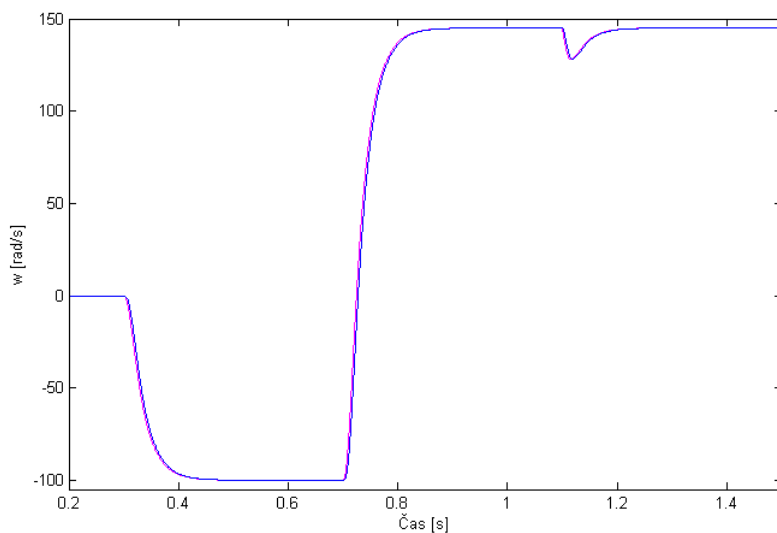
IV. SIMULAČNÉ VÝSLEDKY

Naučená dopredná neurónová sieť bola simulačne otestovaná v štruktúre vektorového riadenia s asynchrónnym motorom, pričom výstup z neurónového estimátora bol privádzaný do riadenia (obr. 3).

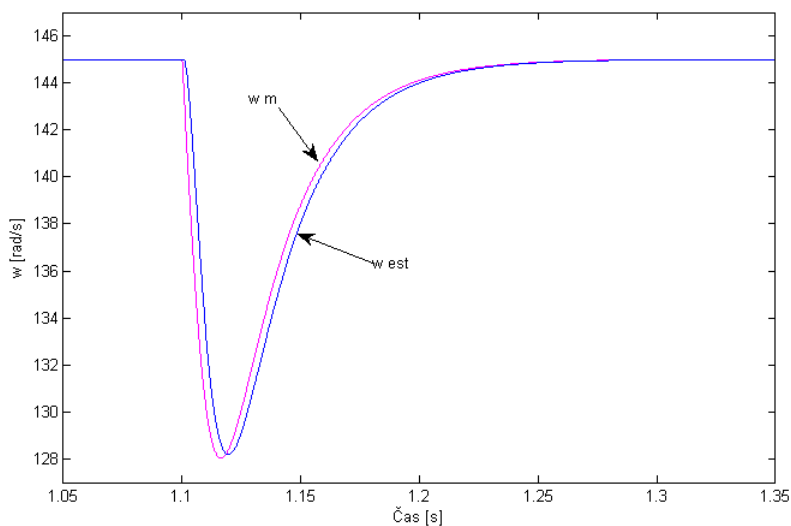


Obr. 3 Bloková schéma pre testovanie neurónového estimátora uhlovej rýchlosti

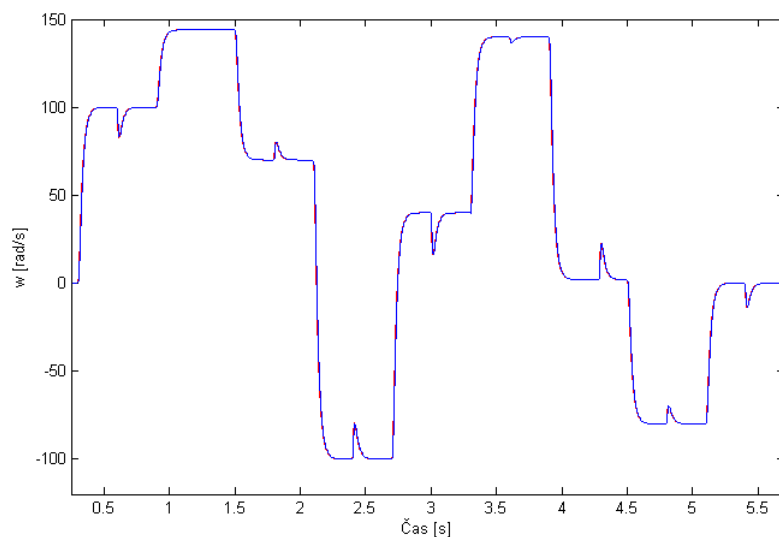
Na obr. 4 je zobrazený priebeh rýchlosti motora a rýchlosti odhadnutej neurónovým estimátorom. V čase 0,35 s bola zadaná zmena žiadanej uhlovej rýchlosti AM z 0 rad/s na -100 rad/s, v čase 0,7 s na +145 rad/s (reverzácia) a v čase 1,1 s bol motor zaťažovaný nominálnym momentom $M_z = 5$ Nm. Obr. 5 zobrazuje detailný pohľad na priebehy rýchlostí z obr. 4 počas zaťaženia, a z obrázka je vidieť určité malé oneskorenie odhadnutej rýchlosti počas prechodného deja. Toto oneskorenie je spôsobené použitím vyhladzovacieho filtra na výstupe z neurónového estimátora. Posledný obrázok zobrazuje priebehy skutočnej a odhadnutej uhlovej rýchlosti pri rôznych hodnotách rýchlosti a zaťaženia motora.



Obr. 4 Priebeh skutočnej a odhadnutej rýchlosti s nominálnym zaťažením v čase 1,1 s



Obr. 5 Detail skutočnej a odhadnutej rýchlosti motora počas zaťaženia momentom $M_z = 5 \text{ Nm}$



Obr. 6 Priebehy skutočnej a odhadnutej rýchlosti pri rôznych hodnotách žiadanej rýchlosti a rôznom zaťažení motora

V. ZÁVER

Príspevok demonštruje výborné aproximačné vlastnosti umelých neurónových sietí. Pre navrhnuté vektorové riadenie bol navrhnutý neurónový estimátor uhlovej rýchlosti. Estimátor tvorí dopredná neurónová sieť s jednou skrytou vrstvou a jeho vstupmi sú súčasné a oneskorené zložky napätia u_{1y} a prúdu i_{1y} . Simulačné výsledky potvrdzujú dobrú presnosť odhadu uhlovej rýchlosti motora pri malých aj vysokých rýchlostiach a pri rôznom zaťažení. Pre odhad polohy rotorového toku je vo vektorovom riadení použitý napäťový model, ktorý však pre praktické aplikácie vyhovuje iba oblastiam stredných a vysokých rýchlostí. Preto sa ďalší výskum sústreďí aj na nahradenie napäťového modelu iným typom pozorovateľa rotorového toku.

POĎAKOVANIE

Úloha bola riešená v rámci grantu VEGA č. 1/0006/10.

LITERATÚRA

- [1] F. Blaschke, "The principle of field orientation as applied to the new transvector closed loop control system in a PWM inverter induction motor drive," Siemens Rev., vol. 39, no. 5, 1972, pp. 217-220.
- [2] J. Vittek, "Vybrané metody riadenia elektrických pohonov v prostredí MatLab-Simulink", Trenčín, ACB, 2004, ISBN: 80-8075-039-4.
- [3] L. Zboray, F. Ďurovský, J. Tomko, "Regulované pohony", Košice, Viena, 2000, ISBN: 80-88922-13-5.
- [4] P. Girovský, "Využitie umelých neurónových sietí vo vektorovom riadení asynchrónneho motora", Košice, 2009, Doktorandská dizertačná práca.
- [5] J. Žilková, J. Timko, P. Girovský, "Bezsúťažové vektorové riadenie asynchrónneho motora", AT&P journal PLUS 2, 2006.
- [6] P. Vas, "Artificial-intelligence-based electrical machines and drives", OXFORD UNIVERSITY PRESS, 1998, ISBN: 0 19 859397 X.
- [7] J. Žilková, J. Timko, "On-line estimation of quantities using artificial neural networks", Acta Technica ČSAV 47, no. 3, 2002, pp. 305-315, ISSN- 0001-7043.
- [8] J. Vittek, P. Bris, M. Štulrajter, M. Pácha, "Chattering free sliding mode control law for position control of the drive employing induction motor", Power Engineering Conference, 2008. AUPEC '08. Australasian Universities, vol., no., pp.1-6, 14-17 Dec. 2008.

Nepriamy jednosmerný menič s mäkkým spínaním so sekundárnym riadeným usmerňovačom

Marcel BODOR, Jaroslav DUDRIK

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jaroslav.dudrik@tuke.sk, marcel.bodor@tuke.sk

Abstrakt — V článku je opísaný jednosmerný nepriamy menič s mäkkým spínaním. Mäkké spínanie tranzistorov mostíkového striedača zabezpečuje riadený usmerňovač s bezstratovou odľahčovacou sieťou na sekundárnej strane výkonového transformátora. Odľahčovacia sieť znižuje aj vypínacie straty tranzistorov usmerňovača. Použitím inovatívneho riadenia sa v meniči neobjavujú okruhové prúdy.

Kľúčové slová — mäkké spínanie, riadený usmerňovač, jednosmerný menič, odľahčovacia sieť

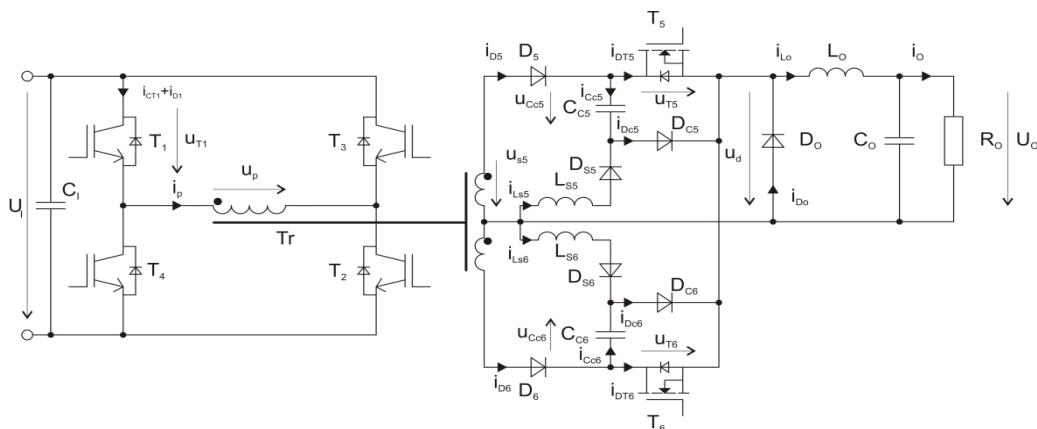
I. ÚVOD

Existuje veľa rôznych zapojení pre dosiahnutie mäkkého spínania polovodičových súčiastok v meničoch. Väčšina nepriamych jednosmerných meničov s mäkkým spínaním využíva pre riadenie výstupných veličín (napätie, prúd) striedač, neriadený usmerňovač a rôzne odľahčovacie siete či už na primárnej alebo sekundárnej strane meniča [1]-[3], [9]. Iné využívajú rezonančný princíp [4], [6], [7].

Ďalej opísaný menič využíva riadený usmerňovač, ktorý má za úlohu zabezpečiť mäkké spínanie tranzistorov striedača a riadiť výstupné veličiny meniča. Odľahčovacia sieť zabezpečí mäkké spínanie tranzistorov riadeného usmerňovača. Vďaka použitiu len reaktívnych súčiastok a diód je táto sieť bezstratová. Celá energia absorbovaná sieťou je odovzdaná na výstup meniča do záťaže [1].

II. VÝKONOVÁ ČASŤ MENIČA

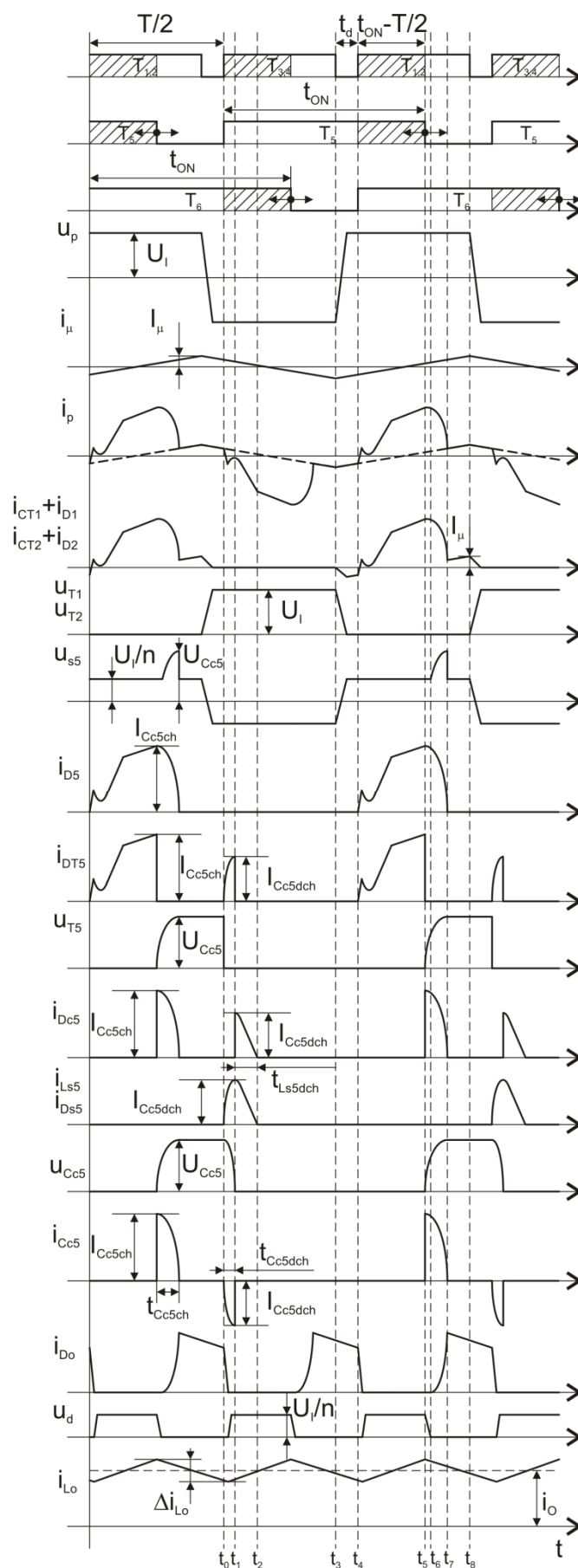
Na Obr. 1 je schéma meniča, ktorý pozostáva z mostíkového striedača (T_1 - T_4), vf transformátora s vyvedeným stredom (Tr), riadeného usmerňovača tvoreného tranzistormi (T_5 , T_6) doplnenými sériovo diódami (D_5 , D_6) pre zabezpečenie záverných vlastností tranzistorov, bezstratovej odľahčovacej siete (D_{S5} , D_{C5} , C_{C5} , L_{S5} , D_{S6} , D_{C6} , C_{C6} , L_{S6}) a výstupného filtra (L_o , C_o , D_o).



Obr. 1 Schéma výkonovej časti meniča

III. PRINCÍP ČINNOSTI MENIČA

Teoretické priebehy napätí a prúdov sú na Obr. 2. Činnosť meniča je opísaná po intervaloch.



Obr. 2 Teoretické priebehy prúdov a napätí meniča

Interval t_0-t_1 :

V čase t_0 zapne tranzistor T_5 , čo spôsobí vybíjanie kondenzátora C_{C5} do záťaže. Indukčnosť L_{S5} obmedzuje strmosť nárastu vybíjacieho prúdu a tým zabezpečí mäkké zapnutie. V čase t_0 zapnú aj tranzistory T_3 a T_4 , tranzistor T_6 je zapnutý. Prúd týchto tranzistorov je zmenšený o vybíjací prúd kondenzátora C_{C5} .

Interval t_1-t_2 :

Na začiatku tohto intervalu nastáva komutácia prúdu z kondenzátora C_{C5} na diódu D_{C5} , ktorou doznieva prúd indukčnosti L_{S5} . Jeho energia sa odovzdáva do záťaže.

Interval t_3-t_4 :

Vypnutím tranzistorov T_3 , T_4 hlavná indukčnosť transformátora L_{lh} magnetizačným prúdom vybíja výstupné kapacity C_{OSS} tranzistorov T_1 , T_2 a nabíja tranzistory T_3 , T_4 . Rozptylové indukčnosti obmedzujú strmosť nárastu prúdu. Tým je umožnené mäkké zapnutie tranzistorov T_1 , T_2 v čase t_4 .

Interval t_4-t_5 :

V čase t_4 zapnutím tranzistorov T_1 , T_2 a T_6 nastáva komutácia prúdu z nulovej diódy na tieto tranzistory a tranzistor T_5 . Prúd tranzistora T_5 je zmenšený o prúd tranzistora T_6 (vibíjací prúd kondenzátora C_{C6} a doznievajúci prúd indukčnosti L_{S6}).

Interval t_5-t_6 :

V čase t_5 sa vypne tranzistor T_5 . Prúd skokom prevezme kondenzátor C_{C5} (a dióda D_{C5}) a obmedzuje strmosť nárastu napätia. Energia rozptylových indukčností je absorbovaná jednak kondenzátorom, jednak záťažou.

Interval t_6-t_7 :

V čase t_6 usmernené napätie u_d dosiahne nulu (prúd komutuje na nulovú diódu), čo spôsobuje zmenu priebehu nabíjania kondenzátora. Energii rozptylových indukčností absorbuje už len kondenzátor. Pre napätie kondenzátora platí:

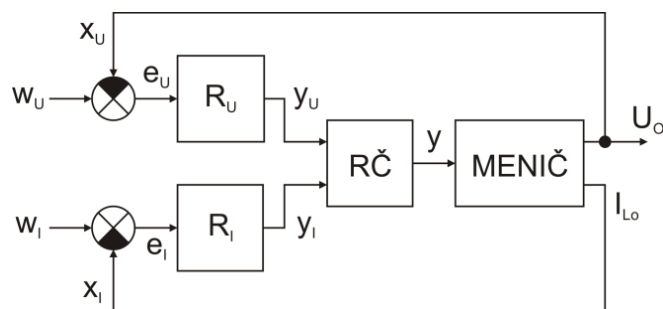
$$U_{C_{C5}} \approx \frac{U_I}{n} + \left(i_o + \frac{\Delta i_{L_o}}{2} \right) \sqrt{\frac{(L'_{1\sigma} + L_{2\sigma})}{C_{C5}}} \quad (1)$$

Interval t_7-t_8 :

Primárnou stranou tečie len magnetizačný prúd a prúd vyhladzovacej indukčnosti sa uzatvára cez záťaž a nulovú diódu.

IV. RIADIACA ČASŤ MENIČA

Riadenie meniča zabezpečuje mikroprocesor. Výstupné veličiny meniča sú riadené šírkovým riadením tranzistorov usmerňovača. Generovanie impulzov zabezpečuje paralelná štruktúra PI regulátora prúdu a PI regulátora napätia. Regulátor napätia R_U reguluje výstupné napätie na žiadanú hodnotu w_U s nulovou regulačnou odchýlkou bez zásahu regulátora prúdu R_I . Ak je prekročený maximálny prúd vyhladzovacej indukčnosti (žiadaný prúd w_I), regulátor prúdu bude znižovať akčný zásah y_I . Rozhodovací člen RČ určuje, ktorý z regulátorov má byť aktívny. Vyberá ten z regulátorov, ktorý má menší akčný zásah. Na Obr. 3 je bloková schéma zapojenia regulátorov. [8]

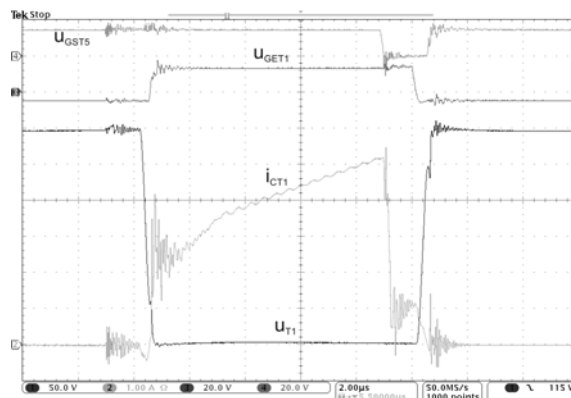


Obr. 3 Bloková schéma paralelnej štruktúry PI regulátorov

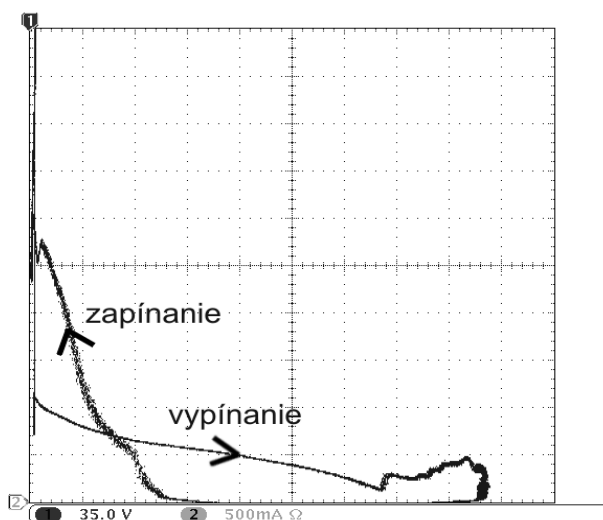
V. EXPERIMENTÁLNE VÝSLEDKY

Činnosť meniča bola overená na laboratórnom modeli s výkonom 1,2kW, výstupným napätím 40V, prúdom 30A pri napájacom napätí 300V. Spínacia frekvencia meniča je 50kHz. Priebehy napätí a prúdov sú namerané pri menovitom výstupnom napätí a prúde 25A.

Na Obr. 4 je priebeh napätia a prúdu tranzistora striedača T_1 . Po vypnutí tranzistora usmerňovača T_5 klesne prúd tranzistora na hodnotu magnetizačného prúdu transformátora. Mäkké spínanie dokazuje aj pohyb pracovného bodu tranzistora na Obr. 5. Pracovný bod tranzistora sa pohybuje v oblasti nízkych strát.

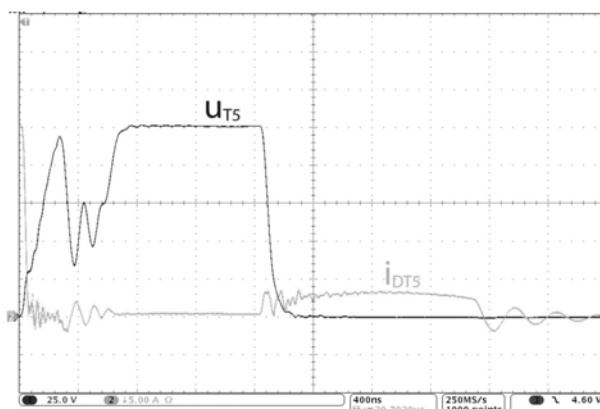


Obr. 4 Spínanie tranzistora striedača T_1

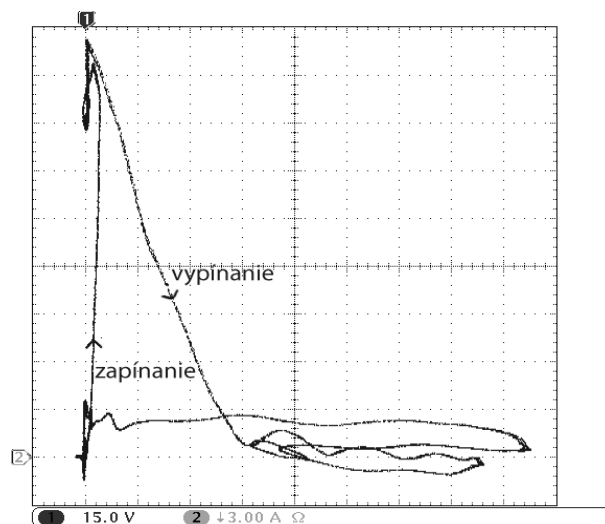


Obr. 5 Pohyb pracovného bodu tranzistora striedača T_1

Spínanie tranzistora usmerňovača je na Obr. 6. Pri vypínaní tranzistora prúd skokom preberá kondenzátor odľahčovacej siete a znižuje strmosť nárastu napätia. Po zapnutí tranzistora sa tento kondenzátor vybíja do záťaže. Pohyb pracovného bodu tranzistora riadeného usmerňovača je na Obr. 7.

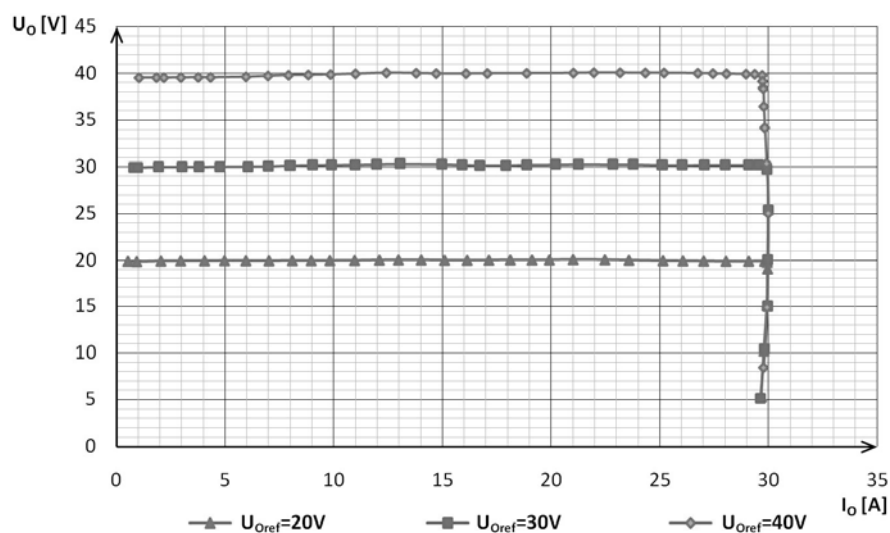


Obr. 6 Spínanie tranzistora usmerňovača T_5

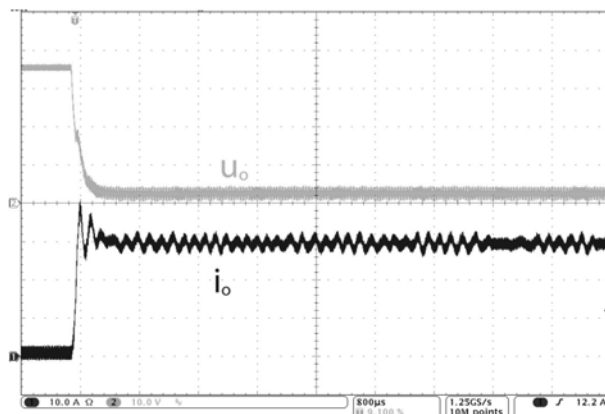


Obr. 7 Pohyb pracovného bodu tranzistora usmerňovača T_5

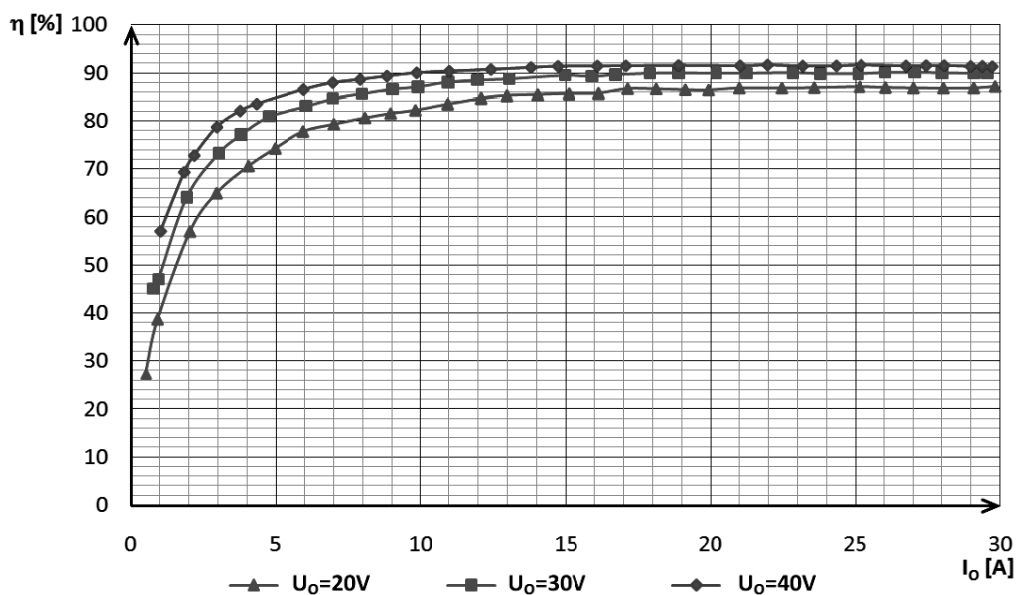
Pre overenie činnosti riadiacej časti meniča boli namerané zaťažovacie charakteristiky (Obr. 8) a priebeh výstupného napätia a prúdu pri skrate výstupu (Obr. 9). Účinnosť meniča dosahuje pri menovitom napätí 92% (Obr. 10).



Obr. 8 Zaťažovacie charakteristiky meniča



Obr. 9 Reakcia prúdového regulátora na skrat



Obr. 10 Účinnosť meniča

VI. ZÁVER

Použitie riadeného usmerňovača vynára ďalšie možnosti riešenia výkonovej časti ako aj spôsobu riadenia jednosmerného nepriameho meniča. Mäkké spínanie je dosiahnuté v celom pracovnom rozsahu meniča. Magnetizačný prúd transformátora je využívaný na nabíjanie resp. vybíjanie výstupných kapacít tranzistorov striedača, čo umožňuje ich mäkké spínanie. Navrhnutá nová odľahčovacia sieť má za úlohu absorbovať energiu rozptylových indukčností transformátora a odovzdávať ju bez strát do záťaže a zároveň umožňuje mäkké vypínanie tranzistorov usmerňovača.

POĎAKOVANIE

Práca bola realizovaná s podporou projektu Vedeckej grantovej agentúry Ministerstva školstva VEGA 1/0099/09.

LITERATÚRA

- [1] J. Dudrik, V. Ruščin, "Voltage Fed Zero-Voltage Zero-Current Switching PWM DC-DC Converter," *Power Electronics and Motion Control Conference, 2008. EPE-PEMC 2008. 13th*, vol., no., pp.295-300, 1-3 Sept. 2008.
- [2] J. Dudrik, *Výkonové vysokofrekvenčné DC-DC meniče s mäkkým spínaním*, Košice: Elfa, 2007.
- [3] J. W. Baek, J. G. Cho, D. W. Yoo, G. H. Rim, H. G. Kim, "An improved zero voltage and zero current switching full bridge PWM converter with secondary active clamp," *Power Electronics Specialists Conference, 1998. PESC 98 Record. 29th Annual IEEE*, vol.2, no., pp.948-954 vol.2, 17-22 May 1998.
- [4] S. Petrov, *Perspektívy primenenia resonansnych probrazovatelej v kačestve istočnikov svaročnovo toka*. Schemotechnika, No. 7., 2006, pp. 30-33.
- [5] P. Chlebiš, *Polovodičové měniče s měkkým spínaním*. Ediční středisko VŠB Ostrava, 2004.
- [6] K. W. Seok and B. H. Kwon, "An improved zero-voltage and zero-current-switching full-bridge PWM converter using a simple resonant circuit," *Industrial Electronics, IEEE Transactions on*, vol. 48, no. 6, pp. 1205–1209, Dec. 2001.
- [7] Muhammad H. Rashid, *Power Electronics Handbook*, Academic Press, 2001.
- [8] L. Zboray, F. Ďurovský, J. Tomko, *Regulované pohony*, Košice: Vienaľa, 2000.
- [9] J. Dudrik, J. Oetter, "Soft-Switching PWM DC-DC Converter for High Power Applications," *Power Electronics and Motion Control Conference, 2006. EPE-PEMC 2006. 12th International*, Portorož, Slovenia, CD, pp. 739-744.
- [10] J. Dudrik, P. Špánik, D. N. Trip, "Zero Voltage and Zero Current Switching Full-Bridge DC-DC Converter with Auxiliary Transformer," *IEEE Trans. on Power Electronics*, Vol. 21, No.5, 2006, pp. 1328 – 1335.

Obojsmerný menič pre hybridnú batériu v elektromobiloch

Tomáš BÉREŠ, Jaroslav DUDRIK, Erik EÖTVÖS

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

tomas.beres@tuke.sk, jaroslav.dudrik@tuke.sk, erik.eotvos@tuke.sk

Abstrakt — V tomto článku je opísaný koncept hybridnej batérie pre elektromobily, ktorý využíva pre svoju činnosť obojsmerný DC/DC menič. Menič je riadený pomocou štruktúry znižujúcej spínacie straty na tranzistoroch. Konceptu hybridnej batérie sa venuje úvod práce, v ďalších častiach tejto práce je opísané riadenie ako aj výkonová časť meniča. Experimentálnym výsledkom v závere práce bola potvrdená správna činnosť riadiacej časti meniča.

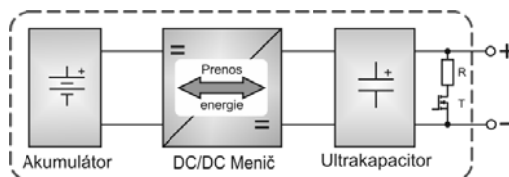
Kľúčové slová — hybridná batéria, obojsmerný DC/DC menič, šírko-impulzová modulácia(PWM)

I. ÚVOD

Snaha zlepšiť vlastnosti akumulátorov a priblížiť sa vlastnostiam ideálneho napájacieho zdroja prináša rôzne spôsoby riešenia tohto problému. Hlavnými nedostatkami akumulátorov z hľadiska elektromobilov je nízka dynamika vstupného resp. výstupného výkonu a nízke nabíjacie prúdy. Jedným z riešení ako odstrániť tieto nedostatky, je použitie paralelného zapojenia akumulátora a ultrakapacitora prostredníctvom obojsmerného DC/DC meniča. Zapojenie týchto elementov tvorí nový druh napájacieho zdroja – hybridnú batériu (HB), ktorej úlohou je zvýšiť dynamiku napájacej sústavy elektromobilu [2], [5], [8].

II. HYBRIDNÁ BATÉRIA

Hybridná batéria je sekundárny napäťový zdroj s vylepšenými parametrami, ako je dynamika výkonu a životnosť, oproti klasickým akumulátorom. Bloková schéma hybridnej batérie pozostávajúcej z akumulátora, obojsmerného DC/DC meniča, ultrakapacitora a obvodu na premenu nadbytočnej energie pri brzdení elektromobilu na teplo je znázornená na obr. 1 [8].



Obr. 1 Bloková schéma hybridnej batérie

Vlastnosti akumulátora ovplyvňujú celkovú kapacitu a životnosť hybridnej batérie. Vysokú dynamiku vstupného/výstupného výkonu zabezpečuje ultrakapacitor. Obojsmerný DC/DC menič je hlavnou časťou hybridnej batérie. Pri nabíjaní HB zabezpečuje prenos energie z ultrakapacitora do akumulátora, pri vybíjaní HB udržiava konštantné napätie batérie a teda energiu prenáša z akumulátora do ultrakapacitora. Batéria obsahuje tiež obvod na pozostávajúci z rezistora R a polovodičového spínača T , ktorý v prípade potreby zabráni napätiu batérie prekročiť zvolenú maximálnu hodnotu pri rekuperácii energie počas brzdenia elektromobilu [5].

III. NÁVRH MENIČA PRE HYBRIDNÚ BATÉRIU

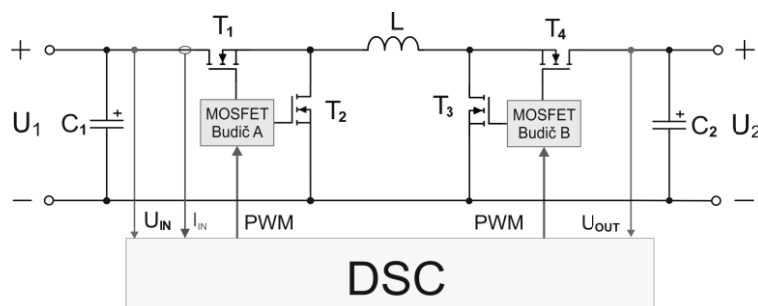
Obojsmerný DC/DC menič v hybridnej batérii plní funkciu medzičlánku medzi akumulátorom a ultrakapacitorom, a jeho hlavnou úlohou je udržiavať napätie na svorkách hybridnej batérie v určitom požadovanom rozmedzí. Jeho ďalšou úlohou je šetriť akumulátor

tak, že pomocou obojsmerného DC/DC meniča sa obmedzuje nabíjaci a vybíjaci prúd akumulátora a tým sa predlžuje jeho životnosť.

Menič použitý na tento účel musí spĺňať nasledujúce požiadavky ako sú:

- menič musí byť schopný znižovať a zvyšovať napätie v oboch smeroch prevádzky
- veľké prúdové zaťaženie
- jednoduchosť zapojenia
- vysoká účinnosť

Najvhodnejším zapojením, ktoré spĺňa tieto podmienky je neinvertujúci kaskádový znižovaco-zvyšovací menič. Jeho bloková schéma je znázornená na obr.2. [1], [7].



Obr. 2 Schéma obojsmerného DC/DC meniča pre HB

Výkonová časť meniča pozostáva zo štyroch výkonových tranzistorov typu MOSFET zapojených do dvoch polomostíkov, ktoré sú vzájomne prepojené induktorom[2]. Na vstupnej aj na výstupnej strane sú použité kondenzátory, ktoré vyhladzujú výstupné napätie. Riadenie meniča bude realizované pomocou mikrokontroléra (DSC).

Veľkosť výstupného napätia meniča je vo všeobecnosti daná vzťahom :

$$V_{OUT} = V_{IN} \frac{D_1}{D_2} \quad (1)$$

kde

$$D_1 = \frac{t_{T1(ON)}}{T} \quad D_2 = \frac{D_{T4(ON)}}{T} \quad (2)$$

$t_{T1(on)}$, $t_{T4(on)}$ vyjadrujú čas zapnutia tranzistora T_1 resp. T_4 počas periódy T .

Tento menič môže pracovať v troch základných režimoch :

- a) Znižovací režim ($V_{IN} > V_{OUT}$)

V tomto režime je tranzistor T_4 zapnutý, T_3 vypnutý. Tranzistory T_1 a T_2 spínajú v protifáze. Menič pracuje ako znižovací impulzový menič. Veľkosť výstupného napätia meniča je daná vzťahom:

$$V_{OUT} = V_{IN} \cdot D_1 \quad (3)$$

- b) Znižovaco-zvyšovací režim ($V_{IN} \approx V_{OUT}$)

Ak je veľkosť vstupného a výstupného napätia približne rovnaká, menič pracuje v znižovaco-zvyšovacom režime, tj. spínajú všetky štyri tranzistory $T_1 - T_4$. Veľkosť výstupného napätia meniča je vyjadrená vzťahom :

$$V_{OUT} = \frac{D_1}{D_2} \quad (4)$$

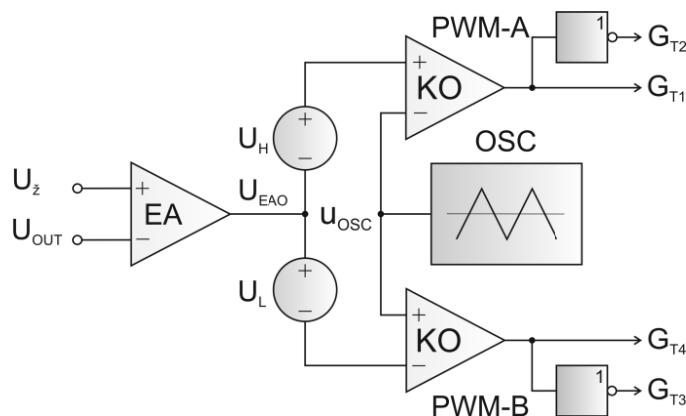
- c) Zvyšovací režim ($V_{IN} < V_{OUT}$)

V tomto prevádzkovom režime je tranzistor T_1 je zapnutý, T_2 je vypnutý a tranzistory T_3 a T_4 tvoria zvyšovací impulzový menič. Výstupné napätie je dané vzťahom :

$$U_{OUT} = \frac{U_{IN}}{D_2} \quad (5)$$

IV. RIADENIE MENIČA

Pre riadenie obojsmerného DC/DC meniča je použitá riadiaca štruktúra zobrazená na obr.3. Použité riadenie umožňuje meniču pracovať v troch pracovných režimoch a to v znižovacom, znižovaco – zvyšovacom a zvyšovacom režime. V znižovacom a zvyšovacom režime spínajú pri navrhnutom riadení iba dva tranzistory a tým sa dosiahne zníženie spínacích strát tranzistorov, čo má za následok zvýšenie účinnosti meniča [3],[6],[9].



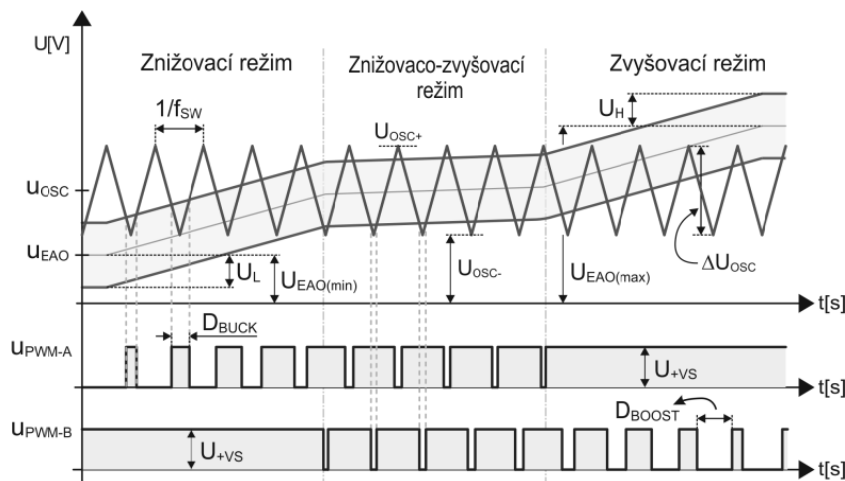
Obr. 3 Bloková schéma riadiacej časti meniča

Menič pracuje v jednotlivých režimoch prevádzky, v závislosti od veľkosti referenčných napätí U_H a U_L nasledovne:

- Ak $U_{EAO} - U_L < U_{OSC-}$ a $U_{EAO} + U_H < U_{OSC+}$ menič pracuje v znižovacom režime
- Ak $U_{EAO} - U_L > U_{OSC-}$ a $U_{EAO} + U_H < U_{OSC+}$ menič pracuje v znižovaco-zvyšovacom režime
- Ak $U_{EAO} - U_L > U_{OSC-}$ a $U_{EAO} + U_H > U_{OSC+}$ menič pracuje vo zvyšovacom režime

kde U_{EAO} je výstup z napäťového PI regulátora (regulačná odchýlka) a U_{OSC+} , U_{OSC-} je veľkosť amplitúdy oscilátora.

Pre lepšie pochopenie je funkcia riadiaceho obvodu zobrazená na obr. 4.

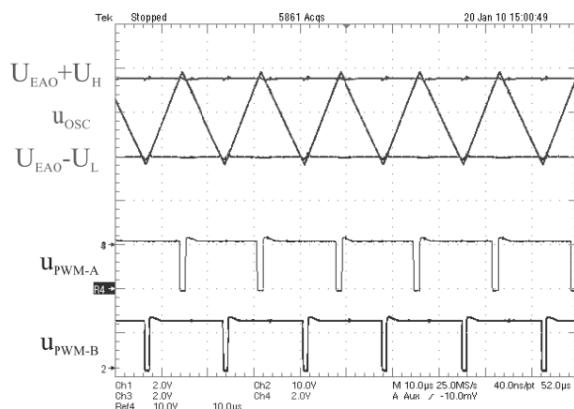


Obr. 4 Funkcia riadiaceho obvodu meniča

Napätie u_{PWM-A} je vstupným signálom pre budič tranzistorov T_1 a T_2 , u_{PWM-B} pre budič tranzistorov T_3 a T_4 .

V. EXPERIMENTÁLNE VÝSLEDKY

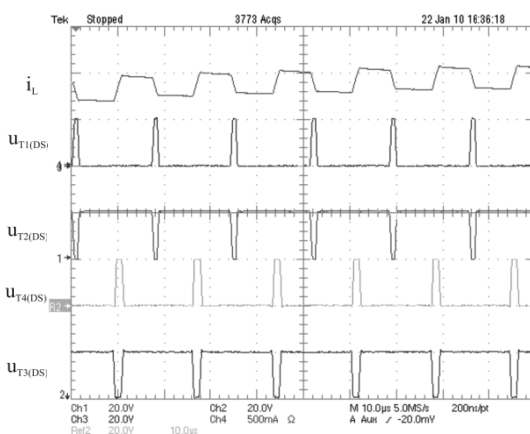
Funkcia opisovaného meniča bola overená použitím fyzikálneho modelu. Bolo použité riadenie podľa obr.3. Na obr.5 je znázornená činnosť riadiaceho obvodu v znižovaco-zvyšovacom režime pri $U_{IN} = 25V$, $U_{OUT} = 24V$.



Obr. 5 Tvorba PWM pre budiče tranzistorov v znižovaco - zvyšovacom režime

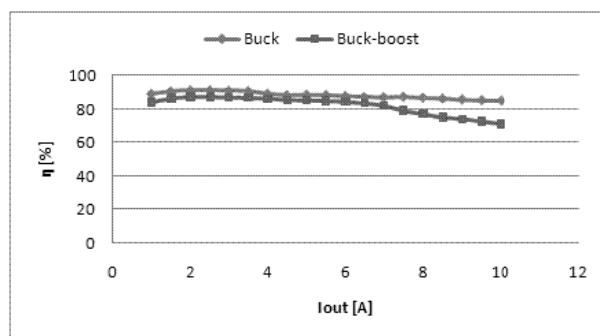
Priebehy na obr. 5 zobrazujú tvorbu PWM signálov pre budiče tranzistorov. Tieto priebehy sa zhodujú s teoretickým predpokladom, ktorý je zobrazený na obr. 4.

Na nasledujúcom oscilogramе sú zobrazené priebehy napätí na jednotlivých tranzistoroch ako aj priebeh prúdu induktorom L.



Obr. 6 Priebehy napätí na tranzistoroch T1 - T4 a prúdu induktorom L

Účinnosť meniča je výrazne ovplyvnená počtom spínaných tranzistorov v meniči. Preto, ak je to možné vhodným nastavením referenčných napätí U_H a U_L prevádzkujeme menič v znižovacom alebo zvyšovacom režime, kde spínajú iba dva zo štvorice tranzistorov. Porovnanie účinnosti meniča v znižovacom a znižovaco-zvyšovacom režime pri $U_{IN} = 20V$, $U_{OUT} = 14V$ je znázornené na obr. 7.



Obr. 7 Porovnanie účinnosti meniča v znižovacom a znižovaco – zvyšovacom režime

Z porovnania účinnosti meniča je vidieť, že menič má v celom rozsahu zaťažovania vyššiu účinnosť v znižovacom oproti účinnosti v znižovaco-zvyšovacom režime, keďže pri spínaní dvoch tranzistorov vznikajú menšie spínacie straty ako pri spínaní štyroch tranzistorov.

VI. ZÁVER

V tomto článku je opisovaný obojsmerný DC/DC menič pre hybridnú batériu v elektromobiloch. Činnosť výkonovej a riadiacej časti meniča bola overená pomocou fyzikálneho modelu meniča. Meraním bolo dokázané, že použitím novej riadiacej štruktúry umožňujúcej činnosť meniča v troch úsporných režimoch bola dosiahnutá vyššia účinnosť meniča. Prechod medzi jednotlivými režimami činnosti je plynulý, daný len veľkosťou referenčných signálov.

POĎAKOVANIE

Práca bola realizovaná s podporou projektu Vedeckej grantovej agentúry Ministerstva školstva VEGA 1/0099/09.

LITERATÚRA

- [1] J. Hamar, B. Buti, I. Nagy, "Dual Channel Resonant DC-DC Converter Family". In: *EPE Journal*, Volume 17, Issue 3, Jul-Sep. 2007, pp.5-15.
- [2] J. Leuchter, P. Bauer, P. Bojda, V. Řeřucha, "Bi-directional DC-DC Converters for Supercapacitor Based Energy Buffer For Electrical GEN-SETS". In: *Proceedings of the 12th European Conference on Power Electronics and Applications (EPE 2007)*, Aalborg: >> EPE Press, 2007, p. P.1 - P.10.
- [3] P. Špánik, B. Dobrucký, M. Frivaldský, P. Drgoňa, "Experimental analysis of commutation process of power semiconductor transistor structures". In: *Acta Technica CSAV*, Vol. 52, No. 4., 2007.
- [4] M. Gaboriault, A. Notman, "A High Efficiency, Noninverting, Buck-Boost DC-DC Converter". In: *APEC '04. Nineteenth Annual IEEE*, Volume 3, 2004.
- [5] T. Markel, M. Zolot, S. Sprik, "Ultracapacitors and Batteries in Hybrid Vehicles". In: *National Renewable Energy Laboratory (NREL) PR-540-38484*, 8.2005, pp.7-15.
- [6] J. Dudrik, *High Frequency Soft Switching DC-DC Power Converters*. Monograph, ELFA, Košice 2007.
- [7] J. Hamar, I. Nagy, P. Stumpf, H. Ohsaki, E. Masada "New Dual Channel Quasi Resonant DC-DC Converter Topologies for Distributed Energy Utilization". In: *13th International Power Electronics and Motion Control Conference (EPE-PEMC 2008)*, pp. 1778-1785.
- [8] J. Leuchter, P. Bauer and Ahmed F. Zobaa, "Power Electronics and Energy Management of Hybrid Power Sources with Supercapacitors". In: *Proceedings of the 5th IET International Conference on Power Electronics, Machines and drives (PEMD 2010)*, Brighton: 2010, pp. 6.
- [9] N. D. Trip, "Analysis and Experimental Results of an Active Snubber for Boost Converters". In: *Proc.of the 6th Int.Conf. on Renewable sources and Enviromental Electro-Technologies - RSEE 06*, Oradea, pp. 125-128.

Jednosmerný rezonančný menič pre obnoviteľné zdroje

Erik EÖTVÖS, Jaroslav DUDRIK, Tomáš BÉREŠ

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

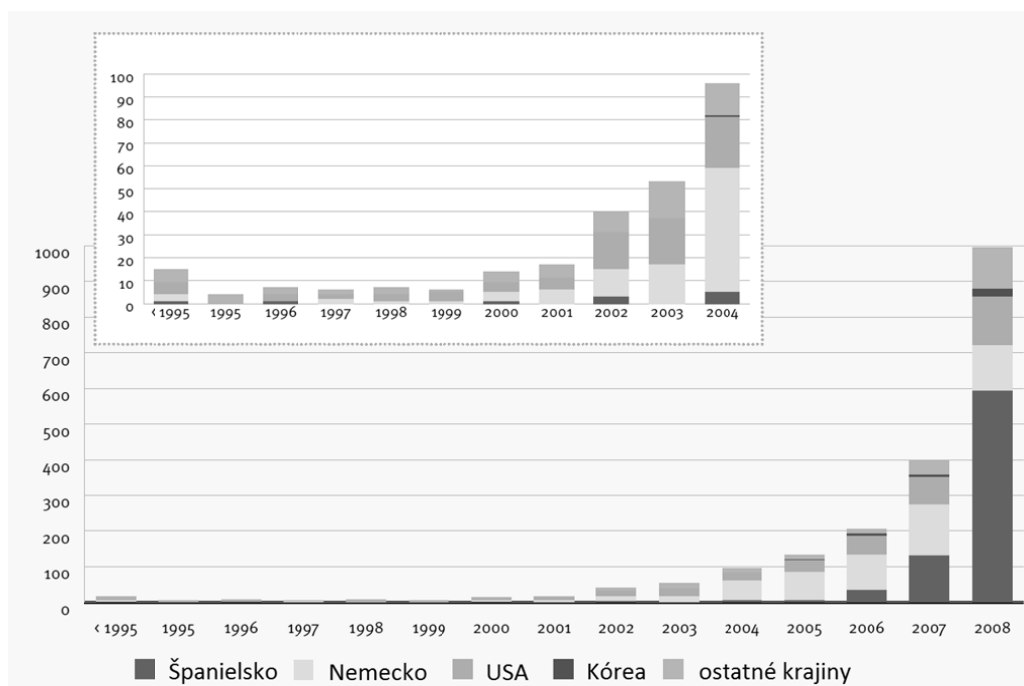
erik.eotvos@tuke.sk, jaroslav.dudrik@tuke.sk, tomas.beres@tuke.sk

Abstrakt — Táto práca popisuje jednosmerný menič, ktorého úlohou je zvyšovať nízke napätie získané z fotovoltického panela (FV) na úroveň, ktorá je potrebná na správnu činnosť pripojeného striedača. Menič zabezpečuje galvanické oddelenie FV panelov od elektrickej rozvodnej siete. Použitá architektúra meniča je typu LLC. Polovodičové prvky meniča spínajú v nule napätia (zero-voltage switching) ZVS a v nule prúdu (zero-current switching) ZCS. Výhodou použitej architektúry je jednoduchosť riadenia a vysoká reálne dosahovaná účinnosť, ktorá sa pohybuje na úrovni 95%.

Kľúčové slová — FV panel, LLC rezonančný menič, ZVS, ZCS

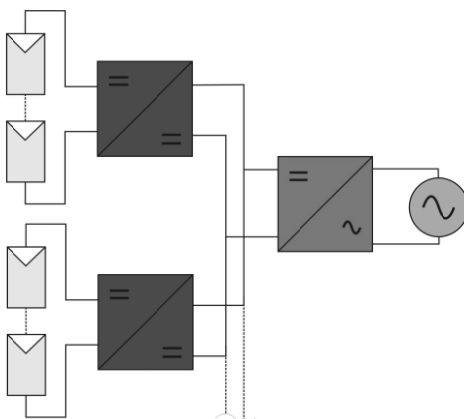
I. ÚVOD

V súčasnosti je spotreba fosílnych palív na maximálnej možnej úrovni, a keďže nové náleziská ropy, či zemného plynu sa objavujú už len sporadicky, je dôležité uvažovať nad náhradou týchto zdrojov energie. Jednou z možností je využitie obnoviteľných zdrojov. V súčasnosti sa najviac využíva vodná a veterná energia. V poslednom období sa čoraz viac využíva aj slnečná energia, a to priamou premenou slnečného žiarenia na elektrickú energiu pomocou FV panelov. Súvisí to s nárastom účinnosti FV panelov, ktorá sa dnes pohybuje na úrovni 12-15%, a v laboratórnych podmienkach dokonca až 42,4%. Na prudkom rozvoji fotovoltiky sa podpísali aj nižšie výrobné náklady pri výrobe FV článkov. V Európe pomáhajú rozvoju fotovoltiky aj rôzne štátne dotácie, napríklad v forme výrazne vyšších výkupných cien takto vyrobenej energie v porovnaní s trhovou cenou el. energie. V grafe (Obr. 1) je uvedený celkový ročný inštalovaný výkon FV panelov za roky 1995 – 2008 [1].



Obr. 1 Graf celkového inštalovaného výkonu FV panelov v rokoch 1995-2008

Na to, aby sme mohli využívať energiu získanú z FV panelov, je potrebné k panelu pripojiť vhodný menič. Typ meniča sa volí podľa spôsobu využívania FV panelov. Ak ide o zapojenie v tzv. ostrovnom systéme (off-grid), je potrebné použiť obojsmerný menič. Ten zabezpečuje dobíjanie akumulátorov v prípade ak FV panel dodáva elektrickú energiu. V čase keď panel nedodáva dostatok energie alebo v prípade vyššej spotreby pracuje menič v opačnom smere, a odoberá energiu z akumulátorov. Omnoho častejšie sa však používajú systémy pripojené na elektrickú rozvodnú sieť (on-grid). V tomto prípade sa najčastejšie používajú striedače, ktoré menia priamo jednosmerné napätie FV panelov na striedavé napätie. Ak sa použijú amorfné alebo tenkovrstvové FV panely, tak dochádza k toku kapacitného prúdu cez menič, čo môže spôsobovať problémy. Preto je potrebné FV panely galvanicky oddeliť od elektrickej siete. Jedným zo spôsobov je galvanické oddelenie pomocou transformátora. Ak by bolo galvanické oddelenie riešené na strane siete, tak by musel byť použitý sieťový transformátor, ktorý je značne nevýhodný kvôli vysokej hmotnosti a vyššej obstarávacej cene. Ako výhodnejšie riešenie sa javí použitie jednosmerného nepriameho meniča, ktorý využíva vysokofrekvenčný transformátor, ktorý je výrazne menší, ľahší a aj cenovo dostupnejší. Zaradením jednosmerného meniča pred striedač vznikne topológia, ktorá je uvedená na Obr. 2. Pri takejto architektúre je možné použiť aj viac samostatných jednosmerných meničov, každý pre jeden reťazec FV panelov. Tým sa zabezpečí aj lepšie využitie FV panelov, keďže každý reťazec môže pracovať pri maximálnom možnom výkone.

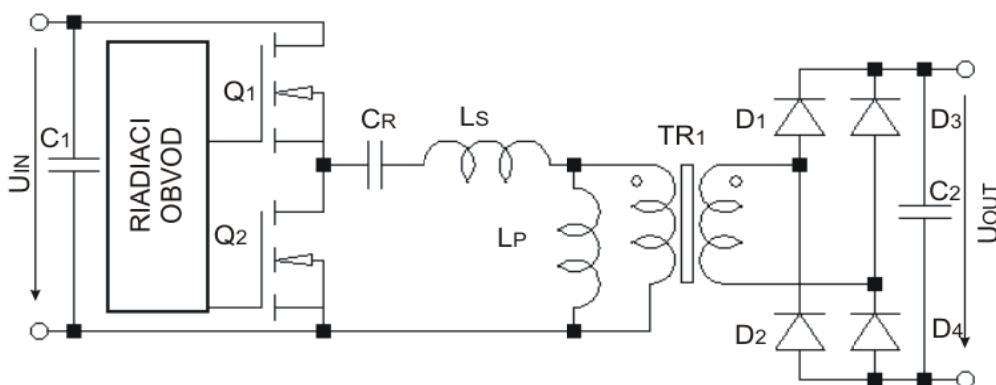


Obr. 2 Topológia zapojenia meničov pre FV systémy

Keďže v tomto zapojení sa nachádzajú dva meniče v sériovom zapojení, tak celková účinnosť systému bude súčinom účinnosti jednosmerného meniča a striedača, čiže bude nižšia. Preto je potrebné zabezpečiť čo najvyššiu účinnosť jednosmerného meniča. Ako najvýhodnejšie sa javí použitie meniča s mäkkým spínaním, ktorý dosahuje vysoké účinnosti. Do kategórie meničov s mäkkým spínaním spadajú aj meniče pracujúce na rezonančnom princípe. Takýchto meničov existuje niekoľko typov a líšia sa najmä rezonančným obvodom, ktorý môže byť sériový, sério-paralelný alebo paralelný [2]. Na základe analýzy [3] bola použitá topológia so sériovým rezonančným obvodom.

Požadované parametre meniča sú nasledovné. Vstupné napätie U_{IN} : 60-100V, výstupné napätie U_{OUT} : 400V a maximálny výstupný výkon P_{OUT} : 600W.

Menich sa skladá z polomostíkového striedača, rezonančného obvodu, ktorý je tvorený rezonančným kondenzátorom, sériovou a paralelnou cievkou Obr. 3.



Obr. 3 Schéma zapojenia obvodu rezonančného meniča LLC

Skratka LLC naznačuje, že rezonančný obvod je tvorený dvoma cievkami s indukčnosťami L_S a L_P a jedným kondenzátorom s kapacitou C_R . Nasleduje ideálny transformátor. Na sekundárnej strane sa nachádza rýchly diódový mostíkový usmerňovač a kapacitný filter.

II. PRINCÍP ČINNOSTI

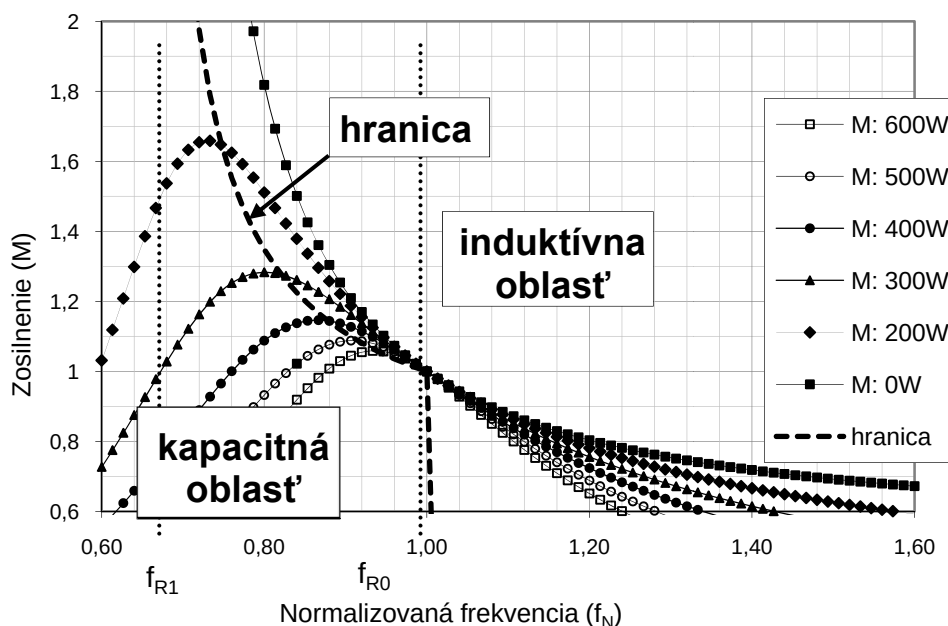
Spínače polomostíkového striedača sú spínané so striedou $z=0,5$ s premenlivou frekvenciou. Rezonančný obvod má dve rezonančné frekvencie, pričom platí $f_{R0} > f_{R1}$. Frekvencia f_{R0} je rezonančnou frekvenciou sériovej cievky L_S a rezonančného kondenzátora C_R . Nižšia rezonančná frekvencia f_{R1} závisí aj od indukčnosti paralelnej cievky L_P rezonančného obvodu. Ak menič pracuje nad frekvenciou f_{R0} , pracuje v induktívnej oblasti, pričom platí, že prúd sa oneskoruje za napätím, a tým je zabezpečené zapínanie spínačov v nule napätia. Taktiež sa dosahuje vypínanie usmerňovacích diód v nule prúdu. Pri frekvencii nižšej ako f_{R1} sa menič dostáva do kapacitnej oblasti, čo spôsobuje predbiehanie prúdu. V tomto prípade by spínače vypínali nulový prúd, avšak pri zapínaní by prúd narastal skokom, čím by vznikali vysoké straty. Preto sa pri riadení musíme vyhnúť tejto pracovnej oblasti. Oblasť medzi frekvenciami f_{R0} a f_{R1} je rozdelená na kapacitnú a induktívnu oblasť, pričom platí že pri zvýšení zaťaženia meniča, musíme spínaciu frekvenciu znižovať a naopak. Preto je potrebné navrhnuť rezonančný obvod tak aby sa menič nemohol dostať do kapacitnej oblasti. Na Obr. 4 sú uvedené krivky zosilnenia rezonančného obvodu pri rôznych zaťaženiach, kde je zobrazená kapacitná a induktívna oblasť rozdelená čiarkovanou čiarou. Tieto krivky boli vypočítané pomocou rovnice (1) pre rôzne zaťaženie rezonančného obvodu. λ je pomer indukčností sériovej L_S a paralelnej L_P cievky (2). Normalizovaná frekvencia f_N je pomer spínacej f_{SW} a rezonančnej f_{R0} frekvencie (3). Faktor kvality rezonančného obvodu sa vypočíta podľa (4) a závisí od impedancie rezonančného obvodu Z_0 , záťaže meniča P_{OUT} , transformačného pomeru n a výstupného napätia V_{OUT} .

$$M = 1 / \sqrt{\left(1 + \lambda - \frac{\lambda}{f_N^2}\right) + Q^2 \left(f_N - \frac{1}{f_N}\right)^2} \quad (1)$$

$$\lambda = L_S / L_P \quad (2)$$

$$f_N = f_{SW} / f_{R0} \quad (3)$$

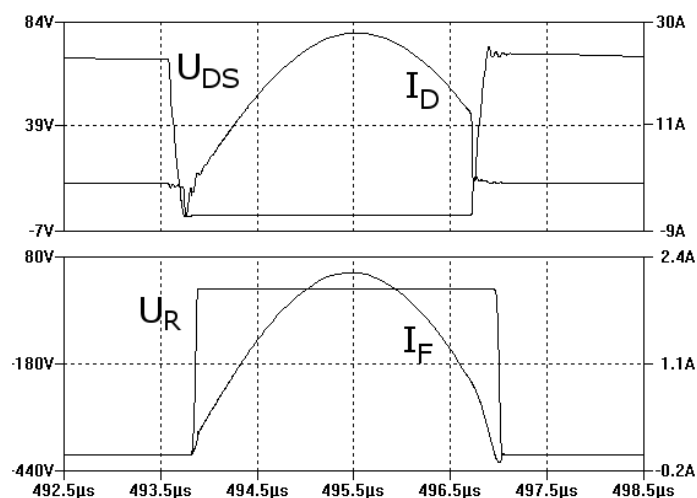
$$Q = (\pi^2 Z_0 P_{OUT}) / 8 n^2 V_{OUT}^2 \quad (4)$$



Obr. 4 Charakteristické zosilnenie rezonančného obvodu meniča

III. SIMULÁCIA

Činnosť meniča bola overená v programe LTSpice IV. Vstupné napätie bolo stanovené na 70V. Výstupný výkon bol 600W a pracovná frekvencia bola pri týchto podmienkach 150kHz. Na Obr. 5 sú zobrazené simulované priebehy meniča.



Obr. 5 Priebehy simulácie napätí a prúdov MOSFET spínača a diódy usmerňovacieho mostíka

V hornej časti je zobrazený priebeh napätia U_{DS} a prúdu I_D spínača. Tu môžeme pozorovať, že ešte pred zapnutím spínača poklesne napätie na nulovú hodnotu, čiže spínač je zapnutý bez strát (ZVS). Pri vypínaní spínača vznikajú minimálne straty, keďže nárast napätia spínača je spomaľovaný nabíjaním jeho výstupnej kapacity. V simulácii boli zahrnuté aj parazitné prvky ako indukčnosť prívodov. To sa prejavilo kmitmi, ktoré sa v priebehoch objavujú. V spodnej časti obrázka sú priebehy napätia U_R a prúdu I_F diódy usmerňovača. Priebeh prúdu diódou má podobný tvar ako prúd spínačov na primárnej strane. Preto pri začiatku vedenia prúdu diódou vznikajú len minimálne straty a dokonca pri vypínaní poklesne prúd na nulovú hodnotu, čím je zabezpečené vypínanie v nule prúdu (ZCS).

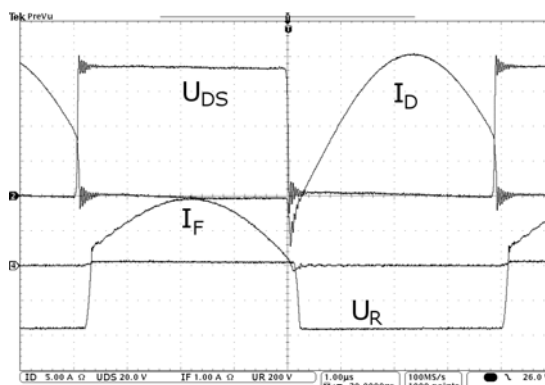
IV. LABORÁTORNY MODEL MENIČA

Na základe výpočtov a simulácií bol postavený laboratórny model meniča. Parametre modelu sú: $U_{IN}=60-100V$, $U_{OUT}=400V$, $P_{OUT}=600W$, $f_{sw}=150-400kHz$, $L_s=2,5\mu H$, $L_p=25\mu H$ a $C_R=540nF$.

Keďže menič pracuje s vysokou spínacou frekvenciou, a s pomerne vysokými hodnotami prúdov na strane striedača, tak bolo potrebné tomu prispôsobiť väčšinu súčiastok. Na polovodičové súčiastky bol kladený dôraz, aby mali čo najmenšie straty v zapnutom stave a čo najvyššie rýchlosti spínania. Preto v prípade striedača boli použité spínače MOSFET, ktoré sa vyznačujú vysokou rýchlosťou spínania. Výhodou MOSFET tranzistorov je ich nízky odpor v zapnutom stave, tým pádom aj nízky úbytok napätia na ňom. Na druhej strane, MOSFETy sa vyznačujú veľkými výstupnými kapacitami, ktoré je nutné vybíjať pri každom zapnutí spínača, čo zvyšuje straty pri spínaní. Avšak pri tomto type meniča sú výstupné kapacity prirodzene vybíjané pred každým zapnutím spínača. V mostíkovom usmerňovači boli použité rýchle diódy D_1-D_4 s nízkym úbytkom napätia. Vstupný filtračný kondenzátor C_1 a najmä rezonančný kondenzátor C_R musia zvládnuť vysoké prúdové zaťaženie pri vysokej frekvencii. Preto boli použité kvalitné bezindukčné kondenzátory typu KPI. Na transformátor bola kladená požiadavka, aby v sebe integroval sériovú aj paralelnú indukčnosť rezonančného obvodu. Správnym návrhom transformátora bolo docieľené, že hlavná indukčnosť predstavovala paralelnú, a rozptyľová sériovú indukčnosť rezonančného obvodu. Vďaka tomu, že menič pracuje v celom pracovnom rozsahu pri každom cykle bez potreby kontroly napätia a prúdu spínačov, tak na riadenie mohol byť použitý 8-bitový mikrokontrolér. Ten zabezpečuje generovanie riadiacich impulzov pre oba spínače s premenlivou frekvenciou v rozsahu 150-400kHz.

V. EXPERIMENTÁLNE VÝSLEDKY

Menič bol napájaný zdrojom jednosmerného napätia s hodnotou 72V a bol zaťažovaný výkonovými rezistormi v rozsahu od 10% do 100% nominálneho výkonu.



Obr. 6 Namerané priebehy napätí a prúdov MOSFET spínača striedača a diódy usmerňovacieho mostíka

Na Obr. 6 je reálny priebeh napätia U_{DS} a prúdu I_D spínača Q_1 a napätie U_R a prúd I_F usmerňovacej diódy D_2 . Vidíme, že predtým než spínač Q_1 začne viesť prúd, poklesne na ňom napätie na nulovú hodnotu. Tým je zabezpečená podmienka spínania v nule napätia, ZVS. Pri vypínaní spínača vznikajú straty, ktoré sú znížené spomaleným nárastom napätia na spínači, ale aj nižším prúdom ktorý je vypínaný. V prípade usmerňovacej diódy je situácia obdobná, pričom pri začiatku vedenia diódy vznikajú minimálne straty, avšak vypínanie je prakticky bezstratové, keďže napätie na dióde narastá až po poklese prúdu na nulovú hodnotu, ZCS. Nasledovne bola odmeraná účinnosť meniča v rozsahu od 10% do 100%. Výsledky sú uvedené v tabuľke Tabuľka I.

Tabuľka I
 VYJADRENIE ÚČINNOSTI V ZÁVISLOSTI OD ZAŤAŽENIA MENIČA

Zaťaženie (%):	10	16	33	55	66	83	100
Účinnosť (%):	0,89	0,92	0,946	0,955	0,953	0,949	0,947

VI. ZÁVER

Opisovaný rezonančný menič sa vyznačuje mnohými výhodami ako sú kompaktné rozmery, jednoduchá konštrukcia, jednoduchosť riadenia a najmä reálne dosiahnutá vysoká účinnosť v širokom pracovnom rozsahu. To predurčuje použitie tohto typu meniča na aplikácie ako sú práve fotovoltické systémy.

POĎAKOVANIE

Práca bola realizovaná s podporou projektu Vedeckej grantovej agentúry Ministerstva školstva VEGA 1/0099/09.

LITERATÚRA

- [1] D. Lenardič, "Large-Scale Photovoltaic Power Plants", 2008
- [2] M. Mao, D. Tchobanov, D. Li, M. Maerz, T. Gerber, G. Deboy, L. Lorenz, "Analysis and Design of a 1MHz LLC Resonant Converter with Coreless Transformer Driver"
- [3] Y. Liu, "High Efficiency Optimization of LLC Resonant Converter for Wide Load Range", Blacksburg, Virginia 2007
- [4] M. Frivaldský, P. Drgoňa, A. Prikopová: "Design and modeling of 200kHz 1,5kW LLC power semiconductor resonant converter". In: *Applied electronics 2009*, 9 – 10 September 2009, Pilsen, Czech Republic
- [5] F. K. Wong, B. Eng, M. Phil, "High Frequency Transformer for Switching Mode Power Supplies", Griffith University, Brisbane, Australia, March 2004
- [6] S. Petrov: "Perspektivy primenenia rezonansnych probrazovatelej v kačestve istočnikov svaročnovo toka", *Schemotechnika*, No. 7., 2006, pp. 30-33.
- [7] J. Hamar, B. Buti, I. Nagy: "Dual Channel Resonant DC-DC Converter Family", *EPE Journal*, Volume 17, Issue 3, Jul-Sep. 2007, pp.5-15.
- [8] J. Leuchter, P. Bauer, P. Bojda, V. Řeřucha: "Bi-directional DC-DC Converters for Supercapacitor Based Energy Buffer For Electrical" GEN-SETS. In *Proceedings of the 12th European Conference on Power Electronics and Applications* (EPE 2007), Aalborg: >> EPE Press, 2007, p. P.1 - P.10. ISBN 9789075815108 and IEEE Catalog Number 07EX1656C.
- [9] P. Špánik, B. Dobrucký, M. Frivaldský, P. Drgoňa: "Experimental analysis of commutation process of power semiconductor transistor structures", *Acta Technica CSAV*, Vol. 52, No. 4., 2007, ISSN 0001-7043
- [10] J. Hamar, K. R. Jordan, I. Nagy, H. Ohsaki: "Virtual Laboratory for Combined Solar Energy System. EPE 2007, Aalborg, Denmark, 2-5 September, ISBN: 9789075815108, pp. p.1-p.10 – CD.
- [11] N. D. Trip: "A new active snubber for dc-dc boost converters", 8th International Conference on Engineering of Modern Electric System, Proceedings, EMES'05, Section Electronical Engineering, Oradea, Romania, May 2005, pp. 124-127.

Riadenie viacúrovňového meniča pre fotovoltické aplikácie pomocou programu RT-LAB

Marek PÁSTOR, Jaroslav DUDRIK

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

marek.pastor@tuke.sk, jaroslav.dudrik@tuke.sk

Abstrakt — Článok sa zaoberá riešením riadenia viacúrovňového kaskádového striedača pomocou programu RT-LAB. Stručne opisuje možnosti simulácie v reálnom čase s využitím Hardware In the Loop (HIL) simulácie. Súčasťou návrhu je praktické overenie činnosti regulátora pre kaskádový napäťový striedač s prúdovým riadením. Regulácia využíva predikovanie prúdu záťaže na základe známych parametrov záťaže.

Kľúčové slová — fotovoltika, RCP (Rapid Control Prototyping), RT-LAB, viacúrovňový striedač

I. ÚVOD

Výkonové polovodičové meniče pre oblasť fotovoltiky (FV meniče) predstavujú v súčasnosti z pohľadu riadenia vysoko komplexné systémy. V mnohých prípadoch sú riešené kaskádovým zapojením meničov zabezpečujúcich čiastkové premeny energie, vyžadujú sledovanie maxima výkonu fotovoltického (FV) generátora (MPPT – Maximum Power Point Tracking), spoluprácu s nadradeným systémom a pod. Návrh regulátora pre takýto FV menič, prípade s využitím moderných spôsobov riadenia ako napr. fuzzy riadenie, vyžaduje komplexný návrh a testovanie. Pre uľahčenie návrhu FV meniča je možné použiť rôzne simulačné programy. Nevýhodou čistej simulácie je nedokonalé zohľadnenie vplyvov, ktoré budú v budúcnosti pôsobiť v systéme navrhovaného FV meniča. Možnosťou ako odstrániť tento nedostatok je využitie simulácie, pri ktorej je časť simulovaného systému nahradená reálnym modelom.

II. HARDWARE IN THE LOOP

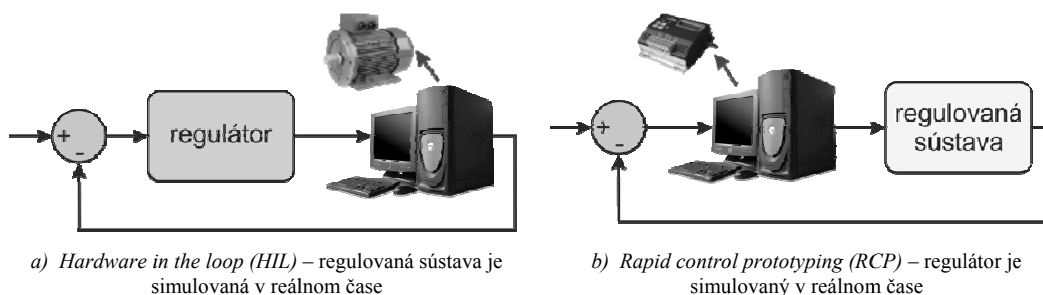
A. Princíp HIL a RCP

Možnosť simulácie matematických modelov v reálnom čase umožňuje nahradenie časti regulačnej slučky reálnym systémom. Vo všeobecnosti sú možné dva prístupy.

Hardware In the Loop (HIL) simulácia sa odlišuje od čisto programovej simulácie v reálnom čase pridaním reálneho modelu do regulačnej slučky. Každý regulovaný systém pozostáva na najvyššej úrovni z regulátora a regulovanej sústavy. Pri čisto programovej simulácii v reálnom čase sú ako regulátor, tak regulovaná sústava, reprezentované pomocou matematických modelov s istou presnosťou a komplexnosťou. HIL simulácia v reálnom čase umožňuje nahradenie regulátora alebo regulovanej sústavy reálnym modelom, čím sa výrazne zvýši presnosť simulácie.

Nahradenie matematického modelu regulátora reálnym regulátorom sa označuje ako HIL simulácia. Matematický model regulovanej sústavy je simulovaný v reálnom čase, čo umožňuje testovanie regulátora bez hrozby poškodenia reálnej regulovanej sústavy v prípade nesprávne pracujúceho regulátora.

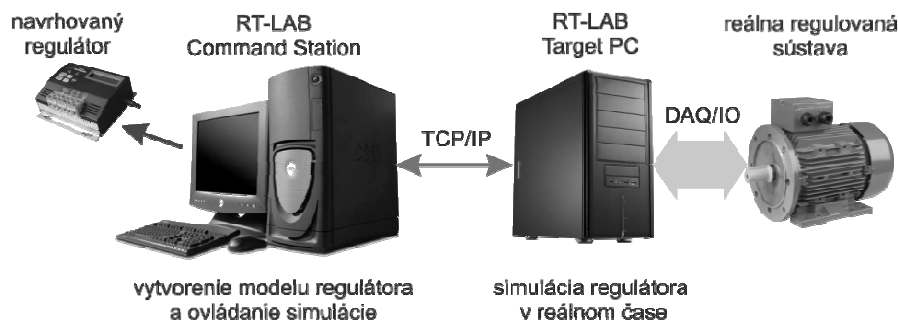
Nahradenie matematického modelu regulovanej sústavy reálnou regulovanou sústavou sa nazýva *Rapid Control Prototyping* (RCP). Regulátor je simulovaný v reálnom čase. Akčný zásah pôsobí na reálnu sústavu a spätná väzba je získaná z reálnym snímačov. RCP umožňuje odladenie regulačnej štruktúry na základe odozvy reálnej regulovanej sústavy. Jedná sa o variantu HIL simulácie.



Obr. 1 Simulácia v reálnom čase s použitím reálneho systému v regulačnej slučke

B. RL-LAB

RT-LAB je distribuovaný systém reálneho času, ktorý umožňuje simulovať modely vytvorené v programe Simulink v reálnom čase s možnosťou využitia HIL simulácie. RT-LAB pracuje s hardvérovou konfiguráciou pozostávajúcou z počítača umožňujúceho editovať modely v Simulinku, ovládať simuláciu v reálnom čase atď. (Command Station), počítača umožňujúceho kompilovanie kódu a simuláciu v reálnom čase (Target PC) a vstupno/výstupných dosiek umožňujúcich komunikáciu systému s okolím (Data Acquisition Board DAQ/IO).



Obr. 2 Distribuovaný systém reálneho času využitý na RCP

Pre simulovanie programovej schémy vytvorenej v Simulinku je potrebné vytvorený model rozdeliť na časť simulovanú v reálnom čase vykonávanú na Target PC a časť slúžiacu na ovládanie simulácie vykonávanú na Command Station (nie v reálnom čase).

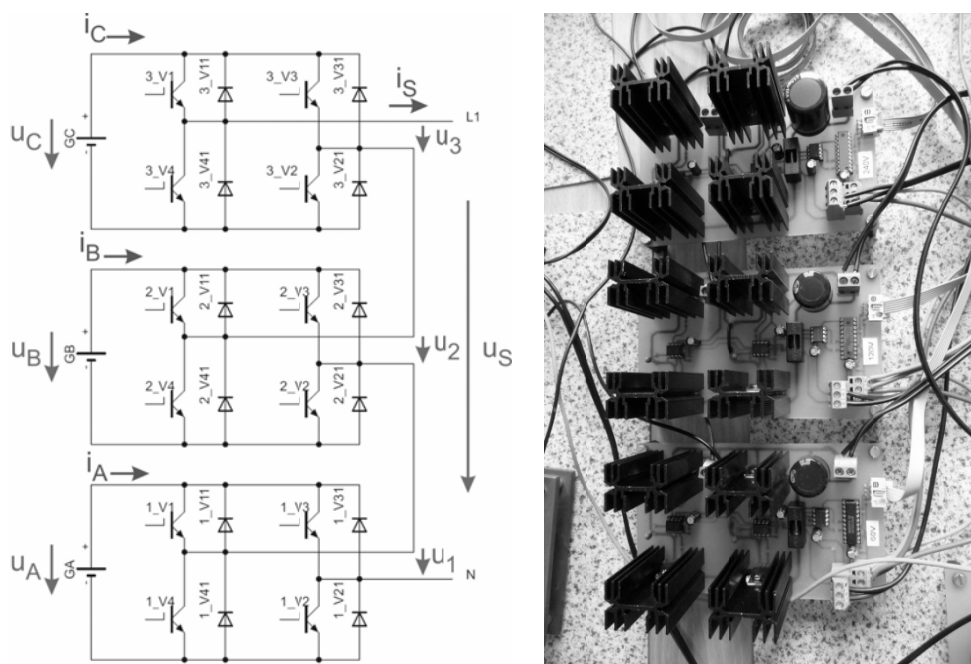
Pre komunikáciu Target PC s okolím bola použitá multifunkčná komunikačná karta NI 6025E. Systém pracuje s pevným krokom výpočtu.

III. KASKÁDOVÝ STRIEDAČ

Viacúrovňové výkonové polovodičové meniče predstavujú riešenie pre meniče vysokých výkonov pracujúcich s vysokým napätím. Viacúrovňové striedače majú v porovnaní s tradičnými dvojúrovňovými meničmi niekoľko výhod [2][10]:

- ich výstupné napätie a prúd majú nižšie harmonické skreslenie, čo umožňuje použitie menších výstupných filtrov,
- použitím nižších napätí majú nižšie du/dt čím sa znižuje kapacitné rušenie a zlepšuje EMC,
- používajú nižšie napätia, čím sa znižuje napäťové namáhanie súčiastok, pracujú pri nižšej spínacej frekvencii, čo má vplyv na zníženie oteplenia súčiastok,
- zníženie súhlasného napätia.

Ich nevýhodou je potreba galvanicky oddelených jednosmerných zdrojov a zložitejšie riadenie. Nutnosť mať k dispozícii niekoľko galvanicky oddelených jednosmerných zdrojov, prípadne rovnaký počet transformátorov, nie je v oblasti fotovoltiky prekážkou. Fotovoltický generátor môže byť ľahko rozdelený na niekoľko samostatných oddelených častí. Problém zložitejšieho riadenia je možné riešiť návrhom regulátora prostredníctvom RCP a simulácie v reálnom čase. Článok sa ďalej zaoberá návrhom riadenia pre 15-úrovňový kaskádový striedač s prúdovým riadením (obr. 3).



Obr. 3 Schéma 15-úrovňového kaskádového striedača a jeho laboratórny model

Počet úrovní výstupného napätia kaskádového striedača pri rôznom napätí jednosmerných zdrojov je definovaný vzťahom:

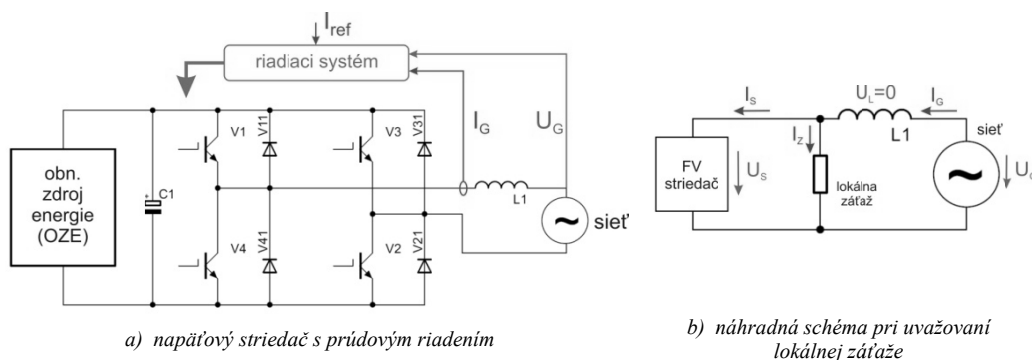
$$n = 2 \left[\sum_{k=1}^d C_k(d) \right] + 1 \quad (1)$$

kde d je počet jednosmerných zdrojov. Navrhnutý kaskádový striedač dokáže na svojom výstupe vytvoriť 15 napätových úrovní (7 kladných, 7 záporných a nulu). Napätia použitých jednosmerných zdrojov sú $U_A = 240 \text{ V}$, $U_B = 120 \text{ V}$, $U_C = 60 \text{ V}$.

IV. NÁVRH RIADENIA KASKÁDOVÉHO STRIEDAČA

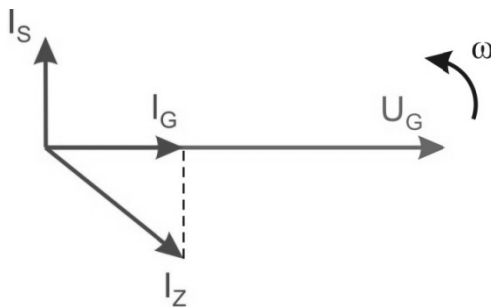
A. Prúdové riadenie

FV striedač s prúdovým riadením (obr. 4) má v porovnaní s FV striedačom s napätovým riadením nezmenený vstup (napájaný je z FV generátora). Pre dosiahnutie prúdového riadenia je potrebné na výstup FV striedača zaradiť tlmivku (alebo iný vhodný filter), keďže samotný napätový striedač je schopný na svojom výstupe vytvárať len príslušné napätie. Cieľom prúdového riadenia je sledovanie referenčnej hodnoty výstupného prúdu aplikovaním vhodného výstupného napätia FV striedača.



Obr. 4 Schéma napätového FV striedača s prúdovým riadením pripojeného k sieti cez filter

V porovnaní so striedačom s napätovým riadením nedokáže striedač s prúdovým riadením pracovať bez prítomnosti siete a neumožňuje ani stabilizáciu napätia siete v bode pripojenia lokálnej záťaže. Má však výhody z pohľadu jednoduchosti riadenia činného a jalového výkonu dodávaného do siete. Umožňuje tak jednoduchú kompenzáciu jalového výkonu odoberaného lokálnou záťažou (obr. 5). Dôvodom je možnosť nezávisle riadiť amplitúdu a fázu výstupného prúdu striedača s prúdovým riadením [7].

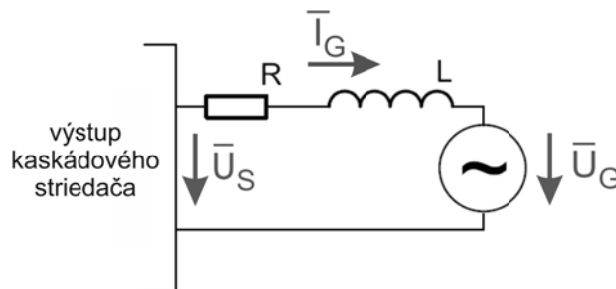


Obr. 5 Fázorový diagram pri kompenzovaní jalového výkonu pomocou napäťového FV striedača s prúdovým riadením

B. Prúdové riadenie s predikciou prúdu siete

V [3] je uvedené prúdové riadenie s predikciou prúdu záťaže pre trojfázový DNPC striedač (Diode Neutral Point Clamped). Toto riadenie bolo upravené pre jednofázový 15-úrovňový kaskádový striedač s prúdovým riadením.

Metóda je založená na predikcii budúceho sieťového prúdu použitím diskretného modelu systému. Regulátor vypočíta nasledovnú hodnotu sieťového prúdu pre všetky možné stavy systému a stav, ktorý vytvára najmenšiu odchýlku medzi požadovanou a skutočnou hodnotou prúdu je vybraný v nasledovnej vzorkovacej perióde.



Obr. 6 Pripojenie kaskádového striedača k sieti

Pre obvod na obr. 6 platí:

$$\bar{U}_S = R\bar{I}_G + L \frac{d\bar{I}_G}{dt} + \bar{U}_G \quad (2)$$

Po diskretizácii s dostatočne malou vzorkovacou periódou T možno vyjadriť prúd ako [3]:

$$\bar{I}_{G(k+1)T} = e^{-\frac{R}{L}T} \bar{I}_{GkT} + \frac{1}{R} \left(1 - e^{-\frac{R}{L}T}\right) (\bar{U}_{SkT} - \bar{U}_{GkT}) \quad (3)$$

Konštanty prúdového regulátora sú:

$$k_1 = e^{-\frac{R}{L}T} \text{ a } k_2 = \frac{1}{R} (1 - k_1) \quad (4)$$

Hodnota sieťového prúdu je predikovaná Lagrangeovou kvadratickou extrapoláciou:

$$\bar{I}_{G(k+1)T}^* = 3\bar{I}_{GkT}^* - 3\bar{I}_{G(k-1)T}^* + \bar{I}_{G(k-2)T}^* \quad (5)$$

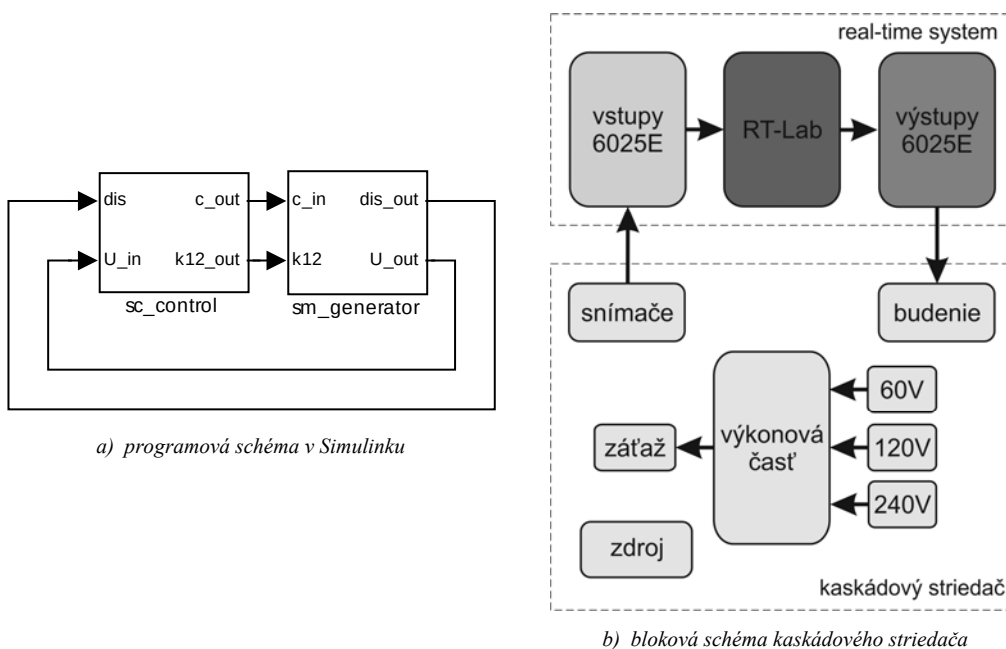
Porovnaním požadovanej a skutočnej hodnoty prúdu vznikne chybová funkcia:

$$z = |\bar{I}_{GkT}^* - \bar{I}_{G(k+1)T}| \quad (6)$$

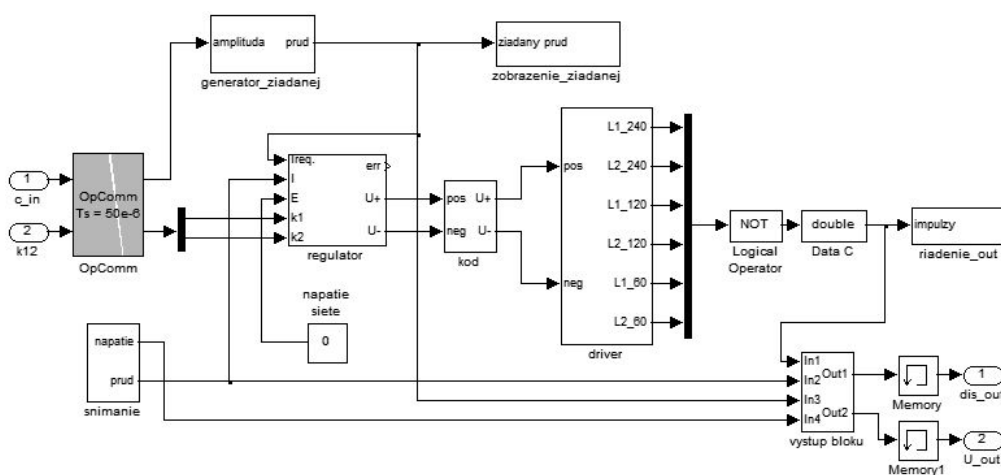
Prúdový regulátor vyberá v každej dobe vzorkovania jeden z 15 napäťových vektorov $\bar{U}_S$ (obr. 6), ktoré je kaskádový striedač na svojom vstupe schopný vytvoriť, tak, aby bola chybová funkcia (7) minimálna.

C. Návrh riadenia pomocou programu RT-LAB

Diskretný regulátor s predikciou prúdu bol realizovaný implementovaním rovníc (3), (5) a (6) pomocou programu Simulink. Konštanty regulátora boli vypočítané pre známe hodnoty odporu a indukčnosti záťaže. Striedač nebol pripojený k sieti, teda napätie $U_G = 0V$. Pevný krok výpočtu modelu regulátora bol nastavený na $T = 50 \mu s$.



Obr. 7 Realizácia riadenia kaskádového striedača prostredníctvom RCP a programu RT_LAB



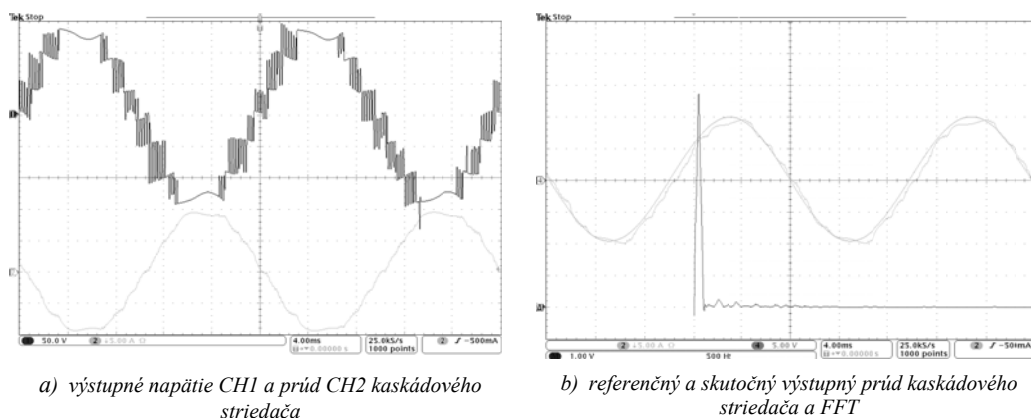
Obr. 8 Blok sm_generator prúdového regulátora

V. OVERENIE ČINNOSTI PRÚDOVÉHO RIADENIA MERANÍM

Činnosť kaskádového striedača s prúdovým riadením bola overená pri napájaní striedača napätím $U_A = 240 \text{ V}$, $U_B = 120 \text{ V}$, $U_C = 60 \text{ V}$. Odpor záťaže bol $13,2 \Omega$ a indukčnosť záťaže 6 mH . Perióda vzorkovania regulátora bola $200 \mu\text{s}$. Maximálna frekvencia spínania polovodičových spínačov je v tomto prípade 5 kHz . Výsledky merania sú zobrazené na obr. 9. Výstupné napätie kaskádového striedača je menené tak, aby prúd záťaže sledoval predpísanú hodnotu prúdu. Zvlnenie výstupného napätia bolo spôsobené nedokonalou filtráciou sieťového napätia napájania laboratórnych jednosmerných zdrojov.

Regulátor dovoľuje doladenie konštánt regulátora, ktoré je potrebné pri zmene parametrov záťaže. Parametre regulátora ovplyvňujú tvar výstupného prúdu kaskádového striedača. Závisí od nich aj celkové harmonické skreslenie výstupného prúdu. Pri uvedených podmienkach bolo celkové harmonické skreslenie 4% pri amplitúde výstupného prúdu 10 A .

Výhodou uvedeného kaskádového striedača je vysoká účinnosť, ktorá dosiahla pri zaťažení striedača na 19% hodnotu 91% . Čím je napájacie napätie jednotlivých mostíkov kaskádového striedača vyššie, tým je frekvencia, s akou príslušný mostík spína, nižšia. Vo výstupnom napätí sa objavujú skoky 60 V na rozdiel od tradičného trojhodnotového striedača, kde by elementárna zmena napätia bola 420 V . Tým sa výrazne znižuje vyžarované rušenie.



Obr. 9 Výstup kaskádového striedača pri prúdovom riadení : ($I_{\max} = 10\text{A}$) ($T = 200\mu\text{s}$, $k_1 = 0,5$, $k_2 = 0,027$, $R = 13,2\Omega$, $L = 6\text{mH}$, $U_A = 240\text{V}$, $U_B = 120\text{V}$, $U_C = 60\text{V}$), účinnosť striedača: 91%

VI. ZÁVER

Návrh regulátora pomocou RCP výrazne zjednodušuje realizáciu prototypu prúdového regulátora. Použitie programu Simulink v spolupráci s programom RT-LAB umožňuje rýchle implementovanie a overenie regulačného algoritmu aj s využitím moderných spôsobov riadenia polovodičových výkonových meničov. Príkladom je aplikovanie adaptivity konštánt prúdového regulátora v závislosti od zmeny hodnôt odporu a indukčnosti záťaže. Použitie Simulinku umožňuje jednoduché implementovanie MPPT algoritmu.

POĎAKOVANIE

Práca bola realizovaná s podporou projektu Vedeckej grantovej agentúry Ministerstva školstva VEGA 1/0099/09.

LITERATÚRA

- [1] A. Ristow, M. Begović, A. Pregelj, A. Rohatgi: "Methodology for Improving PV inverter reliability", *IEEE Transactions on Industrial Electronics*, Vol. 55, No. 7, júl 2008
- [2] S. Khomfoi, L. M. Tolbert: "Multilevel Power Converters", kapitola z H. Rashid: *Power Electronics Handbook*, druhé vydanie, Elsevier, 2007, ISBN10: 0-12-088479-8, strany 451 – 482
- [3] G. S. Perantzakis, F. H. Xepaps, S. A. Papathanassiou, S. N. Manias, "A Predictive Current Control Technique for Three-Level NPC Voltage Source Inverter", [Online]. Dostupné na: http://users.ntua.gr/stpaph/Paper_2.55.pdf
- [4] J. Hamar, R. K. Jordan, I. Nagy, H. Ohsaki: „Virtual Laboratory for Combined Solar Energy System“, EPE 2007 – Aalborg, P.1-P.8
- [5] Opal RT Technologies: "RT-LAB User Guide", [Online]. Dostupné na: <http://www.opal-rt.com/product/rt-lab-professional>
- [6] National Instruments: "6023E/6024E/6025E User Manual", [Online]. Dostupné na: <http://sine.ni.com/nips/cds/view/p/lang/en/nid/10971>
- [7] S. H. Ko, S. R. Lee, H. Dehbonei, Ch. Nayar, "A Comparative Study of the Voltage Controlled and Current Controlled Voltage Source Inverter for the Distributed Generation System" [Online]. Dostupné na: <http://www.itee.uq.edu.au/~aupec/aupec05/AUPEC2005/Volume1/S073.pdf>
- [8] A. Pandey, B. Singh, N. B. Singh, A. Chandra, K. Al-Haddad, D. P. Kothari: "A Review of Multilevel Power Converters", *IEI Journal*, 2006, [Online]. Dostupné na: <http://www.ieindia.org/pdf/86/elp86a163.pdf>
- [9] GPEC: Current Demand of High Performance Inverters for Renewable Energy Systems, 2007, [Online]. Dostupné na: http://www.gpec.dee.ufc.br/publicacoes/Sdaher_Antunes.pdf
- [10] M. Calais, V. G. Agelidis: "Multilevel converters for Single-Phase Grid Connected Photovoltaic Systems- An Overview", 1998, [Online]. Dostupné na: www.ieeexplore.ieee.org/iel4/5707/15282/00707781.pdf
- [11] N. F. Mailh, S. M. Bashi, I. Aris, N. Mariun: "Neutral-Point-Clamped Multilevel Inverter Using Space Vector Modulation", [Online]. Dostupné na: http://www.eurojournals.com/ejsr_28_1_09.pdf

Vektorové riadenie AM s využitím UI

Jaroslav TIMKO, Jaroslava ŽILKOVÁ, Michal KOVÁČ

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jaroslav.timko@tuke.sk, jaroslava.zilkova@tuke.sk

Abstrakt — Príspevok sa zaoberá aplikáciou umelej inteligencie pre vektorovo riadený asynchrónny motor. Riadenie je realizované pomocou dvoch Takagi-Sugeno fuzzy regulátorov. V príspevku je uvedený postup návrhu jednotlivých fuzzy regulátorov, porovnanie ich simulačných výsledkov s klasickým stavovým regulátorom a návrh neurónových pozorovateľov rýchlosti motora.

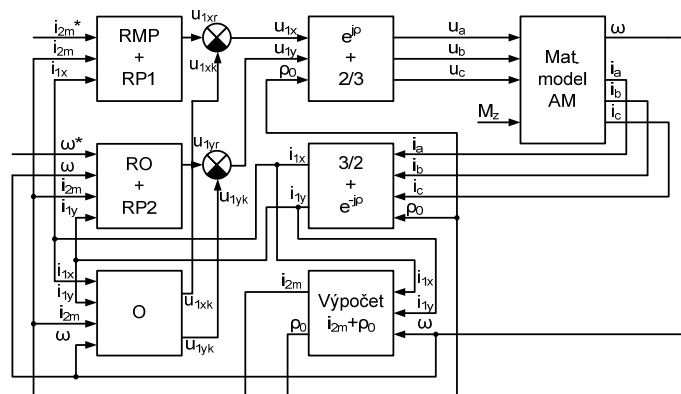
Kľúčové slová — asynchrónny motor (AM), fuzzy regulátor (FR), umelá neurónová sieť (UNS), vektorové riadenie (VR)

I. ÚVOD

Aplikáciou vektorového riadenia v riadení AM sa vytvorili moderné elektrické regulované pohony. Najväčšou nevýhodou konvenčných regulátorov pre tento typ riadenia, je nutnosť poznať všetky parametre a matematický model pohonu, ktorý je následne zjednodušovaný a linearizovaný. Z tohto dôvodu sa v posledných rokoch aj v tejto oblasti začínajú uplatňovať metódy umelej inteligencie, ako je teória fuzzy množín a umelých neurónových sietí, ktoré prinášajú rad nepopierateľných výhod, ako je modelovanie neurčitosti v prirodzenom jazyku a vynikajúce schopnosti aproximácie a tým skvalitňujú a zjednodušujú riadenie pohonov. V tomto príspevku sa zaoberáme návrhom fuzzy vektorového riadenia a neurónovým pozorovateľom uhlovej rýchlosti asynchrónneho motora.

II. NÁVRH FUZZY REGULÁTORA PRE VEKTOROVÉ RIADENIE AM

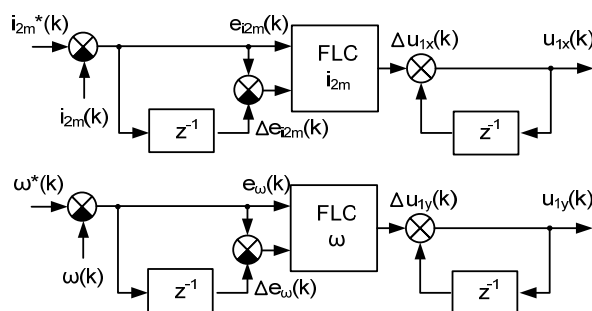
Bloková schéma, priameho vektorového riadenia orientovaného na rotorový tok, je zobrazená na obr. 1.



Obr. 1 Bloková schéma vektorového riadenia AM

Výstupné hodnoty jednotlivých regulátorov sú napätia vyjadrené v súradnicovom systéme x, y otáčajúcim sa synchrónnou rýchlosťou ω_k . Tieto veličiny sú transformované do systému pevných statorových súradníc α, β a následne transformované z dvojfázového súradnicového systému α, β do trojfázového súradnicového systému a, b, c , ktoré následne vstupujú do matematického modelu asynchrónneho motora. Jeho výstupom sú prúdy i_a, i_b, i_c ktoré sú spätne transformované na i_x, i_y . Z týchto prúdov a otáčok asynchrónneho motora sa vypočíta skutočný magnetizačný prúd i_{2m} a uhol ρ_0 potrebný pri jednotlivých transformáciách.

Na riadenie boli použité dva Takagi-Sugeno fuzzy regulátory, ktorých štruktúra je uvedená na obr. 2.



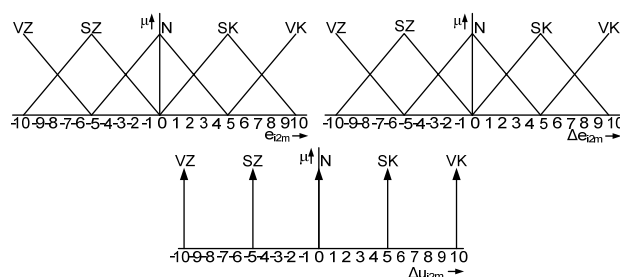
Obr. 2 Schéma PI fuzzy regulátora

Prvý Takagi-Sugeno fuzzy regulátor slúži na reguláciu magnetizačného prúdu a druhý na riadenie uhlovej rýchlosti motora. Regulátory mali dva vstupy a jeden výstup. Ako prvý vstup do regulátorov bola použitá regulačná odchýlka e , ktorá dáva informáciu o rozdiel medzi žiadanou (*) a skutočnou hodnotou vystupujúcou zo systému. Pri regulácii asynchrónneho motora je potrebné poznať aj jej dynamiku a smer. Na to slúži druhý vstup fuzzy regulátora derivácia tejto odchýlky Δe . Výstupom fuzzy regulátora je hodnota akčnej veličiny Δu .

PI fuzzy regulátor s dobou vzorkovania $T=0.001$ s riadi akčnú veličinu na základe informácií o veľkosti regulačnej odchýlky a jej derivácie podľa:

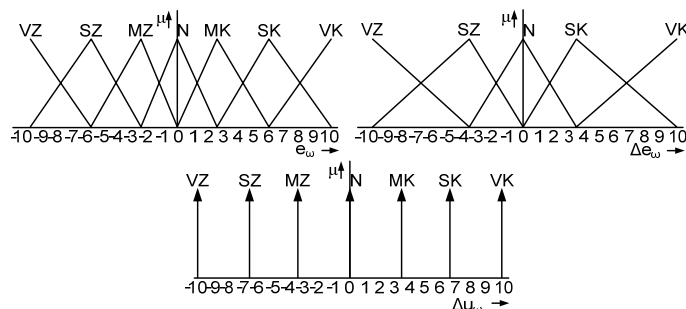
$$\Delta u_{(k)} = K_P \cdot \Delta e_{(k)} + K_I \cdot e_{(k)} \quad (1)$$

Pre regulátor magnetizačného prúdu je použitých pre regulačnú odchýlku a jej deriváciu päť funkcií príslušnosti trojuholníkových tvarov (obr.3). Jednotlivé stavy sú označené (VZ – Veľká Záporná, SZ – Stredná Záporná, N – Nulová, SK – Stredná kladná a VK – Veľká Kladná).



Obr. 3 Funkcie príslušnosti pre Takagi-Sugenov fuzzy regulátora magnetizačného prúdu

V prípade fuzzy regulátora otáčok bolo potrebné zvýšiť počet možných stavov pre zlepšenie rozlišovacej schopnosti v blízkosti ustáleného stavu (obr. 4). Regulačná odchýlka má sedem rôznych stavov a jej derivácia päť s označením (VZ – Veľká Záporná, SZ – Stredná Záporná, MZ – Malá Záporná, N – Nulová, MK – Malá Kladná, SK – Stredná kladná a VK – Veľká Kladná).



Obr. 4 Funkcie príslušnosti pre Takagi-Sugenov fuzzy regulátora otáčok

Na obr. 5 sú uvedené bázy pravidiel pre jednotlivé regulátory, ktoré sú v tvare:

AK $e_{(k)}$ je **A** $\Delta e_{(k)}$ je **POTOM** $\Delta u_{(k)}$ je

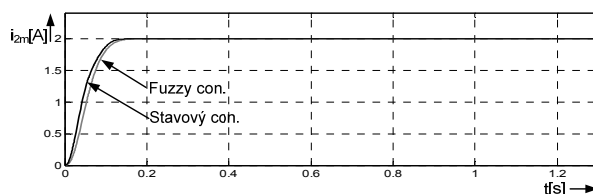
a,						b,							
Δe	VZ	SZ	N	SK	VK	Δe	VZ	SZ	MZ	N	MK	SK	VK
VZ	VZ	VZ	VZ	N	SK	VZ	VZ	VZ	VZ	SZ	N	MK	SK
SZ	VZ	VZ	SZ	SK	VK	SZ	VZ	VZ	SZ	MZ	MK	SK	VK
N	VZ	SZ	N	SK	VK	N	VZ	SZ	SZ	N	SK	SK	VK
SK	VZ	SZ	SK	VK	VK	SK	VZ	SZ	MZ	MK	SK	VK	VK
VK	SZ	N	VK	VK	VK	VK	SZ	MZ	N	SK	VK	VK	VK

Obr. 5 Bába pravidiel pre Takagi-Sugenov fuzzy regulátor magnetizačného prúdu (a,) a uhlovej rýchlosti (b,)

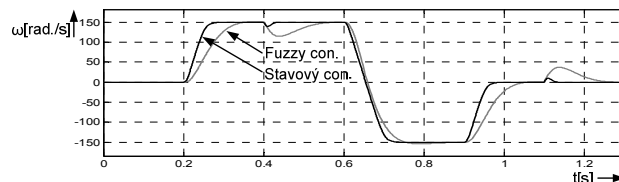
Vektorovo riadenie AM orientované na rotorový tok bolo realizované pomocou navrhnutých fuzzy regulátorov, pričom dosiahnuté simulačné výsledky boli porovnané s klasickým konvenčným stavovým riadením navrhnutým pomocou metódy určenia pólov pre dve vstupné a jednu výstupnú veličinu.

III. DOSIAHNUTÉ VÝSLEDKY

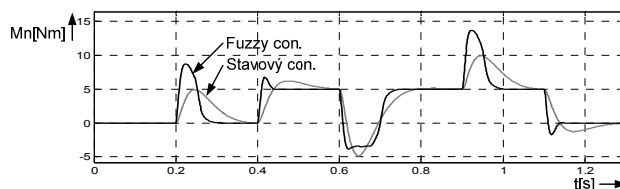
V nasledujúcej časti sú uvedené a porovnané priebehy jednotlivých veličín pohonu pri použití Takagi-Sugenovho fuzzy regulátora a klasického stavového regulátora s predpísaným tmením $d=0,85$ a dobou regulácie $t_r=0,1s$. Simulácie prebehli v programe Simulink (Matlab) na modeli asynchrónneho motora s výkonom 0,75kW, ktorý bol v čase od 0,4s do 1,1s zatážený momentom $M_z=5Nm$.



Obr. 6 Priebeh magnetizačného prúdu i_m fuzzy a stavového reg.



Obr. 7 Priebeh uhlovej rýchlosti ω fuzzy a stavového reg.



Obr. 8 Priebeh momentu motora M_m fuzzy a stavového reg.

Ako je možné vidieť z predchádzajúcich priebehov, fuzzy regulátor reagoval rýchlejšie na prechodové deje, pričom moment motora neprekročil 2,5 násobok nominálneho momentu ($M_N=5,19Nm$).

IV. NÁVRH POZOROVATEĽA UHLOVEJ RÝCHLOSTI

Nevýhodou v súčasnosti používaných snímačov je ich vysoká cena, nároky na údržbu, pokles spoľahlivosti celého systému, preto sa neustále zdokonaľujú metódy bezsnímačového riadenia pohonov.

Pozorovatele na báze umelých neurónových sietí majú rad výhod:

- dokážu aproximovať nelineárne závislosti,

- schopnosť adaptovať sa na základe podnetov okolia,
- UNS je možné trénovať off-line,
- vyhnutie sa zložitým výpočtom pri návrhu klasického pozorovateľa,
- sú robustné.

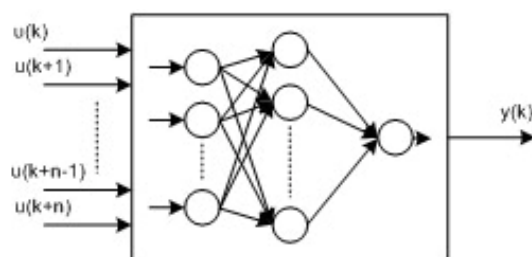
V práci je použitá dopredná neurónová sieť typu 4-7-1 (obr. 9), pričom v prípade vektorového riadenia je vhodné použiť stavové veličiny v súradnicovom systéme x, y. Preto je vstup neurónovej siete vyjadrený :

$$x = [i_{1y}(k), i_{1y}(k-1), u_{1y}(k), u_{1y}(k-1)].$$

Neurónová sieť vykonáva zobrazenie:

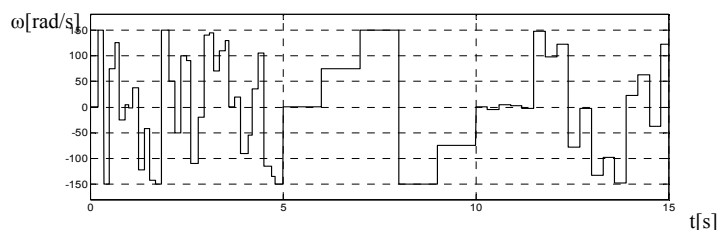
$$\tilde{\omega}(k+1) = f[i_{1y}(k), i_{1y}(k-1), u_{1y}(k), u_{1y}(k-1)].$$

Pre neuróny v skrytej vrstve bola zvolená tansigmoidná a pre výstupný neurón lineárna prenosová funkcia.

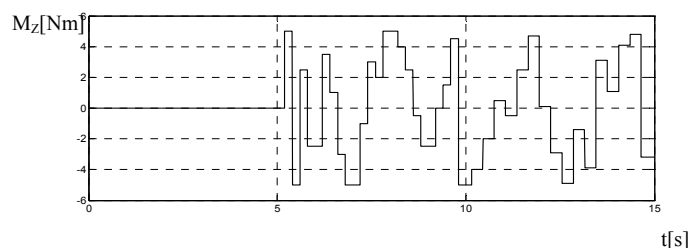


Obr. 9 Štruktúra doprednej neurónovej siete

Učiaci množina (obr. 10 a 11) pokrývala všetky možné stavy AM vo vektorovom riadení s fuzzy regulátormi, v ktorých sa pohon môže nachádzať.



Obr. 10 Priebeh žiadanej uhlovej rýchlosti pri učení UNS

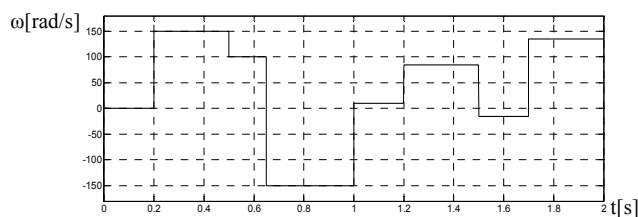


Obr. 11 Priebeh záťažného momentu pri učení UNS

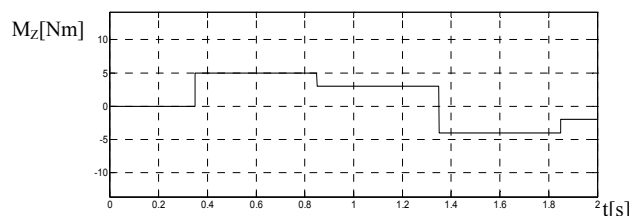
Pri učení bola použitá Levenberg-Marquardtovou metódou učenia, pri ktorej bola dosiahnutá rýchla konvergencia a dobrá presnosť učenia.

V. TESTOVANIE POZOROVATEĽA UHLOVEJ RÝCHLOSTI

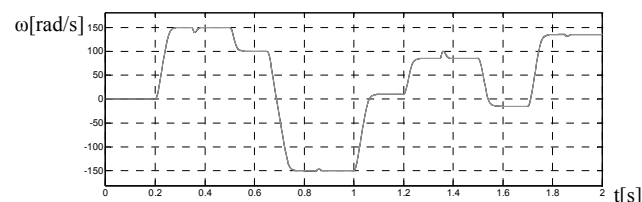
V tejto časti sú uvedené výsledky simulácií neurónového pozorovateľa uhlovej rýchlosti motora, ktoré sú porovnávané so skutočnou uhlovou rýchlosťou z modelu asynchrónneho motora. Simulácie boli uskutočnené v programe Matlab/Simulink s podporou Neural Network Toolboxu. Priebeh žiadanej uhlovej rýchlosti počas testovania je znázornený na (obr. 12) a priebeh záťažného momentu na (obr. 13).



Obr. 12 Priebeh žiadanej uhlovej rýchlosti pri testovaní UNS pozorovateľa uhlovej rýchlosti



Obr. 13 Priebeh záťažného momentu pri testovaní UNS pozorovateľa uhlovej rýchlosti



Obr. 14 Pozorovaná a skutočná ω doprednej UNS 4-7-1

Z priebehov dokumentovaných na obr. 6, 7, 8 a 14 výsledkov je zrejmé, že použitie metód umelej inteligencie zlepšilo vlastnosti vektorového riadenia asynchrónneho motora.

POĎAKOVANIE

Táto práca vznikla s podporou grantu VEGA 1/0006/10.



The authors wish to thank the R&D operational program “Centre of Excellence of Power Electronics Systems and Materials for their Components” for its support. The project is funded by the European Community, ERDF – European Regional Development Fund.

LITERATÚRA

- [1] F. Blaschke, “The principle of field orientation as applied to the new “transvector” closed-loop control system for rotating-field machines”, Siemens Review, 39 (1972), pp.217-220.
- [2] B. K. Bose, “Modern power electronics and AC drives”, Prentice-Hall PTR, Upper Saddle River, NJ 07458 2002.
- [3] P. Brandštetter, L. Štepanec, “Fuzzy logic control of induction motor drive”, IWCIT’01, VŠB-TU Ostrava, Ostrava, 2001.
- [4] K. Hasse, “Zur dynamik drehzahl geregelter antriebe mit stromrichter gespeisten asynchron-kurschlussläufer maschinen”, Techn. Hochsch., Darmstadt, 1969, pp. 74-78.
- [5] J. Timko, J. Žilková, P. Girovský, “Shaft sensorless vector control of an induction motor”, Acta Technica ČSAV, 52, 2007, pp. 81-91.
- [6] J. Timko, J. Žilková, P. Girovský, “Modelovanie a riadenie elektrických pohonov s využitím neurónových sietí”, C-Press Košice, 2009, pp. 201.
- [7] P. Vas, “Artificial-intelligence-based electrical machines and drives”, Oxford University Press, Oxford, 1999.
- [8] J. Vittek, “New trends in electrical drives”, Electrical engineering in practice, 13, 2003, pp.8-11.
- [9] J. Vittek, P. Briš, M. Štulrajter, M. Pácha, “Algoritmus riadenia v kľzavom režime bez kmitania pre polohové riadenie pohonu využívajúceho asynchrónny motor”, AUPEC’08, UNSW, Sydney, Australia, 2008, 12, pp.14 -17.
- [10] J. Vittek, S.J. Dodds, P. Makyš, P. Lehocký, “An Observer Design for Forced Dynamics Control of AC Drives”, Transcom 2007, Žilina, Slovensko, 2007, 06, 25.-27., pp. 207-210.
- [11] J. Žilková, J. Timko, J. Berko, “Speed sensorless control of an induction motor drive using extended Kalman filter”, Acta Technica ČSAV, 50, 2005, pp.279-289.

Analýza kumulatívnej chyby autonómneho vozidla pri inkrementálnom mapovaní priestoru

František BANÍK, Pavol FEDOR, Daniela PERDUKOVÁ

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

frantisek.banik@tuke.sk, pavel.fedor@tuke.sk, daniela.perdukova@tuke.sk

Abstrakt – Pohyb autonómnych vozidiel v neznámom prostredí naráža na problémy s určením správnej polohy vozidla v globálnom súradnicovom systéme. V ostatnom období boli vyvinuté rôzne metódy na odstránenie tejto kumulatívnej chyby výpočtu polohy. Tento článok sa zaoberá prehľadom použiteľných metód pri mapovaní neznámeho prostredia s uložením mapy priestoru v globálnom súradnicovom systéme. Taktiež popisuje experimenty vykonané na reálnom modeli pásového vozidla, ktoré bolo vyvinuté na Katedre elektrotechniky a mechatroniky, Fakulty elektrotechniky a informatiky, Technickej univerzity v Košiciach.

Kľúčové slová – drift correction, kumulatívna chyba, laserový skener, SLAM

I. ÚVOD

Do oblasti mobilných robotov patria autonómne vozidlá. Jednou zo základných úloh týchto vozidiel je prieskum v neznámom prostredí a plnenie dopravných úloh. Mapovaním neznámeho prostredia sa zaoberá mnoho výskumných tímov, pričom využívajú rôzne techniky merania priestoru a následnej analýzy. Základným problémom pri postupnom mapovaní je chyba správneho určenia pozície vozidla pri druhom a ďalšom meraní priestoru. Pre odstránenie tejto chyby boli vyvinuté rôzne metódy, avšak výpočtová náročnosť niektorých metód bráni ich využitiu v praxi.

II. TECHNIKY VYTŮVÁRANIA MÁP A URČENIA POLOHY VOZIDLA

Koncept autonómie mobilných robotov zahŕňa veľa oblastí poznania, metód a taktiež algoritmov navrhnutých pre riadenie trajektórie, vyhýbanie sa prekážkam, lokalizáciu, tvorbu mapy atď.. Úspech pri plánovaní cesty a navigácii autonómneho vozidla závisí od dosiahnuteľnosti oboch, dostatočného určenia pozície vozidla a presnej reprezentácie priestoru.

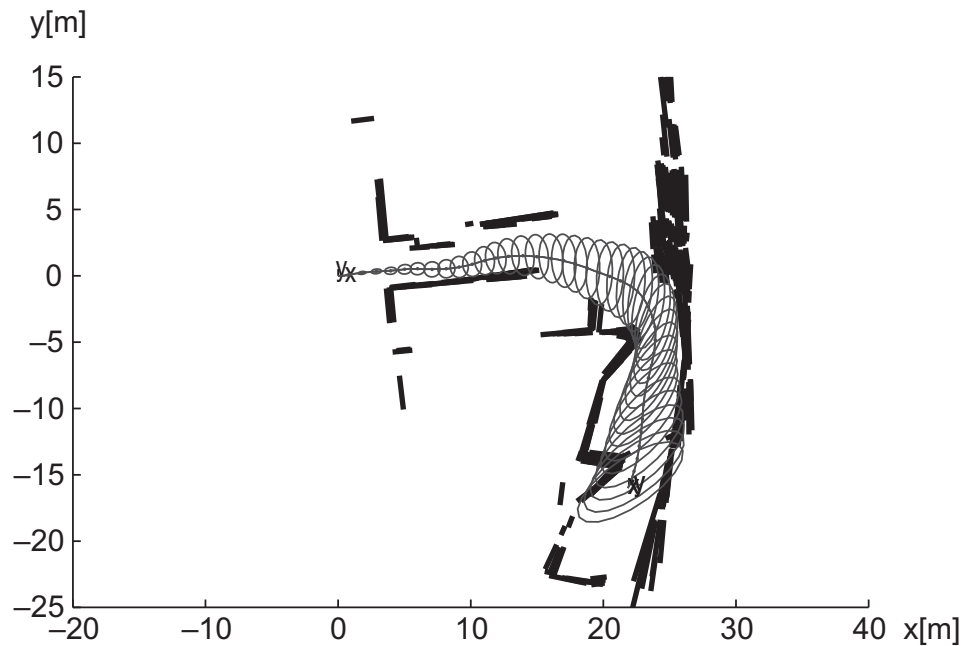
Úloha vytvorenia mapy pozostáva z týchto častí:

- Zosnímanie okolitého prostredia s použitím snímačov na vozidle v čase k , (laserový skener, vizuálny systém, sonar, atď.)
- Reprezentácie zosnímaných dát (fitácia, vykreslenie)
- Integrácie aktuálnych pozorovaní v čase k s už uloženými dátami, ktoré boli zosnímané v čase $k - 1$

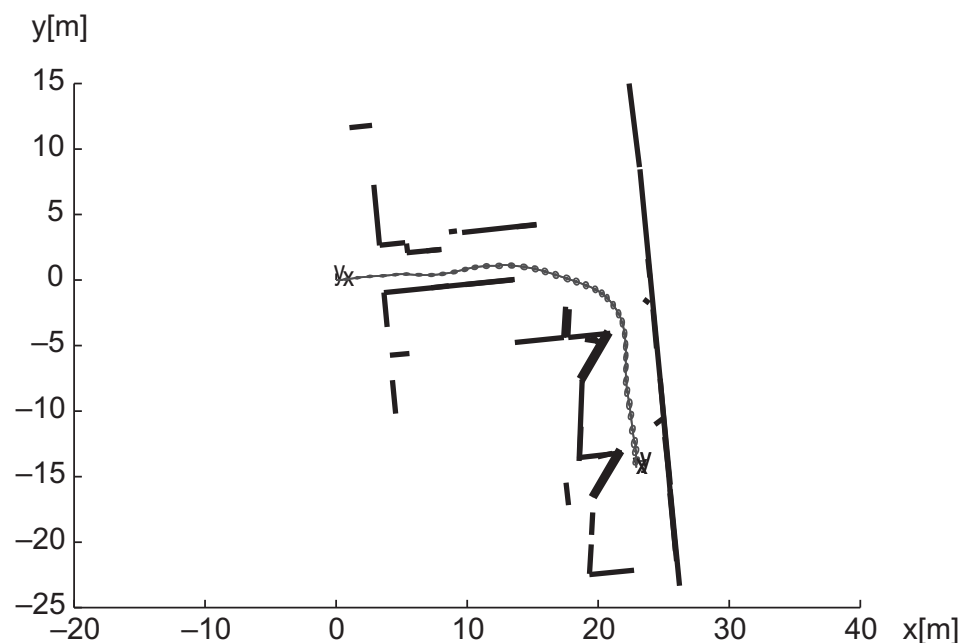
Najjednoduchší prístup k tvorbe mapy je prostredníctvom prístupu dead-reckoning, pre určenie polohy. Táto metóda je založená na výpočte aktuálnej polohy z predchádzajúcej polohy vypočítanej v čase $k - 1$, prejdenej vzdialenosti a zatáčania. Avšak táto metóda je nevhodná pre dlhotrvajúce prieskumné úlohy [1], pretože s narastajúcim časom sa zvyšuje kumulatívna chyba výpočtu polohy, obr. 1.

Problémom je teda spojenie tvorby mapy s výpočtom aktuálnej polohy vozidla. V ostatnej dekáde spopularizoval literatúru o robotike iný prístup nazvaný *Simultaneous Localization And Mapping* v skratke SLAM.

Myšlienku reprezentácie štruktúry priestoru navigácie v diskretnom čase s použitím *state-space* hierarchie prvýkrát publikovali Randall Smith, Mathew Self a Peter Cheeseman v roku 1988 [2]. Autori odprezentovali koncept stochastickej mapy a vyvinuli precízne riešenie pre SLAM problém s využitím rozšíreného Kalmanovho filtra. Spracovanie naskenovaných dát s použitím tejto metódy je na obr. 2, pre porovnanie s nespracovaným skenovaním z obrázka 1.



Obr. 1 Laserové skenovanie priestoru s výpočtom aktuálnej polohy metódou dead-reckoning, kumulatívna chyba polohy



Obr. 2 Využitie metódy SLAM pre úpravu dát z obrázka 1

Popísaných bolo veľa úspešných použití tejto metódy pre vnútorné [1], vonkajšie [3], podvodné [4] a vzdušné nasadenie [5].

Prístup na báze EKF pri SLAM je charakterizovaný rozšíreným stavovým vektorom, ktorý sa skladá z pozície vozidla a pozícií jednotlivých elementov v mape, rekurzívne určených z dostupných pozorovaní v čase k a model pohybu vozidla medzi časovými vzorkami $k-1$ a k . Časové a pamäťové požiadavky pre EKF-SLAM vychádzajú z požiadaviek pre uchovanie plnej kovariančnej matice, čo je $O(n^2)$ kde n je počet obrysov v mape. Pre zníženie výpočtovej náročnosti SLAM vo veľkých priestoroch bolo vynaložené veľké úsilie. Súčasné metódy venované výpočtovej náročnosti sa zaoberajú ohrozenou časťou mapy. Odkladacia metóda [6] a metóda kompresného filtra [3] významne redukujú výpočtovú náročnosť bez zníženia presnosti mapy, aj keď vyžadujú $O(n^2)$ krokov pre celkové vykreslenie. Metóda delenia kovariančných priesečníkov [7] obmedzuje výpočtovú záťaž, ale znižuje presnosť mapy. Metóda rozptýleného rozšíreného informačného filtra [8] je vhodná pre získanie približnej mapy v konštantnom časovom kroku,

okrem uzatvárania slučky v mape. Všetky tieto citované metódy pracujú s jednou reprezentáciou absolútnej mapy čo spôsobuje problémy s nelinearitami a nepresnosťami pri mapovaní veľkých priestorov [9].

Iný prístup, ktorý garantuje statickú nezávislosť, je návrh relatívnej stochastickej mapy k lokálnemu prostrediu. Je to metóda zachytenia lokálnej mapy [10] a metóda viazaného lokálneho *submap* filtra [11]. Obmedzením veľkosti mapy sa tieto operácie stávajú časovo konštantnými pre každý krok. Tieto lokálne mapy sú periodicky zahrnuté do globálnej absolútnej mapy. Stanovením obnovovania tejto lokálnej mapy sa redukuje nežiaduci efekt linearizácie. Každá mapa obsahuje spojenia pre ostatné susedné mapy. Týmto sa formuje stromová štruktúra v ktorej sa však nemôžu nachádzať slučky. Vytvorenie spojenia medzi lokálnymi mapami, založené na prínavosti grafov, je popísané v: Atlase [12], Sieťové spájanie obrysov mapy [13] a SLAM s konštantným časom [14]. Tieto techniky znižujú výslednú optimálnosť globálnej mapy, keďže nezávádzajú slučkovú konzistenciu v grafe. Zavedenie slučkovej konzistencie, pri časovo lineárnej metóde, so zachovaním optimálnosti globálnej mapy je popísané v [15] ako *Hierarchical SLAM*. Iná technika *FastSLAM* [16] využíva EKF filtre pre každý parameter samostatne, ako je trajektória vozidla a samostatne pre polohu každého obrysu mapy. Takéto rozdelenie SLAM medzi lokalizáciu a problém mapovania každého obrysu vyžaduje výpočtovú náročnosť $O(\log(n))$. So zvyšujúcou sa zložitou a zväčšujúcim sa rozmerom prostredia je potrebné meniť mierku delenia obrysov. Obzvlášť pri uzatváraní slučiek v mape.

Iná trieda SLAM techniky je založená na odhade polohy robota, na základe minimalizácie odchýlky medzi prekrývajúcimi sa bodmi pri laserovom skenovaní. Reprezentácia mapy je súbor polôh robota a odpovedajúcich súborov naskenovaných bodov. Využitie párovania naskenovaných bodov medzi jednotlivými polohami robota je popísané v práci Gutmanna a Konoligieho z roku 1999 [17]. Pozície v okolí slučky sú odhadované použitím konzistentného odhadu, ktorého výpočtová náročnosť je $O(n^3)$ a preto je táto metóda nepoužiteľná pre nasadenie v *real-time* systémoch. Podobná metóda pre vytvorenie konzistentnej mapy s mnohými cyklami bola navrhnutá kolektívom autorov pre vytváranie 3D máp v banských dielach [18]. Využitím kvadratickej funkcie sú zhodné body zahrnuté do optimalizačného algoritmu, ktorý prepočítava celkovú trajektóriu.

Existujú dve fundamentálne obmedzenia v porovnaní s EKF SLAM. Prvým je, že neexistuje explicitná reprezentácia neurčitosti v odhade pozícií robota a výslednej mapy. Čoho výsledkom je, že je neznáma konvergencia a konzistentnosť. Druhé obmedzenie je vysoké spoliehanie sa na presnosť laserového merania. To vyžaduje zvýšenie presnosti a hustoty snímania, tak ako je to pri sonaroch, alebo vizuálnych systémoch.

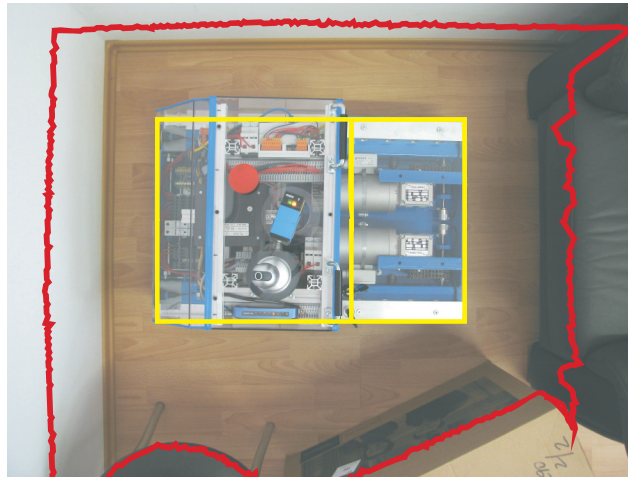
III. LASEROVÝ SKENER

Na Katedre elektrotechniky a mechatroniky, Fakulty elektrotechniky a informatiky, Technickej univerzity v Košiciach, bol vyvinutý model inteligentného pásového vozidla, na ktorom sú testované a vyvíjané metódy pre oblasť autonómnych vozidiel. Pre interakciu vozidla s okolitým prostredím bol zvolený laserový skener. Pri výbere technológie bol kladený dôraz na presnosť interpretácie okolitého "sveta". V súčasnosti sú dostupné rôzne riešenia pre meranie vzdialenosti. Z týchto dôvodov bol vybraný laserový skener. Keďže uvažujeme o testovacom vozidle pre aplikáciu nových teórií, nie je potrebné trojdimenzionálne skenovanie priestoru. Zvolený skener je dvojdimenzionálny. Skener je skonštruovaný kombináciou laserového snímača vzdialenosti a mechanickej otočnej konštrukcie. Laserový snímač má dosah 5,3m a rýchlosť ustálenia meranej vzdialenosti 50ms. Mechanická konštrukcia je zložená z krokového motora a inkrementálneho snímača. Krovový motor má krok 0,72°, čo je 500 skenovaných bodov pre celý priestor 360°. Z týchto údajov je zrejmý celkový čas skenovania priestoru 25s. Pri takomto čase nie je možné využiť tento skener v *real-time* aplikácií, avšak pre laboratórne testovanie to nepredstavuje prekážku. Presný spôsob riadenia tohto laserového skenera je popísaný v [19]. Uvedený laserový skener bol testovaný v priestore kancelárie, pre overenie presnosti vykreslenej mapy.

IV. OVERENIE PRESNOSTI LASEROVÉHO SKENERA

Overenie presnosti získaných bodov bolo vykonané pri skenovaní úzkeho priestoru, kvôli možnosti vnorenia fotografie do vektorovej mapy v rovnakej mierke. Výsledné spojenie fotografie a naskenovanej vektorovej mapy je na obrázku 3.

Mapa a rozmery objektov na fotografii sú v rovnakej mierke. Záver tohto experimentu je, že presnosť skenovania je dostatočná pre aplikáciu autonómnej navigácie na základe takto získaných dát. Presnosť skenera sa znižuje so vzdialenosťou objektu, avšak aj vo vzdialenosti 5,3m je presnosť merania $\pm 1\text{cm}$, čo je odchýlka 0,19%. Takáto chyba je zanedbateľná pri rozmeroch priestoru a vozidla.



Obr. 3 Zúžený priestor kancelárie s vnorenou mapou pri overovaní presnosti skenera

V. OVERENIE PRESNOSTI VÝPOČTU AKTUÁLNEJ POLOHY

Aktuálna poloha vozidla je zadefinovaná parametrami x, y, ψ , podľa matematického modelu [20]. Základom pre výpočet týchto parametrov je prejdená vzdialenosť ľavého a pravého pásu θ_1 a θ_2 . Výpočet je realizovaný v riadiacej jednotke vozidla ako diskretný proces podľa vzťahov:

$$C_k = C_{k-1} + \left(\frac{\theta_1 + \theta_2}{2} \right) \quad (1)$$

$$\psi_k = \arcsin \left[\frac{\theta_1 - (C_k - C_{k-1})}{\sqrt{[\theta_1 - (C_k - C_{k-1})]^2 + 0,09}} \right] \quad (2)$$

$$x_k = x_{k-1} + [(C_k - C_{k-1}) \cos \psi] \quad (3)$$

$$y_k = y_{k-1} + [(C_k - C_{k-1}) \sin \psi] \quad (4)$$

Kde C_k je posun ťažiska celého vozidla a konštanta 0,09 je odvodená od šírky vozidla, resp. od vzdialenosti medzi stredmi pásov. Každá ďalšia poloha vozidla je vypočítaná diskretné v čase a je teda zaťažená kumulatívnou chybou určenia polohy, keďže vychádza z predchádzajúcej polohy ($x_{k-1}, y_{k-1}, \psi_{k-1}$).

Overenie vypočítanej polohy bolo vykonané manuálnym pohybom vozidla v priestore, podľa trajektórie zobrazenej na obrázku 4 modrou farbou. Výsledná vypočítaná poloha vozidla je v bode V a skutočná poloha nameraná voči okolitým objektom je v bode S .



Obr. 4 Porovnanie vypočítanej polohy a skutočnej nameranej polohy vozidla

Chyba sa prejaví hlavne pri zatáčaní vozidla. Pri krátkom presune je vypočítaná poloha relatívne presná, avšak pri dlhodobom pohybe sa kumulovaním chyby úplne vzdiali od reálnej polohy. Tu sa vytvára priestor pre vývoj korekčných pravidiel, buď v závislosti na opätovnom naskenovaní, alebo nelineárnou korekciou priamo pri výpočte.

VI. ZÁVER

V prvej časti je uvedený prehľad metód pre korekciu kumulatívnej chyby výpočtu polohy vozidla. Na reálnom modeli pásového vozidla s implementovaným dvojdimenzionálnym laserovým skenerom boli vykonané experimenty pre overenie presnosti jednotlivých prvkov sústavy. Z týchto experimentov je možné zadefinovať základné závery: laserové meranie a definícia statických prvkov v priestore je dostatočne presné pre aplikáciu v priestore s rozmermi rádovo v metroch. Výpočet polohy vozidla v závislosti od otáčania pásov je zaťažovaný kumulatívnou chybou a nie je vhodný pre dlhodobé určenie polohy vozidla. Riešením korekcie kumulatívnej chyby, v tomto pásovom vozidle, je opätovná korekcia polohy na základe nových skenovaných dát, ktoré je potrebné porovnať s predchádzajúcou polohou objektov, a z tohto rozdielu vypočítať odchýlku polohy vozidla.

POĎAKOVANIE



Tento článok bol vypracovaný s podporou projektu Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy, ITMS 26220220029, ktorý je spolufinancovaný zo štrukturálneho fondu EÚ ERDF v rámci výzvy OPV a V-2008/2.2/01-SORO a prioritnej osi 2 Podpora výskumu a vývoja.

LITERATÚRA

- [1] J. A. Castellanos, J. M. M. Montiel, J. Neira, and J. D. Tardós, "The spmap: a probabilistic framework for simultaneous localization and map building," *IEEE Transactions on Robotics and Automation*, vol. 15, pp. 948–953, 1999.
- [2] R. Smith, M. Self, and P. Cheeseman, "A stochastic map for uncertain spatial relationships," in *Proceedings of the 4th international symposium on Robotics Research*. Cambridge, MA, USA: MIT Press, 1988, pp. 467–474.
- [3] J. E. Guivant and E. M. Nebot, "Optimization of the simultaneous localization and map-building algorithm for real-time implementation," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 242–257, Jun 2001.
- [4] J. L. Leonard, R. N. Carpenter, and H. J. S. Feder, "Stochastic mapping using forward look sonar," *Robotica*, vol. 19, no. 5, pp. 467–480, 2001.
- [5] J. H. Kim and S. Sukkarieh, "Airborne simultaneous localisation and map building," *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 406–411, September 2003, Taipei, Taiwan.
- [6] J. Knight, A. Davison, and I. Reid, "Towards constant time slam using postponement," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 406–412, March 2001, Maui, Hawaii.
- [7] S. J. Julier and J. K. Uhlmann, "Building a million beacon map," *SPIE International Society for Optical Engineering*, vol. 4571, pp. 10–21, Oct. 2001.
- [8] Y. Liu and S. Thrun, "Results for outdoor-slam using sparse extended information filters," *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 1227–1233, Sep. 2003, Taipei, Taiwan.
- [9] J. A. Castellanos, J. Neira, and J. D. Tardós, "Limits to the consistency of ekf-based slam," *5th IFAC Symposium on Intelligent Autonomous Vehicles*, 2004, Lisbon, Portugal.
- [10] J. D. Tardós, J. Neira, P. Newman, and J. Leonard, "Robust mapping and localization in indoor environments using sonar data," *International Journal of Robotics Research*, vol. 21, pp. 311–330, 2002.
- [11] S. B. Williams, G. Dissanayake, and H. F. Durrant-Whyte, "An efficient approach to the simultaneous localisation and mapping problem," *IEEE International Conference on Robotics and Automation, ICRA*, vol. 1, pp. 406–411, 2002, Washington DC.
- [12] M. Bosse, P. Newman, J. Leonard, M. Soika, W. Feiten, and S. Teller, "An atlas framework for scalable mapping," *IEEE International Conference on Robotics and Automation*, vol. 2, no. 3, pp. 1899–1906, Sept. 2003, Taipei, Taiwan.
- [13] T. Bailey, "Mobile robot localisation and mapping in extensive outdoor environments," Ph.D. dissertation, Australian Centre for Field Robotics, University of Sydney, Sydney, August 2002.
- [14] J. J. Leonard and P. M. Newman, "Consistent, convergent and constant-time slam," *International Joint Conference on Artificial Intelligence*, August 2003, Acapulco, Mexico.
- [15] C. Estrada, J. Neira, and J. D. Tardós, "Hierarchical slam: real-time accurate mapping of large environments," *IEEE Transactions on Robotics and Automation*, vol. 21, pp. 588–596, August 2005.
- [16] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "Fastslam: A factored solution to the simultaneous localization and mapping problem," in *Proceedings of the AAAI National Conference on Artificial Intelligence*. Edmonton, Canada: AAAI, 2002.
- [17] J. S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*. Monterey, California: CIRA, 1999, pp. 318–325.
- [18] S. Thrun, D. Hähnel, D. Ferguson, M. Montemerlo, R. Triebel, W. Burgard, C. Baker, Z. Omohundro, S. Thayer, and W. Whittaker, "A system for volumetric robotic mapping of abandoned mines," in *IEEE International Conference on Robotics and Automation*, Taipei, Taiwan, 2003, pp. 4270–4275.
- [19] F. Baník, "2D Laser Scanner for the Navigation of the Autonomous Vehicle," *9th Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics Technical University of Kosice*, no. 1, pp. 114–116, May 2009.
- [20] F. Baník, "Kinematický model pásového podvozka s dvoma pohonmi," in *Electrical Engineering and Informatics*. Faculty of Electrical Engineering and Informatics of Technical University of Kosice, September 2010, pp. 853–856.

Analýza vlastností elektromagnetického nástroja na gravírovavnie kameňa

Mišel BATMEND

Katedra elektrotechniky a mechatroniky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

misel.batmend@tuke.sk

Abstrakt — V predkladanom článku je principiálne popísaná konštrukcia elektromagnetického diamantového nástroja na gravírovanie kameňa a jeho vlastnosti. Ide o elektronicky riadený diamantový hrot gravírujúci nárazom diamantu na povrch tvrdého materiálu. Článok obsahuje merania na prototyp nástroja, ktoré popisujú základné vlastnosti nástroja a možnosti jeho využitia.

Kľúčové slová — gravírovanie kameňa, rytie portrétov, gravírovací nástroj

I. ÚVOD

Portréty gravírované do náhrobných kameňov sú obľúbeným spôsobom uchovania spomienky na zosnulého. V súčasnosti sú portréty ryté prevažne ručne, no postupne sa presadzuje strojová výroba. Ide prevažne o CNC plotre, ktoré používajú ako nástroj na gravírovanie laser [1] alebo frézu [2]. Predkladaný nástroj je ďalším technickým riešením nástroja, ktorý vychádza najmä z nástrojov používaných na ručné gravírovanie. Tie sú zvyčajne konštruované ako vibračné nástroje s hrotom, ktorý udiera o tvrdý materiál a tak vytvára gravírovaný vzor (viď [3]). Ako akčný člen slúži pneumatický pohon alebo elektromagnet napájaný striedavým napätím. Jednotlivé údery sa tak nedajú ovládať, čo však ani nie je potrebné, pretože umelec drží vibrujúci nástroj v ruke a riadi ho pohybom ruky.

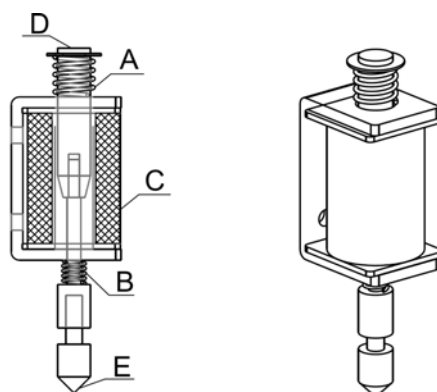
Ak však chceme použiť takýto nástroj na strojové gravírovanie, je výhodné riadiť každý úder samostatne. V spojení s XY plotrom je možné údermi na správne miesta vytvoriť obraz zodpovedajúci čiernobielej predlohe.

Napájanie elektromagnetu jednosmerným impulzovým meničom s reverzáciou napätia umožňuje v princípe meniť dĺžku časov zapnutia a vypnutia elektromagnetu a taktiež veľkosť prúdu tečúceho elektromagnetom. To sú parametre, s ktorými budeme experimentovať, aby sme dosiahli pokiaľ možno optimálnu funkčnosť nástroja.

II. POPIS NÁSTROJA

A. Konštrukcia

Nástroj (Obr. 1) sa skladá z elektromagnetu s pohyblivým jadrom, na konci ktorého je upevnený diamantový hrot a sústavy pružín A a B.



Obr. 1 Gravírovací nástroj: A – vrchná pružina, B – spodná pružina, C – vinutie elektromagnetu. D – pohyblivé iadro. E – diamantový hrot

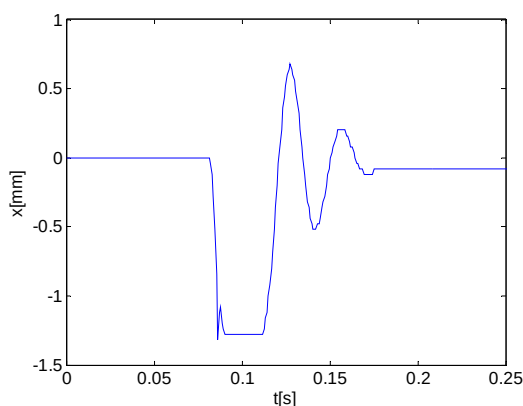
Pružina A zabezpečuje vytiahnutie jadra po skončení pôsobenia elektromagnetickej sily, ktorá jadro vtiahla. Pružina B slúži na tlmenie nárazu pri spätnom pohybe jadra. Taktiež prispieva k zvýšeniu energie úderov. Koeficienty pružnosti oboch pružín sú vo vzťahu:

$$k_A > k_B \quad (1)$$

B. Funkcia

Po pripojení napätia na vinutie elektromagnetu je železné jadro magnetu vplyvom vzniknutého magnetického poľa vtáňované do otvoru vo vinutí. Jadro, pevne prepojené s diamantovým hrotom tak vytvára vertikálny pohyb. Ak do vhodnej vzdialenosti umiestnime leštenú kamennú dosku, udrie diamant na dosku a spôsobí narušenie lešteného povrchu, ktoré sa spravidla javí ako malý svetlý bod s hĺbkou a priemerom niekoľko desiatin milimetra. Jadro je teda vtiahnuté na úroveň, kedy sa diamant dotýka dosky. Pružina A je stlačená (Obr. 2 čas $t = 0,1s$).

Ak odpojíme napájacie napätie vinutia, zanikne magnetické pole a tým aj sila, ktorá vtáňovala jadro. Pružina A vytiahne jadro. Moment zotrvačnosti jadra a sústava pružín A a B spôsobia zákmity okolo rovnovážnej polohy. Následne sa poloha jadra ustáli na pôvodnej úrovni, kedy sú sily medzi pružinami A a B v rovnováhe. Priebeh polohy hrotu v čase je znázornený na obrázku (Obr. 2).



Obr. 2 Priebeh polohy hrotu pri údere

Ide o priebeh nameraný na experimentálnom nástroji zhotovenom podľa Obr. 1. Na meranie polohy sme použili lineárny inkrementálny snímač s rozlíšením 25 impulzov/mm. Perióda vzorkovania v časovej osi je 0,819 ms. Snímač bol upevnený na jadre tak, aby sa jeho vplyv (pridaná hmotnosť) na meranie minimalizoval.

Z merania je vidieť, že sa hrot po prvom styku s doskou ($t=0,08s$) od dosky odrazí a následne opäť dopadne. Pre narušenie povrchu dosky je rozhodujúca energia prvého úderu. Kvôli dosiahnutiu najvyššej frekvencie úderov je vhodné nastaviť časovanie tak, aby hrot uderel na dosku len raz a energia vzniknutá odrazom sa využila na vrátenie hrotu do nulovej polohy ($x=0$). Pri vhodnom načasovaní nasledujúceho úderu sa dá využiť aj energia zákmitu okolo nulovej polohy.

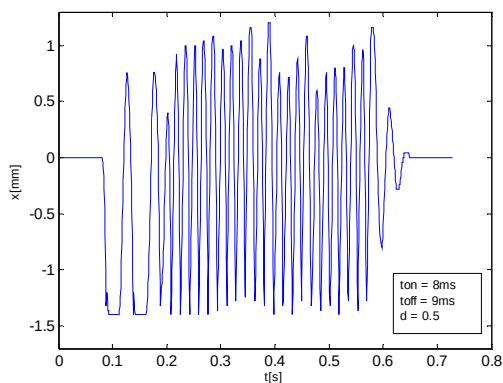
III. MERANIE NA NÁSTROJI

Aby sme zistili maximálnu frekvenciu úderov a ďalšie vlastnosti nástroja, urobili sme sériu meraní. Elektromagnet je budený impulzovým meničom s reverzáciou napätia, ktorý umožňuje nastaviť veľkosť prúdu vinutím a taktiež čas zapnutia a vypnutia elektromagnetu. Experimentovaním s nastavením týchto troch parametrov sme urobili nasledujúce merania. Obr. 3 ukazuje meranie pri striede $d=0,5$, čase zapnutia $t_{on}=8ms$ a vypnutia $t_{off}=9ms$. Prvé dva údery sú pre porovnanie urobené s nižšou frekvenciou ($t_{on}=30ms$, $t_{off}=20ms$).

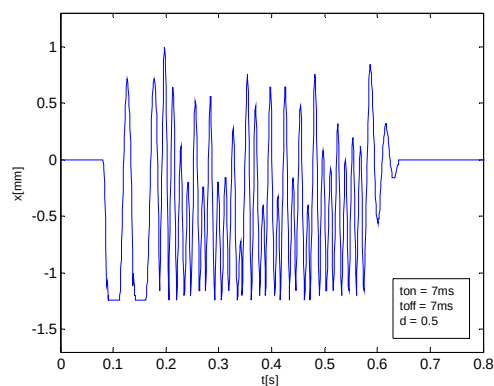
Skracovaním času zapnutia a vypnutia pri rovnakej striede dochádza k nepravidelnosti úderov. Hrot sa nestihne vrátiť do nulovej polohy predtým než príde ďalší úder (Obr.4). Pri striede $d=0,5$ je teda maximálna frekvencia úderov:

$$f_{d0,5} = \frac{1}{t_{on}+t_{off}} = \frac{1}{8 \cdot 10^{-3} + 9 \cdot 10^{-3}} = 58,8Hz \quad (2)$$

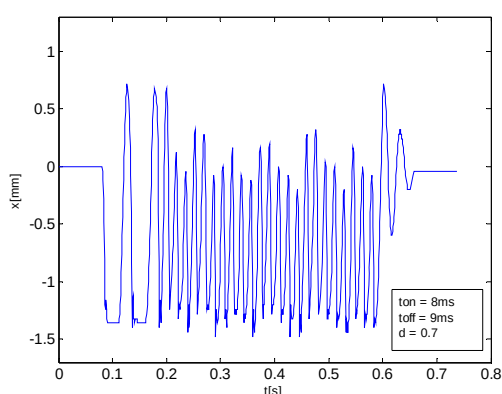
Zväčšením striedy napätia zvýšime prúd pretekajúci vinutím elektromagnetu, a teda aj silu ktorá pôsobí na pohyblivé jadro. To vedie k silnejšiemu úderu hrotu na dosku, čo sa prejaví väčším priemerom a hĺbkou vytvoreného svetlého bodu. Obr. 5 ukazuje meranie so striedou $d=0,7$. Pri čase zapnutia $t_{on}=8ms$ a vypnutia $t_{off}=9ms$ však badať efekt odrazu a opätovného dopadu hrotu na dosku. Aby sme tento efekt odstránili, skrátili sme čas zapnutia na $t_{on}=7ms$ (Obr. 6).



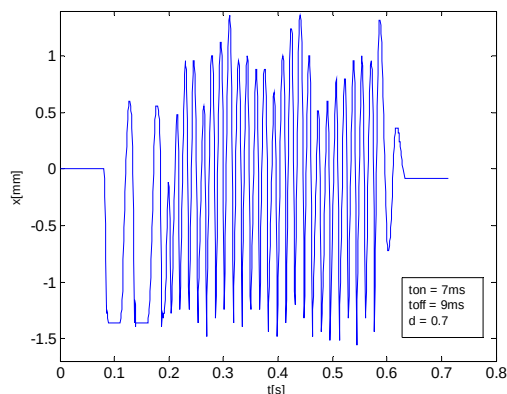
Obr. 3 Poloha hrotu pri úderoch so striedou 0,5. Maximálna frekvencia úderov, bez prelínania jednotlivých úderov.



Obr. 4 Poloha hrotu pri úderoch so striedou 0,5. Pri vyššej frekvencii sú úderu nepravidelné, hrot sa nestihne vrátiť do nulovej polohy.



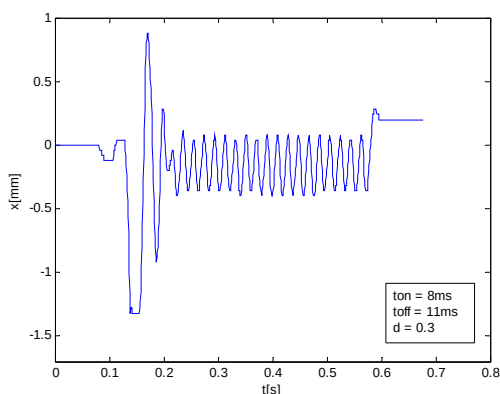
Obr. 5 Poloha hrotu pri úderoch so striedou 0,7. Pri nižšej frekvencii badať zdvojenie úderu odrazom a opätovným dopadom hrotu.



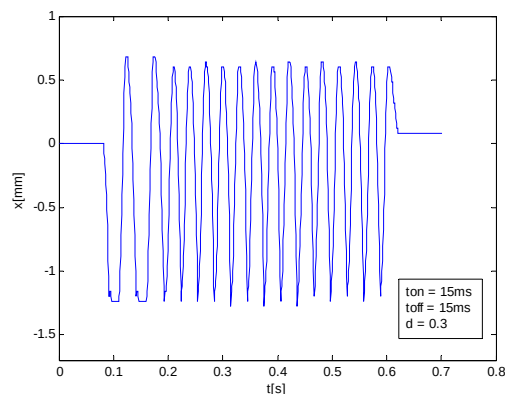
Obr. 6 Poloha hrotu pri úderoch so striedou 0,7. Zvýšením frekvencie sa zdvojenie úderu odstráni.

Zmenšením striedy napätia dosiahneme jemnejší úder hrotu čo sa prejaví menším svetlým bodom na kamennej doske. Na Obr. 7 vidieť, že pri striede $d=0,3$ a časoch $t_{on}=8ms$ a $t_{off}=11ms$ nedochádza k styku hrotu s doskou, pretože čas zapnutia je príliš krátky na to aby sa hrot k doske dostal. Zväčšením časov zapnutia a vypnutia sa tento problém odstráni (Obr. 8). Frekvencia úderov sa však zníži:

$$f_{d0,3} = \frac{1}{30 \cdot 10^{-3}} = 33,33Hz \quad (3)$$



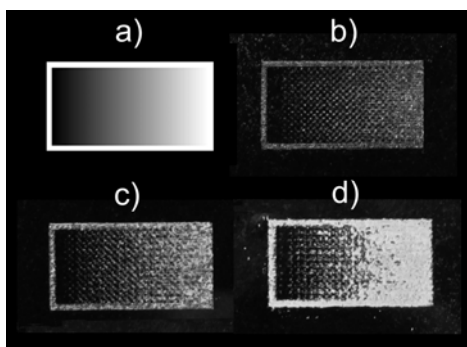
Obr. 7 Poloha hrotu pri úderoch so striedou 0,3. Budiaca sila je príliš malá na to aby hrot stihol prejsť celú dráhu až ku kamennej doske.



Obr. 8 Poloha hrotu pri úderoch so striedou 0,3. Znížením frekvencie úderov dosiahneme plný rozkmit hrotu.

Aby sme overili správnosť nastavenia jednotlivých časov pre rôzne frekvencie, vygravírovali sme jednoduchý obrazec (Obr. 9a). Ide o škálu v odtieňoch sivej s hĺbkou 8 bitov. Zľava doprava postupne rastie svetlosť od najtmavšej po najsvetlejšiu. Obrázok sme prekonvertovali na čiernobiely s hĺbkou 1 bit. Algoritmus pre konverziu sme zvolili podľa [4]. Ide

o poltónovanie so zoskupením bodov. Následne sme obrazec vygravírovali s rôznymi striedami a im prislúchajúcimi časmi zapnutia a vypnutia. Po gravírovaní sme obrázok pretreli bielou farbou, ktorá sa usadí v hlbších bodkách a tak zlepši kontrast medzi obrazom a kamennou doskou.



Obr. 9 Obrázok gravírovaný s rôznymi nastaveniami striedy a časov zapnutia a vypnutia. Mierka 1:1.

Na obrázku (Obr. 9b) je použitý slabší úder ($d=0,3$), na Obr. 9c) stredne silný úder ($d=0,5$) a na Obr. 9d) silný úder ($d=0,7$). Odstupňovanie svetlosti jednotlivých obrazov je jasne viditeľné. Body najtmavšieho obrazu (b) sú také jemné, že sa v nich neusadila žiadna farba. V prípade (c) sú svetlé tóny svetlejšie ako v (b). Prechod od tmavých tónov k svetlým je stále pomerne plynulý. Na obraze (d) sú svetlé tóny jednoznačne najsvetlejšie. Prechod medzi tmavými a svetlými tónmi je však veľmi nepravidelný a skokový. Vidieť, že v svetlých tónoch dochádza k saturácii ešte pred dosiahnutím najsvetlejšieho odtieňu. Je to spôsobené tým, že jednotlivé body sú tak veľké, že sa prelínajú a vytvárajú jednoliaty celok aj v oblastiach, kde by sa mal nachádzať istý počet tmavých bodov.

Z tohto experimentu sa dá usudzovať, že škálovanie gravírovaného obrazu sa dá vylepšiť použitím rôzne veľkej sily úderu. Aby sme dosiahli čo najširšiu škálu odtieňov, je vhodné použiť aj silné údery, ktoré robia body, kde sa usádza väčšie množstvo farby. Treba však odstrániť saturáciu svetlých tónov a zjemniť prechod medzi tmavými a svetlými tónmi.

IV. ZÁVER

Podrobnou analýzou a experimentovaním s predkladaným nástrojom sme získali dobrú predstavu o možnostiach použitia tohto nástroja na gravírovanie portrétov do kamenných dosiek. Došli sme k niekoľkým záverom:

1. Veľkosť a hĺbka gravírovaného bodu závisia hlavne od sily úderu, len málo od dĺžky trvania úderu.
2. Dĺžkou trvania úderu sa treba zaoberať najmä z hľadiska dosiahnutia čo najvyššej frekvencie úderov.
3. Použitie rôznej sily úderu pre rôzne odtiene sa pozitívne prejaví na celkovom kontraste gravírovaného obrazu, aj keď pri použití jednotnej sily úderu sa získajú jemnejšie prechody.

Pre všetky experimenty sme použili konštantnú vzdialenosť nástroja od dosky (1,3 mm). Zmenšením vzdialenosti môžeme dosiahnuť vyššiu frekvenciu úderov. Pre získanie rovnako svetlého bodu je však potrebná väčšia sila úderu. Vzdialenosť približne 1,3 mm sa javí ako vhodná aj preto, že kamenná doska nie je dokonale rovná.

Predkladaný nástroj je oproti iným riešeniam veľmi jednoduchý a lacný. Na rozdiel od diamantových fréz sa neopotrebuje (vydrží najmenej na 120 portrétov, fréza maximálne na 15). Nevýhodou oproti laseru je nižšie rozlíšenie, dané veľkosťou bodu.

LITERATÚRA

- [1] Z12 Monument Etching Laser, dostupné na internete: http://www.kernlasers.com/product_pages/z12.htm
- [2] G10 stone image engraving machine, dostupné na internete: <http://www.acmetek.biz/g10.shtml>
- [3] H. L. K. Johnson, "Improvements in or relating to Electrically Driven Engraving machines", GB536146, May 1941, Patent specification.
- [4] M. Batmend, "An automated headstone photo engraving" (proceedings from conference), 10th Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics Technical University of Košice, I. Edition, pp. 171–173, ISBN 978-80-553-0423-6.
- [5] R.C.Gonzalez,P.Wintz, „Digital image processing“ (second edition), November 1987, ISBN 0-201-11026-1

Dielektrická relaxačná spektroskopia 6 kV statorových cievok bez polovodivej ochrany

Ludovít CSÁNYI, Roman CIMBALA

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

ludovit.csanyi@tuke.sk, roman.cimbala@tuke.sk

Abstrakt — Tento článok sa zaoberá meraním cievok asynchrónneho 6 kV motora bez polovodivej ochrany pomocou dielektrickej relaxačnej spektroskopie. Celkovo bolo skúmaných desať reálnych cievok izolačného systému elektrického stroja s dvoma konfiguráciami elektródového systému. 2 bodový systém uloženia cievky v drážke statora a 4 bodový jej vývoj do čela vinutia motora. Meranie bolo uskutočnené v časovej oblasti pomocou elektrometra Keithley 617 pre meranie extrémne malých prúdov (10^{-16} A) a tvorí základ nedeštruktívnej jednosmernej meracej metódy.

Kľúčové slová — coil, dielectric spectroscopy, insulation system, IRC analysis, rotating machine

I. ÚVOD

V dnešnej dobe je rastúci dopyt po elektrickej energii, tento trend je spôsobený zvyšujúcim sa nárokom spoločnosti na vyšší životný štandard.

Z elektroenergetického hľadiska to znamená, že na vysokonapäťové elektrické stroje sú kladené vyššie nároky zo strany bezpečnosti a najmä spoľahlivosti. Tieto nároky sa vzťahujú najmä na elektroizolačný systém, ktorý patrí medzi najdôležitejšie časti elektrických strojov. Stále asi jedna štvrtina porúch súvisí so zlyhaním izolácie týchto strojov [1]. Jav, ktorý má na izoláciu najväčší vplyv a dá sa minimálne ovplyvniť je starnutie izolácie ktoré má podľa Obr. 1 na príčine vyše 30% porúch. Ako druhý rozhodujúci faktor porúch je porucha vo vinutí s podielom 25%.

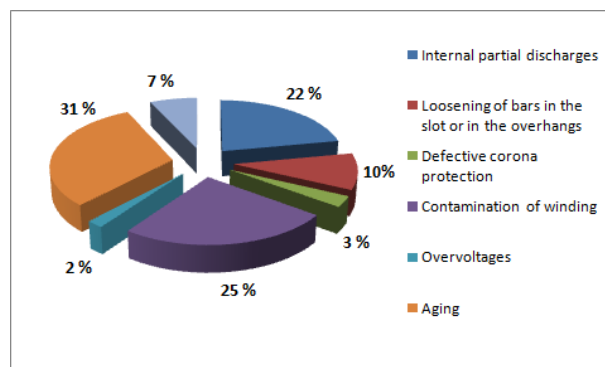


Fig. 1 Hlavné príčiny porúch izolačného systému [2]

Jednou z metód na zisťovanie stavu elektroizolačného systému je dielektrická relaxačná spektroskopia (DRS), ktorá patrí medzi nedeštruktívne metódy a je založená na širokom rozsahu meraní v časovej a vo frekvenčnej oblasti.

II. POLARIZÁCIE DIELEKTRIKA

Pri zisťovaní stavov izolačných systémov elektrotechnických zariadení sa zameriavame najmä na včasné zistenie zdrojov možných porúch týchto zariadení. Je nutné sledovať elektrické alebo fyzikálne veličiny, ktoré popisujú daný stav sledovaného zariadenia. Každá z

týchto veličín poukazuje na určité zmeny, ktoré sa výrazne prejavajú počas chodu elektrického zariadenia (stroja) a u ktorých nárast poukazuje na vhodnosť výberu elektrickej alebo fyzikálnej veličiny.

Treba povedať, že vývoj poznatkov o fyzikálnej podstate dejov v izolácii v elektrickom poli nedospel doteraz do štádia, v ktorom by bola v odbornej a vedeckej literatúre dostatočná zhoda názorov a jednota v názvoch dielektrickej polarizácie [3]. Podľa [3, 4] delíme polarizácie na:

Elektrónová polarizácia : Pri tejto polarizácii dochádza k posunutiu elektrónového obalu vzhľadom na jadro. Posunutím ťažiska nábojov na určitú vzdialenosť získa atóm indukovaný dipólový moment [5]. Elektrónová polarizácia má vysokú rýchlosť ustálenia (10^{-16} až 10^{-13} s).

Iónová polarizácia : Pôsobením vonkajšieho elektrického poľa dochádza k pružnému posunutiu iónov kryštalickej mriežky. Ióny sú v kryštalickej mriežke viazané v uzloch silou a preto sa nazývajú pružné. Doba ustálenia sa pohybuje medzi 10^{-13} až 10^{-12} s.

Iónová relaxačná polarizácia : V normálnom stave sú ióny uložené na nižších energetických hladinách. Vplyvom elektrického poľa dochádza ku prechodu iónov do ďalšieho zachytného bodu po prekročení potenciálovej bariéry a tým vzniká relaxačná polarizácia. Je spojená so stratou energie.

Dipólová relaxačná polarizácia : Pokiaľ na látku nepôsobí žiadne elektrické pole je orientácia jednotlivých dipólov náhodná. Ak vložíme látku do elektrického poľa dochádza k zmene orientácie v smere siločiar a tým ku vzniku dipólového momentu. Je spojená so stratou energie.

Migračné polarizácia: (polarizácia vznikajúca priestorovým nábojom): Čas ustálenie migračnej polarizácie je v rozsahu od 10^{-2} až 10^{-4} s. V minulosti bola táto polarizácia považovaná skôr za okrajovú a bola uvádzaná viac-menej pre úplnosť. V súčasnosti je naopak priestorový náboj predmetom intenzívneho výskumu [6, 7, 8].

Medzivrstvová polarizácia : K polarizácii dochádza v dôsledku nehomogenosti dielektrika. V dielektriku sa môže na povrchu rôznych izolačných materiálov alebo nehomogenít hromadiť ióny a tým vytvárať dipóly.

Hraničná polarizácia : Objemový náboj sa môže hromadiť pri elektródach a tak môže vytvárať polarizačný jav.

III. DIELEKTRICKÁ SPEKTROSKOPIA

Na diagnostiku elektrotechnických zariadení existuje niekoľko metód. Jednou s týchto metód je aj dielektrická relaxačná spektroskopia. Jedná sa o nedeštruktívnu metódu, ktorá je založená na širokom rozsahu meraní v časovej a frekvenčnej oblasti a spadá do oblasti na pomedzi elektrotechniky, fyziky, chémie a materiálového inžinierstva [9]. Vo všeobecnosti možno povedať, že DC vodivosťou, funkciou dielektrickej odozvy a vysokofrekvenčnou zložkou relatívnej permitivity možno charakterizovať správanie sa dielektrického materiálu [10].

Tento článok sa zaoberá len meraním časovej odozvy. Pri hodnotení časovej odozvy sa na hodnotenie systému používa polarizačný a depolarizačný prúd.

A. Spektroskopia v časovej oblasti

Ideálne dielektrikum tvorí látka, ktorá obsahuje elektrické náboje viazané elektrostatickými silami. V prípade ideálnych dielektrík však existujú aj voľné nosiče náboja.

Dielektrická polarizácia je jav, ktorý vzniká pri pôsobení elektrického poľa a spôsobuje pohyb voľných nosičov náboja.

Ak do elektrického poľa o intenzite $E(t)$ vložíme dielektrický materiál, tento materiál bude polarizovaný. Elektrická polarizácia $P(t)$ v dielektrickom materiály môže byť rozdelená na dve časti a to na rýchlu a na pomalú polarizáciu [10].

Pomalé polarizačné procesy sa merajú pomocou nabíjacieho a vybíjacieho prúdu.

Prúd $i(t)$, ktorý preteká dielektrikom po pripojení elektrického poľa s intenzitou $E(t)$, môže byť vyjadrený ako:

$$i(t) = C_0 \left[\frac{\sigma_0}{\varepsilon_0} u(t) + \varepsilon_\infty \frac{du(t)}{dt} + \frac{d}{dt} \int_0^t f(t-\tau) u(\tau) d\tau \right] \quad (1)$$

kde: C_0 – geometrická kapacita meraného objektu

Polarizačný prúd $i_{pol}(t)$ môžeme vyjadriť ako:

$$i_{pol}(t) = C_0 U_c \left[\frac{\sigma_0}{\varepsilon_0} + f(t) \right] \quad (2)$$

Polarizačný prúd je v dôsledku jednosmerného napätia U_C zložený z dvoch častí. Prvá časť súvisí z vodivosťou skúmaného objektu a druhá časť súvisí s aktiváciou rôznych polarizačných procesov vo vnútri skúmaného materiálu [10, 11].

Ak je jednosmerné napätie U_C v čase T odstránené a materiál je uzemnený, potom začne pretekať depolarizačný prúd $i_{depol}(t)$.

$$i_{depol}(t) = -C_0 U_C [f(t) - f(t + T_c)] \quad (3)$$

kde: $f(t)$ – monotónne klesajúca funkcia

Druhú časť rovnice môžeme zanedbať pre veľké hodnoty T_c . Depolarizačný prúd sa stane úmerným k dielektrickej odozve [12]:

$$f(t) \approx -\frac{i_{depol}}{C_0 U_C} \quad (4)$$

Jednosmernú konduktivitu σ_0 je možné odhadnúť z rovníc (2) a (3):

$$\sigma_0 \approx \frac{\epsilon_0}{C_0 U_C} [i_{pol}(t) + i_{depol}(t)] \quad (5)$$

IV. EXPERIMENT

Meranými objektmi boli 6 kV satorové cievky bez polovodivej ochrany (Obr. 2). Účelom experimentu bolo zistiť vplyv konfigurácie elektródového systému vytvoreného na skúmaných cievkach satorového vinutia na výsledky dielektrickej spektroskopie v časovej oblasti. Technológia uloženia materiálu pre rovnú časť systému, ktorá je uložená v drážke (4 bodový elektródový systém) a pre časť naväzujúcu na čelo vinutia (2 bodový elektródový systém) je rozdielna.



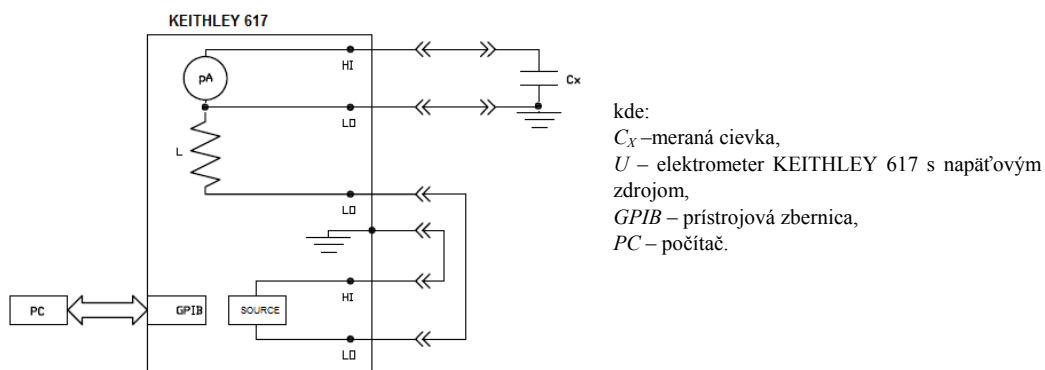
Obr. 2 Satorová cievka 6 kV

Technické parametre motora, pre ktoré boli cievky vyrobené sú: výkon - 200kW, max. otáčky - 740 ot/min, napätie - 6000 V, prúd - 27,6 A, počet závitov - 13 závitov, dĺžka drážky - 590mm.

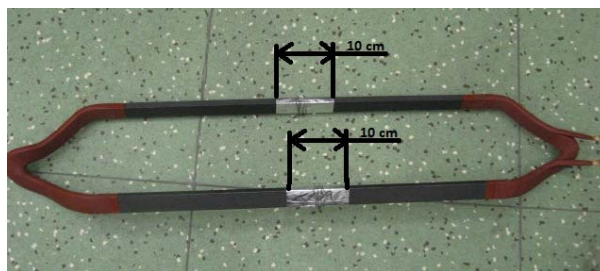
A. Popis merania

Cieľom merania bolo porovnať závislosť nabíjacieho prúdu pri dvoch elektródových systémoch (2 bodový a 4 bodový elektródový systém), ktoré sú rozmerovo rovnaké ale majú rozdielny spôsob technológie uloženia. Počet meraných cievok bol 5.

Na meranie bola použitá metóda dielektrickej relaxačnej spektroskopie. Pri meraní bol použitý už spomínaný elektromer KEITHLEY 617 [13] a softvér Agilent HP VEE Pro. Schéma merania je zobrazená na (Obr. 3).

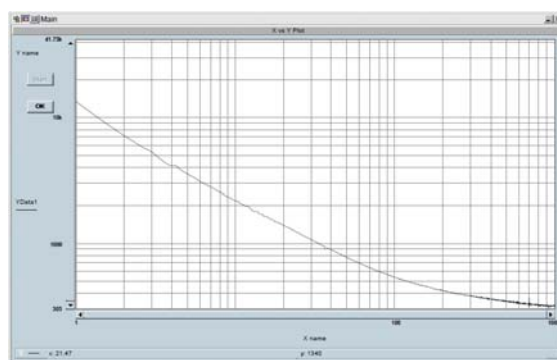


Obr. 3 Schéma zapojenia meracieho systému



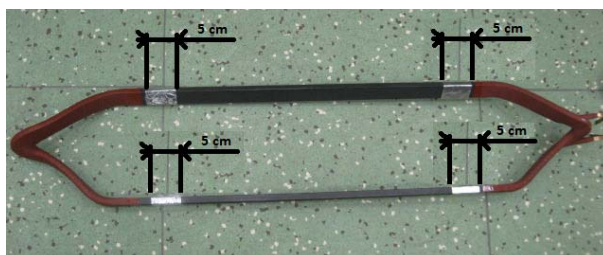
Obr. 4 Dvojbodový elektródový systém

Pri 2 bodovom elektródovom systéme (Obr. 4) bola na strednú časť cievky navinutá vodivá vrstva s dĺžkou 10 cm ($\pm 0,2$ cm), ktoré bola pre potreby merania skratovaná a uzemnená. Na grafe na Obr. 5 je vidieť závislosť cievky č. 2 pri dvoj bodovej uzemňovacej elektróde.



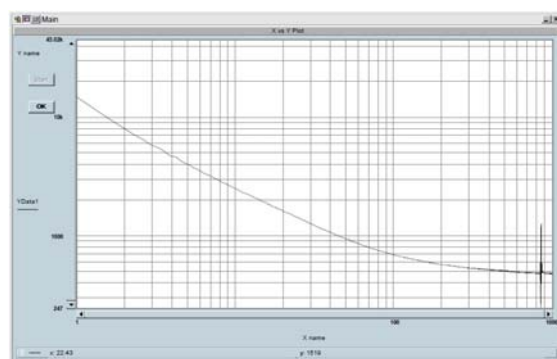
Obr. 5 Cievka č. 2 - 2 body uzemnenia

Nasledovalo merania zo štyrmi bodmi uzemnenia (Obr. 6). Pri 4 bodovom uzemnení sa na konci drážkovej časti navinul vodivý materiál o šírke 5 cm ($\pm 0,2$ cm) a tieto body následne skratované. Zabezpečila sa tým rovnaká kapacita elektródového systému, nakoľko efektívna plocha pokrývajúcej sa časti uzemňovacích elektród a medeného jadra cievky tvoriaceho druhú elektródu bola rovnaká pre oba prípady.



Obr. 6 Štvorbodový elektródový systém

Na nasledujúcom obrázku je vidieť závislosť prúdovej odozvy na čase pre štvorbodový elektródový systém (Obr. 7).



Obr. 7 Cievka č. 2 - 4 body uzemnenia

Celkový polarizačný prúd je možné na základe Debyeovej teórie o polarizácii v dielektriku nahraďovať sumou nezávislých elementárnych prúdov, ktoré reprezentujú jednotlivé druhy polarizácie vyskytujúce sa v materiáli podľa rovnice (6)

$$i_{pol}(t) = \frac{U}{R_i} + \sum_{i=1}^n I_{mi} \exp\left(\frac{-t}{\tau_i}\right) \quad (6)$$

Počet členov rozvoja n je závislý od štruktúry a homogenity dielektrika a od času, v ktorom sa polarizačné deje merajú. V tabuľke Tab. 1 je uvedené číselné zobrazenie rozvoja nameraného polarizačného prúdu pri aproximácii do rozvoja pre sedem členov.

Tabuľka 1
 ROZVOJ PRÚDOVEJ ODOZVY DO 7 DEBYEOVÝCH ELEMENTOV

2 body		4 body	
τ_i (s)	I_{mi} (pA)	τ_i (s)	I_{mi} (pA)
0.18	178128	0.16	213311
0.73	21306	0.67	28804
2.92	6100	2.68	7456
14.57	1950	12.85	2365
52.86	617	48.42	955
308.45	185	251.79	223
1090906.64	310	113291.63	483

V. ZÁVER

Výsledkom práce je porovnanie dvoch rôzne umiestnených elektródových systémov na 6 kV statorovej cievke pomocou dielektrickej relaxačnej spektroskopie. Dielektrická relaxačná spektroskopia bola uskutočnená pomocou prístroja KEITHLEY 617. Získané výsledky predstavujú nabíjacie prúdy a z nich získaná relaxačné časy a elementárne prúdy. Výsledky sa nachádzajú v časovej oblasti. Vzhľadom na to, že hodnoty elementárnych polarizačných prúdov sú závislé od veľkosti meraného objektu a majú priamy vzťah na jeho kapacitu, pre analýzu kvality materiálu je podstatná analýza časových konštánt ustálenia elementárnych polarizačných prúdov. Porovnaním časových konštánt z Tab. 1 vyplýva, že geometrické usporiadanie elektródového systému nemá vplyv na ich hodnotu. Zaznamenaný rozptyl je v súlade s výsledkami získanými na merania v reálnej praxi [14]. Technológia použitá pre rovnú časť cievky a pre časť naväzujúcu na čelo vinutia nemá vplyv na výsledky dielektrickej spektroskopie v časovej oblasti. Uvedené konštatovanie je rozhodujúce pre navrhovanie diagnostických systémov izolačných strojov, nakoľko pri aplikovaní vyššieho skúšobného napätia je dvojelektrodový systém výhodnejší pre meranie závislostí, kde je riziko vzniku povrchových výbojov. Je teda možné jednu a tú istú konfiguráciu elektródového systému použiť pre meranie jednosmerných charakteristík skúmajúcich polarizačné deje v časovej oblasti a aj pre meranie striedavých závislostí aplikovaním vysokého napätia, akým je napr. meranie fázovej analýzy čiastkových výbojov.

POĎAKOVANIE

Tento článok bol vypracovaný s podporou projektu Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy, ITMS 26220220029, ktorý je spolufinancovaný zo štrukturálneho fondu EÚ ERDF v rámci výzvy OPV a V-2008/2.2/01-SORO a prioritnej osi 2 Podpora výskumu a vývoja.

LITERATÚRA

- [1] Thorsen, O.V. Dalva, M.: "A survey of faults on induction motors in offshore oil industry, petrochemical industry, gas terminals and oil refineries," IEEE Trans. Ind. Appl. vol. 31, no. 5, pp. 1186-1196, Sep./Oct. 1995.
- [2] CIGRE Study Committee SC11, EG11.02, "Hydrogenerator Failures – Results of the Survey," 2003.
- [3] Hassdenteufel, J. – kol.: Elektrotechnické materiály, ALFA SNTL, Praha, 1978.
- [4] Marton, K.: Elektrotechnické materiály (Fyzika dielektrík), ESŠT, Košice, 1971.
- [5] Mentlík, V.: Dielektrické prvky a systémy, BEN – technická literatúra, Praha, 2006, ISBN 80-7300-189-6.

- [6] Mizutani, T.: Space charge measurement techniques and space charge in polyethylen , IEEE Transaction on Dielectrics Electrical Insulation, Vol. 1, No. 5, pp. 923-933, 1994.
- [7] Suzuoki, Y. aj.: Study of space-charge effects on dielectric breakdown of polymers by direct probing, IEEE Transactions on Electrical Insulation, Vol. 27, No. 4, 758 – 762, 1944.
- [8] Mentlík, V. – Lang, J.: Studium polarizačních jevů v nehomogenním dielektriku. (Závěrečná výzkumná zpráva úkolu 0432 FR MŠMT) ZČU Plzeň 1998.
- [9] Palai-Dany, T.: Dielektrická spektroskopie karboxymetylcelulózy v časovej oblasti, Zkrácená verze Ph.D. Thesis, 2009, CVUT Brno, FEKT, Ústav fyziky.
- [10] Farahani, M., Borsi, H., Gockenbach, E. Dielectric spectroscopy in Time and Frequency Domain on Insulation System of High Voltage Rotating Machines. In 8th IEEE International Conference on Solid Dielectrics (ICSD). Toulouse, France, 2004.
- [11] Walter S. Zaengl.: Dielectric Spectroscopy in Time Domanie for HV Power Equipment, Part I: Theoretical Considerations, IEEE, vol. 19, no. 5, september-october 2003.
- [12] V. Der Houhanessian, „Measurement and Analysis of dielectric response in Oil-Paper Insulation System“, PhD thesis, Swiss Federal Institute of Technology, ETH, No. 12832, Zurich, 1998.
- [13] Model 617 Programmable Electrometer, Instruction manual, Keithley Instruments, Inc., Cleveland, Ohio, USA, 617-901-01 Rev.G, 1984.
- [14] R. Cimbala, “Ageing of high voltage insulation systems”, book, TU Košice 2007.

Jednosmerné charakteristiky nanokeramík

Juraj KURIMSKÝ

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

juraj.kurimsky@tuke.sk

Abstrakt — V príspevku je popísaný princíp javu jednosmernej odozvy nanokeramík. Sú uvedené niektoré používané spôsoby prípravy a sintrácie keramik, ktoré výrazne vplyvajú na elektrické vlastnosti ZnO a SiC a princíp merania jednosmerných charakteristík sintrovaných keramik. Sú prezentované experimentálne výsledky z meraní nabíjajúcich prúdov v ZnO vzorkách. Získané výsledky sú dôležitými poznatkami pri sledovaní makroskopických prejavov zmien v kryštalickej štruktúre ZnO, ale aj SiC varistorov.

Kľúčové slová — sintrovaná karamika, jednosmerné charakteristiky, ZnO

I. ÚVOD

Vývoj materiálov používaných v elektroenergetike v súčasnosti umožňuje ich maximálne využitie aj v extrémnych podmienkach. Preto je dôležité sledovať parametre poukazujúce na ich aktuálny stav, a to najmä z hľadiska zvyškovej životnosti.

Sintrovaná keramika sa vyrába práškovou metódou a jeho názov je odvodený od spôsobu prípravy sintráciou a vlastností sintrovaných polovodičov. Technologický postup je založený na spracovaní predmetov z prášku. Základný materiál pomelie na jemný prach, následne sa zahrieva nad bod tavenia (sintrácia pevných materiálov). Sintrácia sa najčastejšie využíva pri výrobe keramických predmetov a takisto aj pri práškovej metalurgii. Práca je orientovaná na chemické prvky zinok a kremík, pretože sú základnými prvkami sintrovaných keramik. Najznámejším polovodičovým materiálom je kryštalická forma prvku zo skupiny IV periodickej tabuľky prvkov, kremík Si. Ďalším známym polovodičom z tejto skupiny je germánium Ge a perspektívny výskum prebieha na zlúčeninách SiGe a SiC. Významné miesto majú zlúčeniny pozostávajúce z prvkov, ktorých fyzikálne vlastnosti ich predurčujú na použitie v rýchlej elektronike a optoelektronike [1]. Základná štruktúra a vlastnosti sú v ďalšom texte vysvetlené na príklade oxidu zinočnatého, ZnO.

A. Oxid zinočnatý

Varistory vyrobené na báze ZnO klasifikujeme podľa šírky zakázaného pásma ako polovodiče. ZnO varistory sa vyznačujú nelineárnou závislosťou prúdu od napätia; jedná sa o vodivosť typu n (nosičmi náboja sú pohybujúce sa elektróny vo vodivostnom pásme). Nelineárne vlastnosti ZnO varistora sa využívajú v zariadeniach na ochranu pred prepätiami. Vyrábajú sa na báze cca 90% oxidu zinočnatého (ZnO) vo funkcii keramického základu a 10% prísad, ktoré slúžia na rast zŕn a tvorbu hradlovej vrstvy medzi zŕnami ZnO. Pri zlisovaní tejto hmoty sa okolo dobre vodivých zŕn ZnO vytvárajú hradlové vrstvy, ktoré sú zrovnateľné so zenerovými diódami. Ich odpor ($>1\text{M}\Omega$) sa pri vzniku prepätí znižuje v priebehu niekoľkých nanosekúnd až na hodnoty niekoľko desiatín ohmu. Využitím celého objemu keramiky pre absorpciu energie je daná vysoká zaťažiteľnosť tohto typu zvodíčov pri ich zaťažovaní impulznými zvodovými prúdmi. Takmer univerzálne možnosti využitia varistorov sú obmedzené iba v oblasti vysokých frekvencií, kde sa negatívne prejavuje ich relatívne vysoká kapacita (jednotky nF). Bolo poukázané na procesy degradácie takto pripravovaných keramik [3].

Oxid zinočnatý ZnO je polovodič s priamym optickým prechodom a charakteristickou šírkou zakázaného pásma. ZnO kryštalizuje v hexagonálnej kryštalickej štruktúre. Prítomnosť elektrónov v nedopovanom ZnO je pripisovaná atómom Zn, kyslíkovým obohatením, alebo

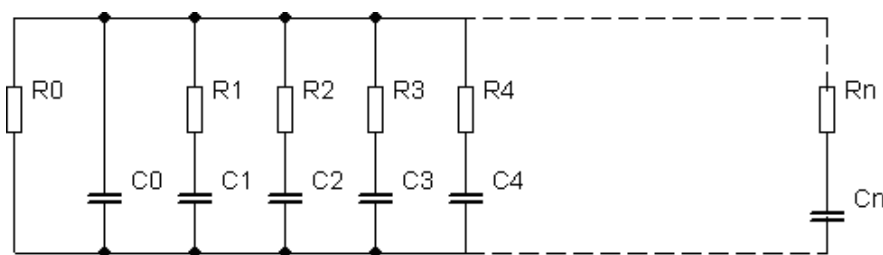
prítomnosti vodíka. Na obohatenie štruktúry ZnO sa používajú najčastejšie prvky Al, Ga alebo In. Vytvorenie *p*-typu polovodiča je problematické, a to nielen pre ZnO, ale aj pre ZnSe (Selén zinočnatý). Podľa principiálnych odhadov by mohli prvky P a As slúžiť ako dopanty pre vytvorenie *p*-typu ZnO. Rôznymi optickými metódami sa potvrdil priamy optický prechod a silne viazaný excitónový stav, kde atómové jadro má vyšší energetický stav ako daný atóm. Zároveň plnia funkciu polovodivej vrstvy spoluvytvárajúcej oblasť ochudobnenia a vstavaný potenciál so vznikom PN prechodu [2].

V práci sú analyzované nabíjacie charakteristiky ZnO keramik, je popísaný merací systém, ktorý bol testovaný na pracovisku. Výsledky, archivované v digitálnej forme, sú analyzované a prezentované v grafickej podobe.

II. EXPERIMENT

A. Jednosmerné charakteristiky

Jednosmerná charakteristika je široký pojem, ktorý môže zahŕňať niekoľko elektrických meracích metód (napr. U-I a E-J charakteristiky a pod.). V tomto texte budeme pod pojmom jednosmerná charakteristika materiálu rozumieť prúdovú odozvu systému na napäťový skok. Princíp merania je popísaný mnohými autormi, pozri napr. [4], [5] a [6]. Pri týchto meraniach predpokladáme, že merané vzorky majú vlastnosti lineárnych dielektrík. Je to nevyhnutná požiadavka pre správnu interpretáciu nameraných hodnôt. Základná myšlienka spočíva v Maxwellovom-Wagnerovom modeli dielektrika [5], ktorý interpretuje dielektrikum ako náhradnú schému zo sústredenými parametrami, pozri obrázok 1. Jednotlivé vetvy tohto náhradného modelu predstavujú možné polarizačné mechanizmy, v ktorých je zahrnutá informácia o čase polarizácie. Polarizačný dej v *i*-tej vetve náhradnej schémy je určený časovou konštantou $t_i = R_i \cdot C_i$. Jednotlivé členy *R-C* prezentujú polarizáciu s dobou ustálenia. Pre účely modelovania sa v praxi používa sedemvetvový náhradný model.



Obr. 1 Maxwell-Wagnerov model

B. Príprava vzoriek

Pre úpravu vlastností vzoriek bol navrhnutý a realizovaný systém pre urýchlené starnutie, ktorý spĺňa nasledujúce požiadavky:

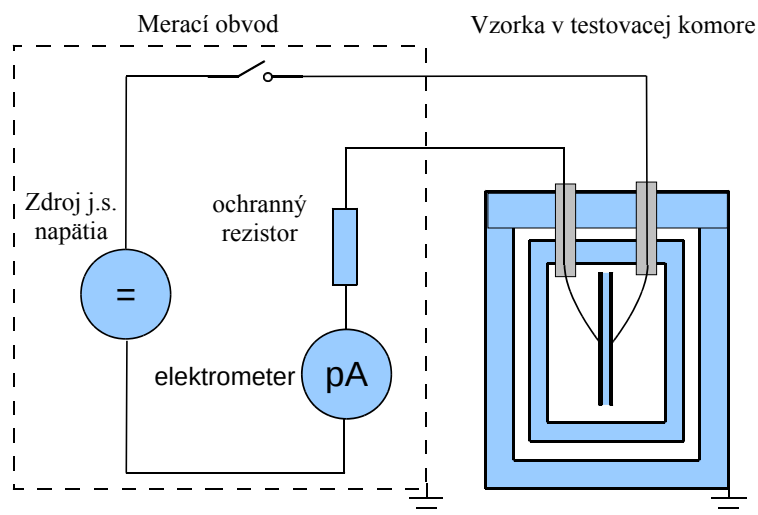
- zabezpečenie proti prehriatiu média a skratu odolnosť pomocou riadiacej elektroniky,
- reguláciu teploty média od 20°C do 110°C,
- riadenie teploty pomocou termočlánku, napr. PTC,
- použitie bezpečnostného teplomeru vo funkcii tepelného relé,
- nádoba s vyhrievacím telesom pre temperovanie média.

C. Požiadavky na meranie jednosmerných charakteristik

Celkové požiadavky na merací systém vyplývajú z hore uvedených skutočností a sú zhrnuté do nasledujúceho zoznamu:

- zdroj jednosmerného napätia so vzdialenou reguláciou výšky napätia,
- merná nádoba (merná komora) s dvojitém tienením,
- vákuové spínacie relé s ovládacou elektronikou cez PC,
- presnosť aparatury na úrovni 1pA,
- zabezpečenie elektromagnetickej kompatibility a odrušenia EM polí,
- galvanické oddelenie koncového meradla.

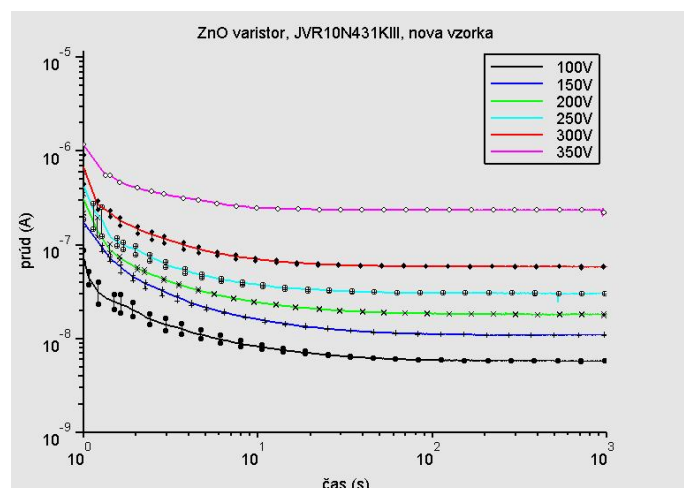
Zapojenie meracieho obvodu je na obr. 2.



Obr. 2 Zapojenie pre meranie jednosmerných nabíjajúcich charakteristík sintrovaných keramik

III. VÝSLEDKY MERANÍ

Popísanou metódou bolo umožnené zmerať nabíjacie charakteristiky na vzorkách ZnO sintrovanej keramiky. Príklad grafickej analýzy vyjadrujúcej časovú závislosť priebehu nabíjacieho prúdu pri rôznych hladinách nabíjacieho napätia je na obr. 3.



Obr. 3 Nabíjacie charakteristiky novej ZnO vzorky

Celkové vyhodnotenie experimentu je v tabuľke 1, pričom E predstavuje intenzitu elektrického poľa v objeme vzorky, I_{pmax} je maximálna hodnota nabíjacieho prúdu a I_{ps} je ustálená hodnota nabíjacieho prúdu.

Tabuľka 1
 MAXIMÁLNE A USTÁLENÉ HODNOTY NABÍJACÍCH PRÚDOV ZNO VZORIEK

ZnO E (V/m)	Nová vzorka		starnuté 330 hodín		starnuté 1168 hodín	
	I_{pmax} (A)	I_{ps} (A)	I_{pmax} (A)	I_{ps} (A)	I_{pmax} (A)	I_{ps} (A)
5,03E+04	7,39E-8	5,78E-9	7,31E-8	5,78E-9	6,95E-8	5,26E-9
7,55E+04	1,75E-7	1,09E-8	1,27E-7	1,10E-8	1,23E-7	9,90E-9
1,01E+05	3,13E-7	1,80E-8	2,44E-7	1,73E-8	1,93E-7	1,59E-8
1,26E+05	4,38E-7	3,03E-8	5,17E-7	2,80E-8	3,39E-7	2,57E-8
1,51E+05	6,84E-7	5,83E-8	1,82E-7	4,88E-8	3,96E-7	4,64E-8
1,76E+05	1,17E-6	2,17E-7	8,21E-7	1,75E-7	8,38E-7	1,62E-7

IV. DISKUSIA

Pozorovaním rôznych parametrov je možné sledovať proces starnutia sintrovaných keramik. Týmto sa dá odhaliť prechodné zlepšenie elektrických vlastností v procese starnutia. Experiment poukazuje na to, že tisíc hodín urýchleného starnutia môže byť v prípade sledovaných vzoriek považovaný za krátkodobé kondicionovanie. Avšak sa predpokladá, že ďalšie – dlhodobé – starnutie má za následok zmenu vnútornej štruktúry, ktorá sa v makroskopickom pohľade prejaví zvýšením meraných hodnôt nabíjajúcich prúdov.

V. ZÁVER

Podľa predpokladov, ZnO vzorky vykazujú vlastný, charakteristický tvar priebehov nabíjajúcich prúdov. Urýchlené starnutie, ktoré trvá do 1000 hodín, nespôsobuje dramatické zmeny v parametroch týchto priebehov. Analýza poukázala na to, že ZnO keramiky, ktoré sú čoraz častejšie používané v elektrických zariadeniach, neprispievajú negatívne na ich celkovú životnosť a spoľahlivosť. Avšak v tomto experimente sa neuvažovali prúdové a prepäťové rázy, ktoré musia byť zahrnuté do komplexnej analýzy životnosti týchto zariadení.

POĎAKOVANIE



Tento článok bol vypracovaný s podporou projektu Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy, ITMS 26220220029, ktorý je spolufinancovaný zo štrukturálneho fondu EÚ ERDF v rámci výzvy OPV a V-2008/2.2/01-SORO a prioritnej osi 2 Podpora výskumu a vývoja.

LITERATÚRA

- [1] Ballato A. Piezoelectricity: old effect, new thrusts., IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control. DOI – 10.1109/58.464826. (1995); 42(5):916–926.
- [2] Norton, D. et al, ZnO: Growth, doping & processing, In Materials today, June 2004, pp. 34-40.
- [3] Mao-hua Wang, Chao Ya, Nan-fa Zhang, Degradation characteristics of low-voltage ZnO varistor manufactured by chemical coprecipitation processing. Department of Chemical Engineering, Jiangsu Polytechnic University, Changzhou 213164, Jiangsu, PR China Changzhou Chuangjie Lightning Protection Co, Ltd., Changzhou 213016, Jiangsu, PR China, (2007).
- [4] Harden, J., Martzloff, A., Moris, K., Golden, M., Matsuoka, M. Littelfuse varistors- Basic properties, terminology and theory, Application note, July 1999 AN9767.1 Junction studies at ZnO Varistors & Experiment.
- [5] Cimbala, R. Starnutie vysokonapäťových izolačných systémov, Technická univerzita v Košiciach, (2007), str. 140-160. ISBN 978-80-8073-904-1
- [6] Artbauer, J., Šedovič, J., Adamec, V. Izolanty a izolácie, SNTL Bratislava, (1969) str. 62-132.

Modelovanie výbojovej činnosti na vn cievke 6kV asynchrónneho motora

Milan KVAKOVSKÝ, Irida KOLCUNOVÁ, Juraj KURIMSKÝ

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

milan.kvakovsky@tuke.sk, iraida.kolcunova@tuke.sk, juraj.kurimsky@tuke.sk

Abstrakt — S rastúcim inštalovaným výkonom rastie aj dôležitosť na bezporuchovú prevádzku elektrických strojov, pretože prípadná porucha, havária alebo nútená odstávka by spôsobila veľké energetické i finančné straty. Výrobcovia a užívatelia majú preto snahu zaistiť čo najväčšiu spoľahlivosť a bezpečnosť svojich strojov použitím najdokonalejších a najprogressívnejších diagnostických metód. V tomto článku sa popisuje realizácia merania čiastkových výbojov priamou metódou na 6kV cievkach točivého elektrického stroja. Na meranie bolo použitých päť nových statorových cievok bez polovodivej ochrany. Článok rieši rozdiel vo výbojovej činnosti týchto cievok.

Kľúčové slová — izolácia, čiastkové výboje, drážka, elektrický stroj točivých

I. ÚVOD

Pri prevádzkovaní strojov, u ktorých ich zlyhanie je sprevádzané buď obrovskými následnými stratami v produkcii alebo finančnými stratami, je veľmi dôležitá diagnostika. Pri diagnostike elektrických strojov točivých treba brať na zreteľ to, že k najviac namáhaným častiam patrí statorová izolačná sústava. Za prevádzky je vystavená kombinácii namáhania elektrického, mechanického, tepelného a chemického. Tieto degradačné vplyvy spôsobujú postupne zhoršovanie všetkých elektrických a mechanických vlastností izolácie, ktorá v krajnom prípade prestane plniť svoju funkciu a dôjde k havárii stroja. Aby sa týmto havarijným stavom dalo predchádzať, vykonávajú sa diagnostické merania, umožňujúce sledovať postup starnutia izolácie. Ak majú byť výsledky týchto meraní výpovede schopné, musia sa tieto merania periodicky opakovať pokiaľ možno pri rovnakých podmienkach merania [1]. Zlyhanie vysokonapäťovej izolácie môže nastať z rozličných dôvodov:

- izolácia je vplyvom prevádzkových namáhání degradovaná až na hranicu svojej elektrickej pevnosti,
- elektrický prieraz nastane v priamom dôsledku mechanického poškodenia,
- elektrický prieraz nastane v dôsledku zvýšenej intenzity v miestach lokálnych nehomogenít elektrického poľa,
- elektrický prieraz nastane v dôsledku výbojovej činnosti v dutinkách v izolácii, ktoré mohli vzniknúť už pri výrobe, v dôsledku prevádzkového namáhania alebo ako výsledok veľkého lokálneho elektrického namáhania. Ak sa zvýši napätie priložené na meraný objekt, objavia sa pri určitom napätí v týchto dutinkách výboje. V miestach so zníženou elektrickou pevnosťou vplyvom vysokej intenzity elektrického poľa v izolácii vznikajú čiastkové výboje [2]. Čiastkové výboje sú výsledkom koncentrácie lokálneho elektrického namáhania v izolácii alebo na povrchu izolácie a vytvárajú prírodné (napäťové) impulzy s dobou trvania menej ako 1 μ s. Tieto výboje predstavujú prieraz v plynom vyplnených dutinkách. Inými slovami, čiastkové výboje premostujú vzdialenosť medzi dvoma elektródami len čiastočne a zvyšok izolácie je schopný vydržať celé prevádzkové napätie [3]. Všeobecne sa jedná o zhoršenie elektroizolačných vlastností, ktoré sa prejaví ako elektrický prieraz v mieste, kde sú k tomu najpriaznivejšie podmienky [4].

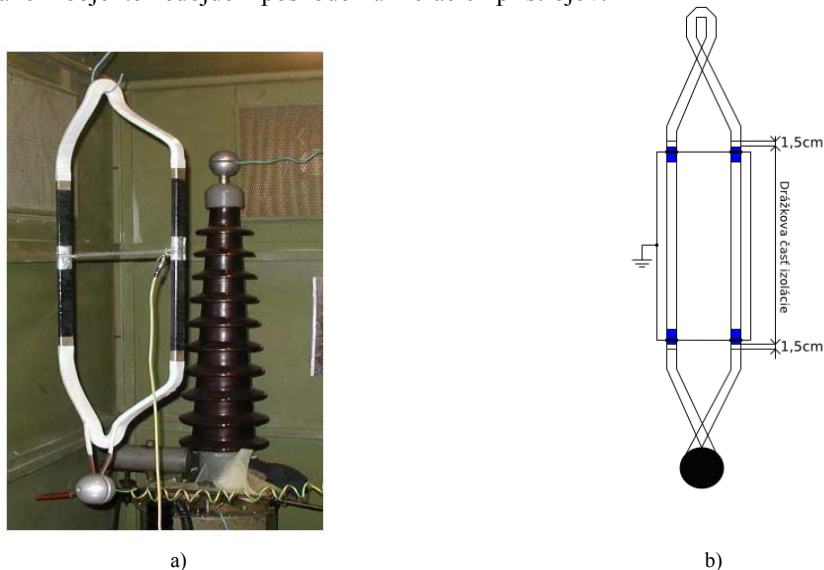
II. METÓDY MERANIA

Meranie čiastkových výbojov v izolácii statorových vinutí elektrických strojov je jednou s diagnostických metód, poukazujúcich na kvalitu izolačného systému. Aby boli výsledky meraní dôveryhodné je potrebné dokonale poznať skúmaný objekt, miesta vzniku výbojovej činnosti a

správne namodelovať poruchový dej. V prípade statorovej izolácie sa môžu objaviť vnútorné alebo vonkajšie čiastkové výboje [5].

V našom prípade sa na skúmanie, vzniku čiastkových výbojov rozvíjajúcich sa pri výstupe cievky z drážky, použilo päť novonavínutých 6kV statorových cievok, bez polovodiovej ochrany, s menovitým fázovým napätím $U_n=3,6\text{kV}$, menovitým prúdom $I_n=27,6\text{A}$ a menovitým výkonom $P_n=200\text{kW}$. Hlavná izolácia meraných cievok bola Samicatherm 366.28, ide o kalcínovo-sľudovú izoláciu spájanú epoxidovou živickou navínutou diskontinuálne. Cievka bola zavesená vo Faradayovej kletke na izolačnom lane (obr. 1a). Konce vinutia sa prepojili homogenizačnou elektródou a uviedli sa na vysokonapäťový potenciál. Drážková časť izolácie bola ovinutá v štyroch špecifických miestach alobalom o šírke 5cm, pričom jednotlivé miesta boli galvanicky spojené vodivou sieťovinou a vyvedené na potenciál zeme (obr. 1b).

Pri meraní bola použitá priama galvanická metóda merania čiastkových výbojov. Blokovaná schéma použitého zapojenia je na obr. 2. Výhodou tohto zapojenia je, že v prípade prirazu na meranom objekte nedôjde k poškodeniu meracích prístrojov.

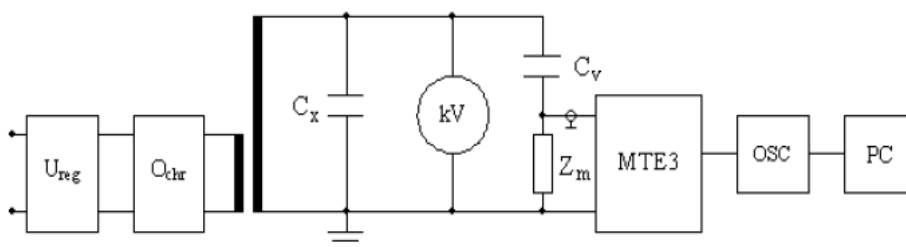


Obr. 1 a) umiestnenie skúšanej cievky pri meraní čiastkových výbojov,
b) uzemnenie cievky prostredníctvom štyroch bodov

III. SCHÉMA ZAPOJENIA

Celý merací systém sa skladá z troch častí:

- vysokonapäťovej: VN transformátor, väzobný kondenzátor CV, skúšaná vzorka CX,
- meracej: meracia impedancia Z_m , merač čiastkových výbojov MTE 3,
- vyhodnocovacej: pamäťový osciloskop, počítač so softvérom.



Obr. 2 Schéma zapojenia pre meranie čiastkových výbojov

Jedná sa o meranie čiastkových výbojov pomocou meracej impedancie, ktorá je zapojená do série s väzobným kondenzátorom. Signály sú registrované použitím väzobného kondenzátora a meracej impedancie. Tvar a poloha impulzov čiastkových výbojov sa sleduje osciloskopom. Každý čiastkový výboj je krátky prúdový impulz šíriaci sa po statorovom vinutí. Čím je tento výboj väčší, tým je väčšia výška impulzu, ktorý sa dosiahne na konci vinutia. Po úprave a zosilnení sú čiastkové výboje zobrazené na osciloskope. Vzťah medzi nábojom, ktorý sa šíri po vinutí a zdanlivým nábojom meraným na konci vinutia je ťažko definovať, pretože cesta prenosu signálu od miesta vzniku až do meracieho bodu môže byť zložitá. No vo všeobecnosti platí, že výška impulzu je priamo úmerná hodnote zdanlivého náboja čiastkového výboja [6].

IV. POSTUP MERANIA

Prvé testovacie napätie sa nastavilo na hodnotu 2kV a pôsobilo na cievku 3 minúty, ak sa pri tomto napätí neobjavila výbojová činnosť hodnota napätia sa zvýšila o 200V s rovnako dlhou dobou pôsobenia. Napätie sa postupne zvyšovalo až po to napätie, kým nevznikli prvé stabilné výboje (napätie počiatočných výbojov), kedy bol spustený program na zaznamenávanie výsledkov merania a uskutočnené prvé meranie čiastkových výbojov. Ďalšie merania sa uskutočnili pri zvyšovaní napätia krokom 200V až po nominálnu hodnotu združeného napätia skúšanej cievky 6kV. Každé meranie trvalo 3 minúty, za ktoré bolo zaznamenané viac než 900 periód aplikovaného napätia. Zvyšovaním napätia krokom 200V od počiatočnej hodnoty vzniku čiastkových výbojov až po nominálnu hodnotu 6kV boli získané napäťové závislosti charakteristických parametrov čiastkových výbojov.

V. NAMERANÉ HODNOTY A ICH DISKUSIA

Na porovnanie nameraných výsledkov zo všetkých vykonaných meraní sa určili tri napäťové hladiny (napätie počiatočných výbojov, nominálne fázové napätie 3,6kV a nominálne združené napätie 6kV) pre ktoré sme vykonali vyhodnotenie. Merania boli vykonávané od napätia počiatočných výbojov po nominálne združené napätie cievky. V tabuľke 1 sú uvedené maximálne hodnoty zdanlivého náboja čiastkových výbojov $q_{\max}$, početnosť čiastkových výbojov n_{str} a fázový uhol vzniku čiastkových výbojov v kladnej polovlne $\varphi+$ a zápornej polovlne $\varphi-$.

Tabuľka 1
 NAMERANÉ HODNOTY

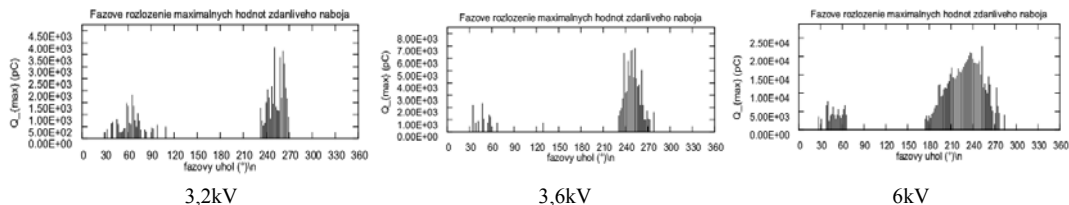
Cievka č.1				
U [kV]	$q_{\max}$ [pC]	n_{str} [-]	$\varphi+$ [°]	$\varphi-$ [°]
3,2	3000	0,14	30-90	230-270
3,6	7000	0,25	30-60	230-270
6	22500	1,5	30-60	180-300
Cievka č.2				
U [kV]	$q_{\max}$ [pC]	n_{str} [-]	$\varphi+$ [°]	$\varphi-$ [°]
2,6	60	0,2	30-70	220-270
3,6	2200	0,15	40-70	210-270
6	16000	0,9	40-60	195-290
Cievka č.3				
U [kV]	$q_{\max}$ [pC]	n_{str} [-]	$\varphi+$ [°]	$\varphi-$ [°]
2,6	300	0,1	30-80	230-260
3,6	4000	0,45	-	210-300
6	19000	1	30-80	180-300
Cievka č.4				
U [kV]	$q_{\max}$ [pC]	n_{str} [-]	$\varphi+$ [°]	$\varphi-$ [°]
2,8	140	0,25	30-90	220-240
3,6	3000	0,08	30-80	210-270
6	20000	0,35	30-80	180-270
Cievka č.5				
U [kV]	$q_{\max}$ [pC]	n_{str} [-]	$\varphi+$ [°]	$\varphi-$ [°]
3,4	1200	0,06	30-80	240-280
3,6	4500	0,17	30-80	220-270
6	19500	1,4	30-80	180-270

Na základe nameraných výsledkov je možné skonštatovať, že hodnota počiatočného napätia (Tab. 1) bola najvyššia pri cievkach č. 1 $U_{\text{poč}} = 3,2\text{kV}$ s maximálnou hodnotou zdanlivého náboja $q_{\max}=3000\text{pC}$ a pri cievke č. 5, kde $U_{\text{poč}} = 3,4\text{kV}$ s maximálnou hodnotou zdanlivého náboja $q_{\max}=1200\text{pC}$. Hodnota počiatočného napätia pri cievke č. 2 a podobne aj pri cievke č. 3 bola $U_{\text{poč}} = 2,6\text{kV}$, hodnota maximálneho zdanlivého náboja bola pri cievke č. 2 $q_{\max}=60\text{pC}$ a pri cievke č. 3 $q_{\max}=300\text{pC}$. Počiatočné napätie na cievke č. 4 bolo $U_{\text{poč}} = 2,8\text{kV}$ a $q_{\max}=140$. Pri týchto cievkach bola pravdepodobne zaznamenaná vnútorná výbojová činnosť. Pri porovnaní hodnôt počiatočného napätia mala teda najlepší výsledok cievka č. 1 a č. 5. Fázové rozloženia všetkých meraných cievok boli podobné t.j. v zápornej polovlne bola nameraná vyššia výbojová činnosť ako v kladnej.

Na obr. 3 až 7 sú znázornené fázové rozloženia maximálnych hodnôt zdanlivého náboja pre tieto napäťové hladiny:

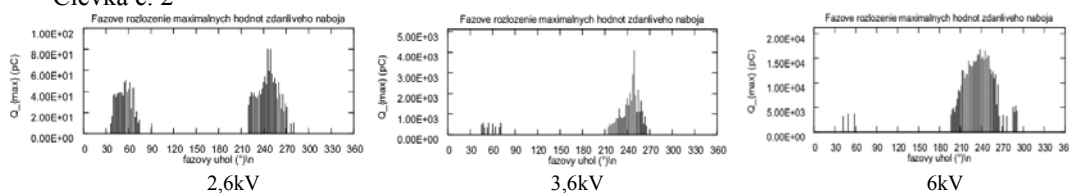
- napätie počiatočných výbojov,
- nominálne fázové napätie 3,6kV,
- nominálne združené napätie 6kV.

Cievka č. 1



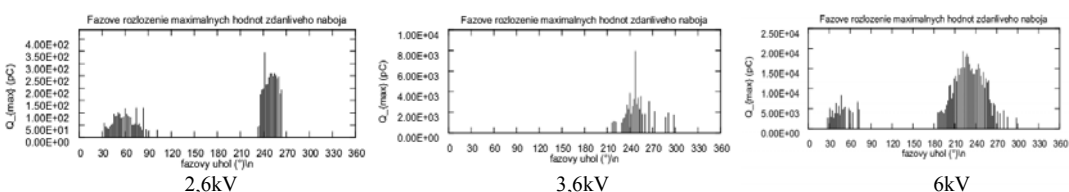
Obr. 3 Fázové rozloženia maximálnych hodnôt zdanlivého náboja pre napäťové hladiny 3,2kV; 3,6kV a 6kV.

Cievka č. 2



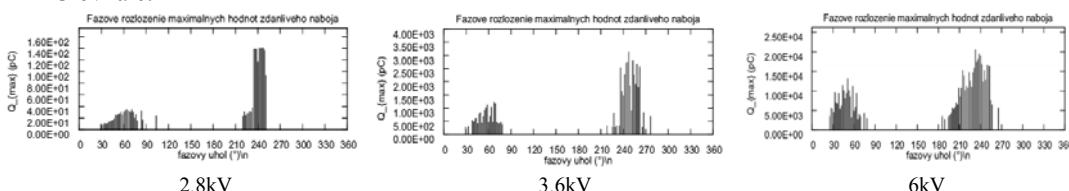
Obr. 4 Fázové rozloženia maximálnych hodnôt zdanlivého náboja pre napäťové hladiny 2,6kV; 3,6kV a 6kV

Cievka č. 3



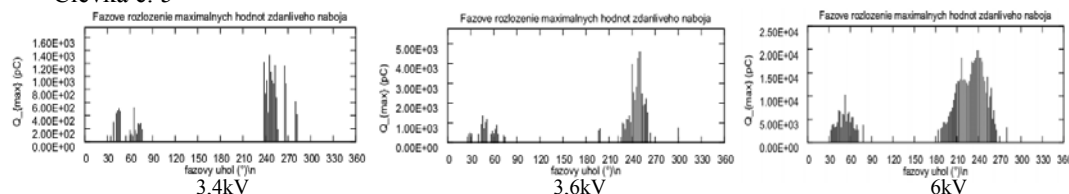
Obr. 5 Fázové rozloženia maximálnych hodnôt zdanlivého náboja pre napäťové hladiny 2,6kV; 3,6kV a 6kV

Cievka č. 4



Obr. 6 Fázové rozloženia maximálnych hodnôt zdanlivého náboja pre napäťové hladiny 2,8kV; 3,6kV a 6kV

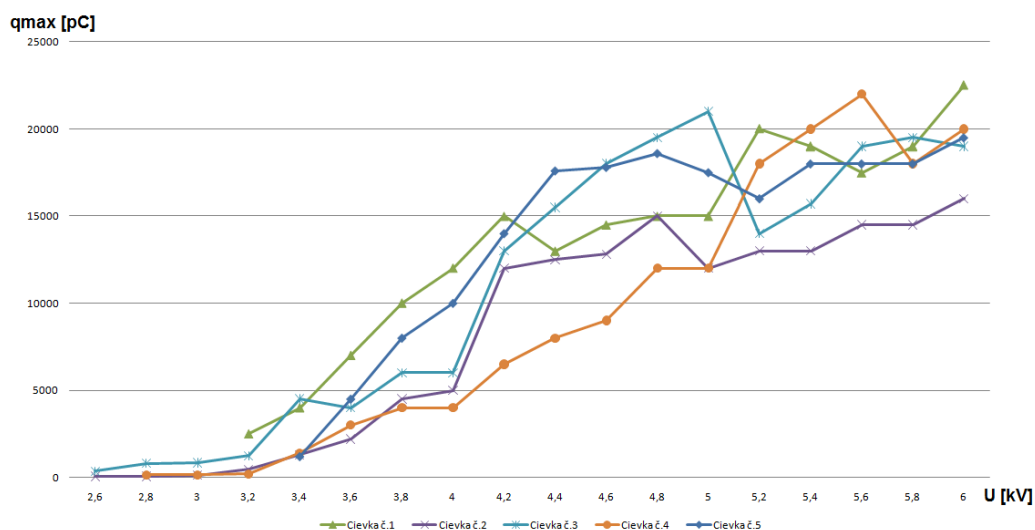
Cievka č. 5



Obr. 7 Fázové rozloženia maximálnych hodnôt zdanlivého náboja pre napäťové hladiny 3,4kV; 3,6kV a 6kV

VI. ZÁVER

Na obr. 8 je graficky znázornená závislosť maximálnej hodnoty zdanlivého náboja od napätia, ktorá poukazuje na to, že trend rozvoja výbojovej činnosti je vo všetkých meraných cievkach podobný. Z tabuľky 1 vyplýva, že namerané maximálne hodnoty zdanlivého náboja sa do výraznej miery neodlišovali a fázové rozloženia všetkých meraných cievok boli podobné. Na úplný záver je možné konštatovať, že kvalita izolačného systému meraných cievok je minimálne na rovnakej úrovni.



Obr. 8 Závislosť maximálnej hodnoty zdanlivého náboja od napätia

POĎAKOVANIE



„Táto práca bola podporovaná Agentúrou Ministerstva školstva Slovenskej republiky pre štrukturálne fondy EÚ na základe projektu *Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy* (číslo projektu: 26220220029, prioritná os 2 Podpora výskumu a vývoja)“.

LITERATÚRA

- [1] Záliš, K.: „Evaluation of Partial Discharge Activity by Expert Systems,” In: 11th International Symposium of High Voltage Engineering, ref. 5.344.P5, London, Sept.1999.
- [2] Kurimský, J., Kolcunová, I., Cimbala, R.: Partial discharge analysis for insulation systems of electric rotating machines with various voltage stress. In: Acta Electrotechnica et Informatica. roč. 8, č. 4 (2008), s. 64-67.
- [3] Cimbala, R.: Absorpčný prúd a jeho vzťah k iným diagnostickým metódam, Diagnostika 93, Plzeň 1993.
- [4] Mentlík, V. : Dielektrické prvky a systémy. Plzeň: BEN, 2006. ISBN 80-7300-189-6.
- [5] Kolcunová, I., Kvakovský, M., Kurimský, J.: Meranie povrchových výbojov na vn cievke točivého stroja. In: Elektroenergetika Journal. č. 4 (2009), ISSN 1337-6756.
- [6] Cimbala, R., Kolcunová, I.: Diagnostika izolačných systémov elektromotorov, Technická univerzita v Košiciach, ISBN 978-80-8073-998-0, 2008.

Analysis of Voltage Variation in Low Voltage Networks

Bystrík DOLNÍK

Dept. of Electric Power Engineering, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovak Republic

bystrik.dolnik@tuke.sk

Abstract – Uninterrupted power supply in the developed countries is already a matter of course. Great emphasis today is laid to meet its quality. One of the monitored parameters which must ensure the electricity supplier are the voltage characteristics. Supply voltage variations may have a significant impact on the life and operation of certain electrical equipments. The article presents some possibilities to describe the characteristics of the point of delivery in terms of voltage characteristics. To reflect the voltage characteristics the tools of mathematical statistics were applied.

Keywords – power quality, voltage variation, statistics, pattern

I. INTRODUCTION

At present uninterrupted supply of electricity becomes a matter of course in developed countries. The emphasis is high requirement for its quality in this case. Small electricity consumers in households use a variety of different electrical appliances with low power. As individual customers do not significantly affect the quality of electricity but together can significantly affect the quality in endpoints of supplied electricity. On the other hand, wholesale customers must ensure the requirements of EMC also on side of their own consumption in order did not influence significantly the characteristics of the supply voltage [1], [2], [3]. The low-voltage power characteristics include:

- network frequency,
- supply voltage,
- voltage variations,
- flicker,
- supply voltages dip,
- short supply interruption,
- long supply interruption,
- temporary power frequency overvoltages,
- transient overvoltages,
- supply voltages unbalance,
- harmonic voltage,
- interharmonic voltage,
- network signaling voltage on supply voltage.

It is clear that listed power characteristics define the supply voltage with respect on frequency, amplitude, waveform and symmetry of three-phase voltages. These characteristics are subject to changes during normal operation of supply system due to load changes, disturbances caused by certain equipment and the occurrence of failures that usually cause external factors. With regard to changes in characteristics, they vary randomly in time and in terms of location. Some phenomena that affect the voltage, can not be significantly predicted. One of the listed power characteristics are supply voltages variations. This is an increase or decrease of voltage normally caused by a change in the overall load in distribution system or part of its section.

Under normal operating condition which excludes voltage interruptions should be in each period of one week 95 % of 10-minute mean r.m.s value of power supply voltage within the range $U_n \pm 10\%$ where U_n is the nominal voltage of the system [4]. Note that all the average r.m.s values of supply voltage in 10-minute measurement intervals must be in the range $U_n + 10\% - 15\%$. The severity of variations in supply voltage has a significant impact on life and/or operation of some appliances and in practice the level of power quality is a compromise between user and supplier [5], [6].

Table 1
 THE NUMBER OF DATA SAMPLES MEASURED DURING THE PERIOD

Period	1999	2000	2002	2003	2007	2008	2009	2010
Spring-time							6 739 200	7 862 400
Summer-time		92 160		431 940	4 147 200	2 678 400	6 739 200	6 048 000
Autumn term						7 689 600	6 912 000	6 652 800
Wintertime	10 080		18 720	449 214	950 400	3 369 600	7 084 800	7 084 800

II. DESIGN AND PREPARATION OF THE EXPERIMENT

Long-term monitoring of the supply voltage can be used for the characterization or classification of the delivery point respectively. Electricity supplier shall supply electricity in the required quality, while the standards allow a defined tolerance for different voltage characteristics. Already mentioned, the characteristics of voltage variation is random, therefore, continuous monitoring of voltage can be considered as a tool that helps both the electricity supplier and the customer.

The experiment was designed so that the point of delivery can be described in the long term by variations in supply voltage. For this purpose a continuous record of r.m.s value of supply voltage was realized. The case of interruption of electricity supply due to prearranged maintenance were not taken into account. To obtain the optimum number of data 60 s recording interval was chosen first. After a partial analysis of the trend in the recorded voltage on reference point of delivery the recording interval was shortened for 10 seconds and finally for 1 second. The experiment began in the winter season in 1999 until 2000, then in 2002–2003 and then continued from 2007 and lasts until now. An overview of the total number of measured data for the period is given in Table 1. 74 960 514 data samples has been acquired together for the period given in Table 1.

Measured data can be evaluated according to the selected time period. As a basis for analyzing voltage variations day cycle was used which begins on the 12th o'clock at night and lasts for 24 hours. The analysis can be used for different periods of time as a weekly, monthly or quarterly in the season (spring, summer, autumn and winter) and so on. As a comparative template of voltage the monthly average was used.

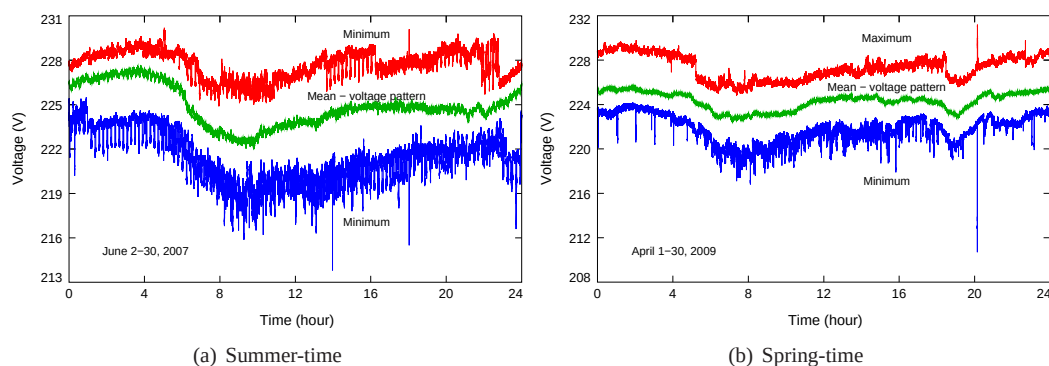


Fig. 1 Time course of calculated monthly template of voltage for the monthly period

In Fig. 1 the time course of calculated voltage template for the monitored period in June 2–30, 2007 (summer-time) and April 1–30, 2009 (spring-time) are presented. The course of comparative voltage template is located in the zone limits which are the minimum and maximum value at any given time. The figure shows that the time record can be divided into several time intervals, which are characterized in terms of the trend of the monitored voltage variations. According to trends of voltage patterns four characteristic periods were selected. The first period starts at midnight and ends around the 6th o'clock in the morning having a balanced trend, i.e. do not appear to extreme voltage fluctuations. The second period lasts from 6 hours to 10 hours in the morning. The voltage has a downward trend with a decrease of several units of volts to the first period. From 10 hours to 16 hours the voltage has an increasing trend, while in some cases exceeds or falls below the mean level of the first period (cf. the time course in Fig. 1a and Fig. 1b). The last, fourth period is characterized by relatively even trend with potentially slight increase. The above examples compare trends of the voltage template in the summer and spring season.

In a similar manner each other individual periods can be compared. It may also be used the annual voltage template by applying of the factor analysis to obtain an overview of the extreme fluctuations to the voltage template. Factor analysis is a statistical method where using a correlation

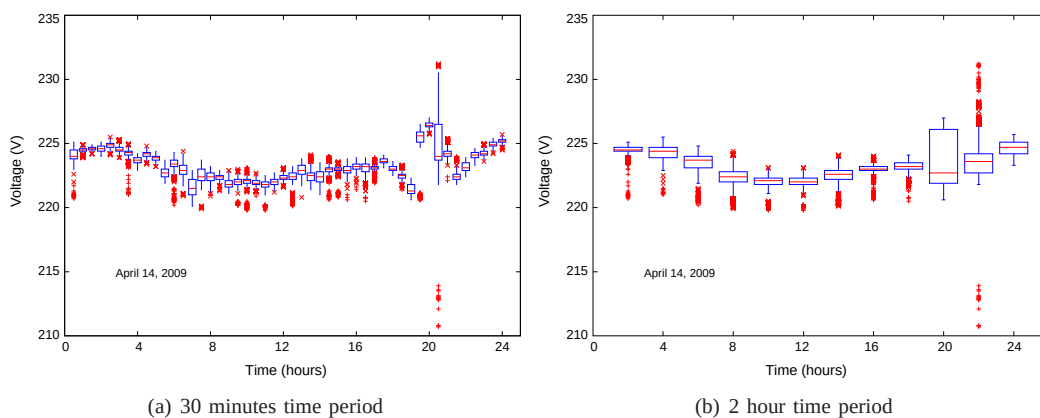


Fig. 2 Box-plot graph of voltage time course for 24-hour period of 14 April 2009

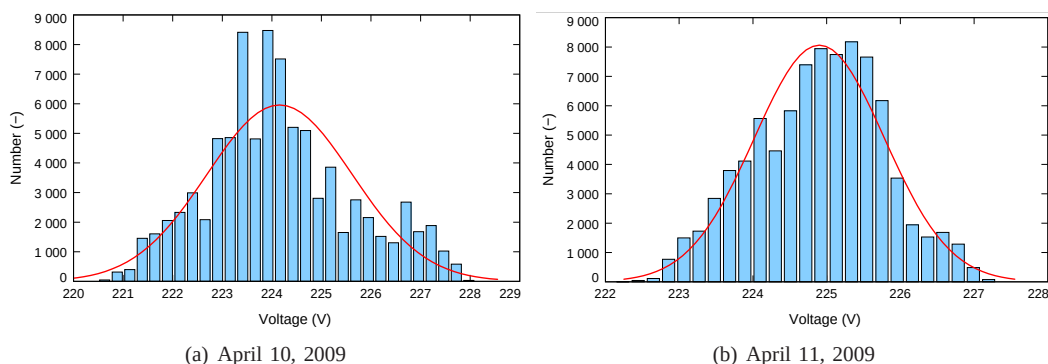


Fig. 3 Histogram of voltage time course for 24-hour period

factor, or for cases illustrative presentation using a correlation diagram can express the correlation between the measured effective voltages. Fig. 2 shows an example of illustrative presentation of measured data (r.m.s supply voltage) using what is known box-plot graph display which offers a basic statistics: Q_2 median (the 50th percentile, horizontal line segment within the rectangle), lower quartile Q_1 (the 25th percentile), upper quartile Q_3 (the 75th percentile), minimum and maximum.

Boxplots are useful for determining where the majority of the data lies. The bottom and top of the box are always the 25th and 75th percentile (the lower and upper quartiles, respectively – the height of the rectangle is equal to the difference $Q_3 - Q_1$ which is the quartile range, which is 50 % of the data), and the band near the middle of the box is always the 50th percentile (the median). Between the median and quartiles Q_1 respectively Q_3 there is 25 % of the data. Vertical lines define the border beginning of the moderate diverted data (in Fig. 2 marked with "x"); the limit is equal to $1.5 \cdot (Q_3 - Q_1)$ starting in Q_2 . The limit of beginning of the strongly diverted data is defined by length of $3 \cdot (Q_3 - Q_1)$ from Q_2 (in Fig. 2 marked with "+"). As can be seen from Fig. 2 extreme changes (voltage variations) become around the 9 hours in the afternoon. The median value is 223.1 V and the mean value is 223.30 V. Quartile range $Q_3 - Q_1$ is 1.8 V. Box-plot in Fig. 2a, although it is less readable, provides more details on the changes in voltage. In contrast, box-plot in Fig. 2b is more readable.

Figure 3 shows the histogram of the voltage record variations for 24-hour period of April 10, 2009 and April 11, 2009. The voltage record was made with 1-second sampling interval. The number of classes for construction of the histogram is 30, while for a 1-second record comes together in 24 hours 86400 data samples. Bell-shaped curve in the graph represents the normal distribution density. The calculated values of 10 April 2009 are as follows: the median has the value of 224 V, the mean value is 224.2 V and the quartile range is 1.8 V. Finally, the calculated values of 11 April 2009 are as follows: the median has the value of 225 V, the mean value is 224.9 V and the quartile range is 1.2 V.

III. CONCLUSION

In the experiment a long-term record of r.m.s. voltage values to classify the point of delivery was realized. For data analysis and classification of the possible consumer site were used tools of

mathematical statistics, which were analyzed in one parameter of the power characteristics – the voltage variations. The comparative 24-hour voltage pattern from a monthly record was calculated. Using templates can make additional calculations and graphically illustrate the relationship between the measured voltages. According to the trends of time-voltage templates four characteristic periods were designed, typical for the monitored point of delivery. The trend of periods can be used for classification of the identifiable segment of power supply in terms of power quality and check back of agreed parameters with the possibility of retroactive financial compensation.

ACKNOWLEDGMENT



We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project “Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky”, ITMS 26220120055.

REFERENCES

- [1] D. L. Brooks, R. C. Dugan, M. Wacławski, and A. Sundaram, “Indices for assessing utility distribution system rms variation performance,” *IEEE Transactions on Power Delivery*, vol. 13, no. 1, pp. 254–259, 1998.
- [2] N. G. Hingorani, “Introducing custom power,” *IEEE Spectrum*, vol. 32, no. 6, pp. 41–48, June 1995.
- [3] C. J. Melhorn and M. F. McGranaghan, “Interpretation and analysis of power quality measurements,” *IEEE Transactions on Industry Applications*, vol. 31, no. 6, pp. 1363–1370, Nov/Dec 1995.
- [4] “STN EN 50160 napäťové charakteristiky elektrickej energie dodávanej verejnými rozvodnými sústavami,” August 1998, Úrad pre normalizáciu, metrológiu a skúšobníctvo SR.
- [5] H. Markiewicz and A. Klajn, “Power quality application guide, voltage disturbances. standard EN 50160 – voltage characteristics in public distribution systems.” European Copper Institute, 168 Avenue de Tervueren, B-1150 Brussels, Belgium, Tech. Rep., July 2004. [Online]. Available: http://www.leonardo-energy.org/webfm_send/2703
- [6] M. H. J. Bollen and I. Yu-Hua Gu, *Signal Processing of Power Quality Disturbances*, 1st ed., ser. IEEE Press Series on Power Engineering. 111 River Street, Hoboken, NJ 07030, USA: John Wiley & Sons, Inc., 2006, ch. ORIGIN OF POWER QUALITY VARIATIONS, pp. 52–67, 82–112.

Príspevok k štatistickému vyhodnoteniu prepätí v sieti nízkeho napätia

Bystrík DOLNÍK

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky, Technická univerzita
v Košiciach, Slovenská republika

bystrik.dolnik@tuke.sk

Abstrakt – Spínacie prepätia spôsobené spínaním elektrických zariadení a ich ochrana pred prepätiami je stále predmetom zvýšeného záujmu pretože prepätia nepriaznivo pôsobia na izoláciu elektrických zariadení, pričom často je táto prepäťová udalosť sprevádzaná elektromagnetickou interferenciou. V článku sú opísané niektoré typické charakteristiky spínacích prepätí namerané na zariadeniach pripojených do elektrickej siete nízkeho napätia. Na vyjadrenie charakteristík sú aplikované základné nástroje matematickej štatistiky.

Kľúčové slová – izolácia, prepätie, štatistika

I. ÚVOD

Elektrické a elektronické zariadenia používané v elektroenergetike sú dnes oveľa citlivejšie ako tomu bolo pred dvadsiatimi a viac rokmi. Na strane odberateľ a sa zvýšili a stále zvyšuje počet zariadení (zát'aže) ako sú: počítače, komunikačné zariadenia, rôzne elektronicky riadené zdroje na napájanie elektrických zariadení, zariadenia používané v zdravotníctve a pod. Spomínané zariadenia reprezentujú typický príklad zát'aží, ktoré sú citlivé na poruchy v napájacej sieti. Treba však poznamenať, že aj samotné zát'aže môžu byť zdrojmi porúch a prepätí. Všeobecne platí, že akékoľvek spínanie (plánované alebo neplánované), vznik poruchy, prerušenie dodávky elektriny v elektrickej sieti, spôsobí prechodný dej, ktorý môže byť sprevádzaný prepätím. S problematikou prepätí zaoberá sa taktiež vedný odbor elektromagnetická kompatibilita. Spínacie prepätia možno opísať rôznymi spôsobmi, pričom najčastejšie sa používajú tieto parametre:

- 1) amplitúda,
- 2) trvanie,
- 3) U_t integrál, ktorý je definovaný nasledovne

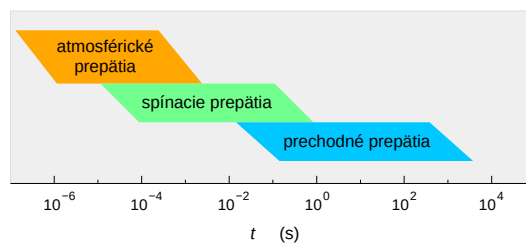
$$U_t = \int U(t)dt, \quad (1)$$

kde t je čas a U je napätie. Hranice integrovania sú zvyčajne: začiatok prepäťovej udalosti $t = 0$ a horná hranica napr. čas, za ktorý napätie dosiahne 10 % maximálnej hodnoty prepätia. Napätie $U(t)$ možno merať alebo od nuly alebo od okamihu odklonu od sínusového napätia v ustálenom stave. Prepätia možno klasifikovať aj podľa ich trvania nasledovne: extrémne krátke, krátke, dlhé a extrémne dlhé. Niektoré zariadenia na monitorovanie kvality elektrickej energie nepretržite zaznamenávajú napätie, prúd, fázový posuv, napäťovú symetriu a pod. s možnosťou zachytenia udalosti, keď sledovaná veličina presiahne vopred stanovenú hodnotu – napäťová podmienka [1].

Prepäťové udalosti sú klasifikované podľa rôznych kritérií, niektoré boli už spomenuté vyššie. Prepätia s extrémne krátkym trvaním a veľkými amplitúdami vznikajú najmä počas búrkovej činnosti pri priamom zásahu blesku do zariadenia alebo vedenia a niektoré prepäťové udalosti spôsobené spínaním (hlavne v zariadeniach s plynovou izoláciou okrem vzduchu, napr. SF₆ alebo dusík). Prepäťové udalosti s krátkym a dlhým trvaním sú typické počas spínania v elektrickej sieti alebo ako reakcia na atmosférickú prepäťovú udalosť.

Na kvalitatívne vyjadrenie alebo vyhodnotenie prepätí používa sa najviac tzv. koeficient prepätia k_p , ktorý sa vypočíta ako podiel amplitúdy prepätia a pomernej jednotky definovanej podľa (3) alebo podľa (4)

$$k_p = \frac{U_{mp}}{p. j.}, \quad (2)$$



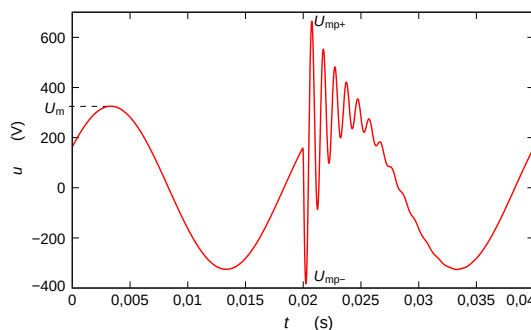
Obr. 1 Klasifikácia prepätí podľa ich trvania

kde k_p je koeficient prepätia, U_{mp} je amplitúda prepätia (pozri obr. 2) a p. j. je pomerná jednotka. Pomerná jednotka v jednofázovej elektrickej sieti sa vypočíta ako súčin efektívnej hodnoty fázového napätia a odmocniny z dvoch (3), čím sa získa amplitúda U_m , pozri obr. 2; v trojfázovej sieti zase ako podiel amplitúdy napätia siete a odmocniny z troch, pozri (4)

$$p. j. = U \cdot \sqrt{2}, \quad (3)$$

$$p. j. = \frac{U \cdot \sqrt{2}}{\sqrt{3}}. \quad (4)$$

V elektrických sieťach nízkeho napätia vyskytujú sa prepätia hlavne z dôvodu spínania, pričom spínanie môže byť mechanické alebo pomocou elektronických spínacích prvkov – charakter spínania má vplyv na prechodný dej. Poznamenajme, že akákoľvek rýchla zmena v elektrickom obvode s kapacitami a indukčnosťami vyvolá prechodný dej, ktorý má často charakter prepätí. Tieto sa ďalej šíria po metalickom vedení k ostatným pripojeným zariadeniam, čím je ich izolácia namáhaná [2], [3]. Krátke vodiče možno modelovať pomocou modelov so sústredenými parametrami, dlhé vodiče (od mn až do zvn) resp. vedenia pomocou modelov s rozloženými parametrami ako vedenia s malými stratami.



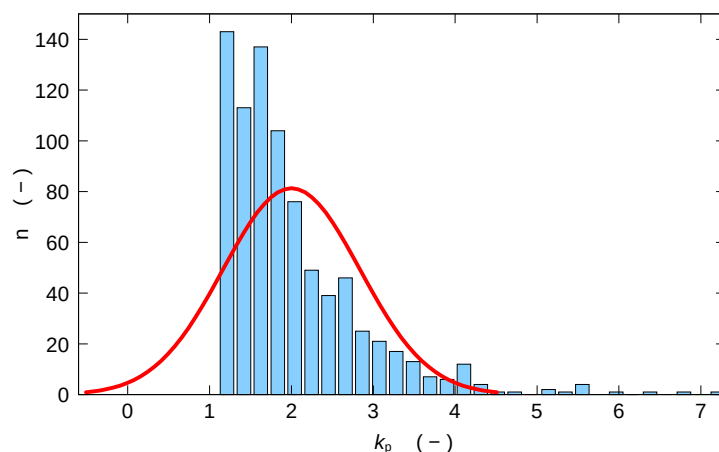
Obr. 2 Časový priebeh napätia so superponovaným spínacím prepätím

II. EXPERIMENT

A. Návrh a príprava experimentu

Cieľom experimentu bolo zistiť vlastnosti spínacích prepätí, ktoré sú generované manuálnym spínaním (zapínanie a vypínanie) elektrických zariadení na nízke napätie s malým výkonom pomocou mechanického spínača. Boli použité zariadenia, ktoré sú bežne používané v spotrebičoch v domácnosti ako: jednofázový sieťový transformátor, osobný počítač, vŕtačka, chladnička, žiarovka, žiarovka a kondenzátor. Poznamenajme, že počet zariadení používajúcich sieťový transformátor ubúda, lebo čoraz viac sa začínajú používať elektronické spínané zdroje kvôli menším rozmerom a pomerne vysokej účinnosti.

Vybrané zariadenia reprezentujú rôzne typy záťaž: ohmická (žiarovka), prevažne indukčná (sieťový transformátor, vŕtačka) a prevažne kapacitná (kondenzátor). Zariadenia boli pripojené do rozvodu elektrickej energie s nízkym napätím 230/400 V, priemyselnou frekvenciou 50 Hz a s konfiguráciou TN-C-S. Časový priebeh napätia bol meraný pomocou odporového deliča napätia s deliacom pomerom 1 000:1 umiestnený blízko zariadenia a pripojený k digitálnemu osciloskopu s cieľom urobiť analýzu nameraných prepätových udalostí pomocou štatistickej analýzy. Spínanie spotrebičov robila obsluha a preto okamih spínania možno považovať za náhodný jav.



Obr. 3 Histogram a graf normálneho rozdelenia koeficientu prepätia k_p všetkých nameraných prepätí

V prípade spínania vŕtačky vznikali prepätia nie len v okamihu zapnutia resp. vypnutia, ale aj počas jej prevádzky. Na analýzu a vyhodnotenie vypočítaných dát použili sa nástroje matematickej štatistiky, predpokladal sa náhodný výber. Pred vlastnou analýzou treba zistiť platnosť základných predpokladov, t. j. nezávislosť, homogenita a normalita výberu. Keď nie sú splnené, potom ďalšie postupy ako výpočet priemeru, intervalu spoľahlivosti, kvantilov a väčšina testov je spochybnená.

B. Vyhodnotenie experimentu

Na vyhodnotenie prepätí, ktoré boli generované spínaním stanovila sa prahová hodnota 10 % nad amplitúdou menovitého fázového napätia v jednofázovej sieti (koeficienty $1 \leq k \leq 1,1$ sa pri vyhodnotení nebrali do úvahy). Počas experimentu bolo celkovo nameraných 825 prepätí udalostí podľa uvedeného kritéria. Amplitúdy nameraných prepätí pohybovali sa od prahovej hodnoty až do jednotiek kilovoltov (1–2,2 kV).

Tabuľka 1
 MAXIMÁLNA AMPLITÚDA PREPÄTIA NAMERANÁ NA VYBRANÝCH ELEKTRICKÝCH ZARIADENIACH

Elektrické zariadenie	Amplitúda prepätia ^a (V)	Koeficient prepätia (–)
Jednofázový elektromotor	1 675,25	5,15
Kondenzátor	1 096,25	3,37
Jednofázový transformátor	2 228,00	6,85
Vŕtačka	2 377,75	7,31
Žiarovka	1 976,25	6,08

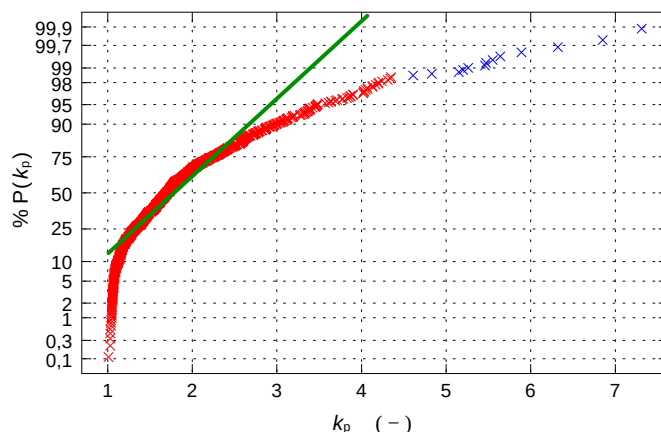
^aPre uvedené amplitúdy prepätí nie je zohľadnená polarita napätia.

Tabuľka 2
 PREHL'AD VYBRANÝCH ŠTATISTÍK KOEFICIENTOV PREPÄTIA

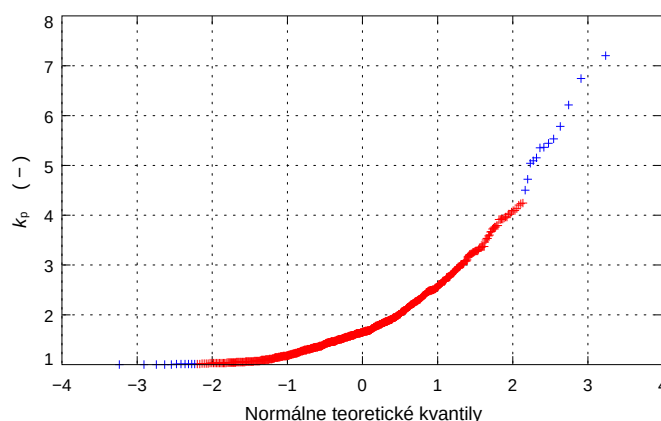
Minimum	1. kvartil	Medián	3. kvartil	Maximum	Priemer	Štand. odchýlka	Šikmost'	Špicatost'
1,11	1,44	1,76	2,33	7,31	2,00	0,84	2,03	6,11

V prípade analýzy prepätí, ktoré sú generované elektrickými zariadeniami treba pripomenúť, že okrem napätí ového namáhania izolácie, prepätia spôsobujú aj rušiacie napätia v sieti cez galvanickú väzbu a rušiacie polia cez elektromagnetickú väzbu. Podobné výsledky prepätí namerali sa na prevažne odporovej záťaži, ktorú reprezentovala žiarovka. Maximálna nameraná hodnota prepätia bola 1,98 kV, pričom prepätia boli generované ako pri zapínaní, tak aj pri vypínaní. Keďže sa jedná o prevažne odporovú záťaž, možno vzniknuté prechodné dej a s tým spojené oscilácie počas spínania pripísať rezonančnému obvodu, ktorý je tvorený parazitnými kapacitami a indukčnosťou napájacích prívodov k žiarovke.

Amplitúda generovaného prepätia je závislá, okrem už spomínaných faktorov, aj od mechanickej konštrukcie použitého spínacieho prvku. Aby sa zabránilo prechodným dejom medzi kontaktami spínača – cyklickému generovaniu prepätí zotaveným napätím počas vypínania, je snaha, aby sa proces vypnutia uskutočnil v čo najkratšom čase.



Obr. 4 Normálny pravdepodobnostný graf pre všetky vypočítané koeficienty prepätia k_p



Obr. 5 Kvantil-kvantilový graf pre všetky vypočítané koeficienty prepätia k_p

Spínanie osobného počítača vyvolávalo iba veľmi malé prepätia, ktoré nemajú podstatný vplyv na namáhanie izolácie pripojených elektrických zariadení. Spínanie jednofázového motora bez mechanickej záťaže a podobne aj spínanie jednofázového transformátora bez záťaže (naprázdno) spôsobilo generovanie prepätí do 2 kV. Krátke zhrnutie maximálnych amplitúd prepätia, ktoré boli počas experimentu namerané na vybraných elektrických zariadeniach, je uvedený v tab. 1.

C. Diskusia k experimentu

Ako už bolo spomenuté, na analýzu a vyhodnotenie dát nameraných v experimente použili sa nástroje matematickej štatistiky: minimum, 1. kvartil, medián, 3. kvartil, maximum, priemer, štandardná odchýlka, šikmosť a špicatosť. Pre štatistickú analýzu bol uvažovaný štatistický súbor, ktorý obsahuje všetky namerané koeficienty prepätia. Prehľad vybraných štatistík je uvedený v tab. 2. Z tabuľky je zrejmé, že hodnota pre kritérium prepätia bola stanovená na 110 %. Podľa (2) vypočítal sa koeficient prepätia. Okrem číselného vyjadrenia štatistík použili sa diagnostické grafy pre účel prieskumnej analýzy. Účelom prieskumnej analýzy dát je odhaliť zvláštnosti a overiť predpoklady vybranej vzorky dát pre ďalšie spracovanie, t. z. napomáha minimalizovať riziko chybných záverov o rozložení, z ktorého daný náhodný výber pochádza. Tento postup sa nazýva štatistická indukcia. Histogram uvedený na obr. 3 ponúka lepšiu interpretáciu vypočítaných koeficientov prepätia. Ako možno vidieť z obrázku, najväčší počet (152) prepätíových udalostí zodpovedá koeficientom prepätia v rozsahu od 1,54 do 1,75. Túto skutočnosť, pre uvažovaný štatistický súbor, viac vystihuje medián, ktorý má hodnotu 1,76 ako priemer s hodnotou 2,00. Krivka preložená cez histogram na obr. 3 reprezentuje graf hustoty normálneho rozdelenia.

Na obr. 4 je zobrazený normálny pravdepodobnostný graf. Jedná sa o graf, ktorý porovnáva namerané hodnoty (v našom prípade vypočítané koeficienty prepätia) s ich očakávanými hodnotami a označuje sa aj ako P-P graf. Na tomto type grafu priamka prechádza cez prvý a tretí kvantil. V prípade normálneho rozdelenia body patriace koeficientom prepätia sa nachádzajú blízko tejto priamky. Na obr. 4 vidno, že usporiadané dvojice (dát) sú zoradené do konkávnej krivky, čo značí, že štatistický súbor má kladnú šikmosť (2,0311 – pozri tab. 2) a zároveň nespĺňajú podmienku pre

normálne rozdelenie. Veľmi výraznú šikmost' má napríklad exponenciálne rozloženie. Odľahlé (vybočujúce) hodnoty sú vyznačené farebne modrou farbou. Medzi parametre tvaru patrí aj špicatost' a vyjadruje koncentráciu rozloženia početnosti okolo priemeru. Pre normálne rozloženie je špicatost' rovná nule. Špicatost' narastie keď sa v štatistickom súbore nachádzajú obojstranne vybočujúce merania. V nami uvažovanom prípade je špicatost' rovná 6,113 (tab. 2) čo je výrazné odchylenie sa od nuly.

Na obr. 5 je zobrazený Q-Q graf (kvantil-kvantilový) koeficientov prepätia pre všetky namerané dáta. V tomto grafe sú vynesené vypočítané dáta (koeficienty prepätia – skutočné hodnoty) voči teoretickým kvantilom (kvantily štandardného normálneho rozloženia s priemerom 0 a štandardnou odchýlkou 1) [4]. Q-Q graf umožňuje posúdiť, či dátový súbor pochádza z nejakého známeho rozloženia. Z jeho tvaru dá sa posúdiť symetria, normalita, špicatost' a homogenita výberu [5]. Z obr. 5 dá sa usúdiť, že dáta nemajú normálne rozdelenie, pretože body neležia na priamke, ale na krivke bez výrazných zlomov a nehomogenít. Majú kladné zošikmenie – jedná sa o konvexnú krivku. Z toho vyplýva, že použitie aritmetického priemeru ako odhadu strednej hodnoty by bolo nesprávne. V takýchto prípadoch je lepšie používať medián.

III. ZÁVER

V experimente som analyzoval výskyt prepätí generovaných spínaním jednofázových elektrických zariadení používaných v sieťach nízkeho napätia ako: osobný počítač, jednofázový sieťový transformátor, vŕtačka, chladnička, žiarovka a žiarivka. Na analýzu nameraných dát som použil nástroje matematickej štatistiky. Priemerná hodnota koeficientu prepätia zo všetkých 825 nameraných udalostí je 2,00. Najväčšia početnosť koeficientu prepätia sa pohybovala od 1,54 do 1,75 pre všetky merané elektrické zariadenia. Pre štatistické vyhodnotenie nameraných dát sa použilo normálne (Gaussovo) rozdelenie, na grafickú interpretáciu dát histogram početností, hustota normálneho rozdelenia pravdepodobnosti, kvantil-kvantilový graf (Q-Q graf) a pravdepodobnostný graf (P-P graf).

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku / Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu „Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky“, ITMS 26220120055.

LITERATÚRA

- [1] M. H. J. Bollen and I. Yu-Hua Gu, *Signal Processing of Power Quality Disturbances*. Wiley-IEEE Press, 2006.
- [2] S. Perera, "Transient overvoltages on the electricity supply network – classification, causes and propagation," Power Quality and Reliability Centre, Tech. Rep., 2005. [Online]. Available: <http://www.elec.uow.edu.au/pqaustralia/files/technote8.pdf>
- [3] —, "Transient overvoltages on the electricity supply network – effects on connected equipment and their mitigation," Power Quality and Reliability Centre, Australia, Tech. Rep., 2006. [Online]. Available: <http://www.elec.uow.edu.au/pqaustralia/files/technote9.pdf>
- [4] J. Beirlanta, W. Glänzelb, and A. Carbonez, "Scoring research output using statistical quantile plotting," *Journal of Informetrics*, vol. 1, no. 3, pp. 185–192, 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/B83WV-4NYSHJH-1/2/d2cd72c0a2a5bfef1872131e16120cd3>
- [5] J. Liang, W. S. Y. Pan, and Z.-H. Yang, "Characterization-based q-q plots for testing multinormality," *Statistics & Probability Letters*, vol. 70, no. 3, pp. 183–190, 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/B6V1D-4DPSVF8-3/2/bfee4a50de503e1fd81ea0944e89efe6>

Meranie výbojovej činnosti na cievkach s termosetickou izoláciou a s polovodivou ochranou

Marián HRINKO, Irida KOLCUNOVÁ, Juraj KURIMSKÝ

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

marian.hrinko@tuke.sk, iraida.kolcunova@tuke.sk, juraj.kurimsky@tuke.sk

Abstrakt — Vzhľadom na smernicu nárastu vývoja nových izolačných materiálov a trendov, ktoré smerujú k minimalizácii elektrických strojov točivých, je nutné mať na vedomí overovanie kvality týchto materiálov. Čím viac hrúbky klesajú, tým sú izolácie vystavené vyššiemu elektrickému namáhaniu. Dá sa povedať, že reciprokej hodnote elektrického namáhania zodpovedá životnosť izolácie, čiže v konečnom dôsledku dochádza k poklesu doby životnosti izolácie. Vhodnou metódou na overovanie kvality izolačných systémov je meranie čiastkových výbojov, ktorá bola použitá v tomto prípade.

Kľúčové slová — statorová izolácia, čiastkové výboje, točivé elektrické stroje

I. ÚVOD

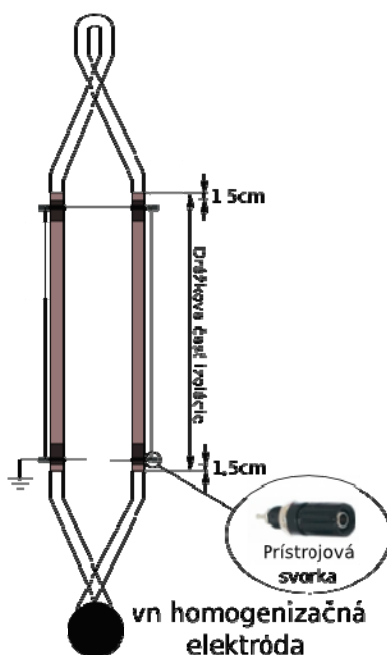
Počas prevádzky stroja izolácia podlieha kombinácii degradačných faktorov, ktoré spôsobujú rôzne zmeny vo vnútornej štruktúre izolačného materiálu, čo sa navonok prejaví znížením kvality celého izolačného systému [4]. Keďže kvalita izolácie je úzko spojená so spoľahlivosťou a bezporuchovosťou celého stroja je nutnou podmienkou tie faktory čo najlepšie eliminovať. S tým súvisí diagnostika elektrických strojov točivých, ktorá zisťuje stav daného zariadenia a ktorej úlohou je zabezpečiť čo najdlhšie bezporuchovú činnosť daného zariadenia v prevádzke [3].

Jednou z diagnostických metód na určovanie stavu izolačného systému je meranie čiastkových výbojov. Výstupom tejto metódy sú rôzne veličiny, ktoré pomáhajú lepšie poznať vnútorný stav izolácie a na základe tejto informácie zhodnotiť či daná izolácia zodpovedá stanovenej kvalite. Na úplnú analýzu týchto veličín potrebujeme navyše faktor, ktorý nie je možné zmerať – a to ľudskú skúsenosť. Takže okrem grafických výstupov z merania je dôležité aj oko, ktoré tieto výsledky vyhodnocuje. V súčasnej dobe je možná alternatíva ľudskej skúsenosti – neurónová sieť. Ale na to, aby dokázala správne vyhodnocovať grafické výstupy z merania je nutné ju najprv naučiť. Poskytnutie vhodných grafických výstupov na učenie neurónovej siete je jedným z cieľov tohto článku.

II. PRÍPRAVA VZORKY A MERANIE

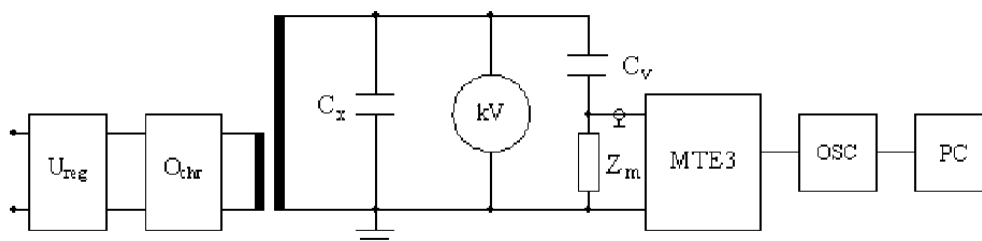
Meranými objektmi boli štyri vysokonapäťové cievky pre asynchrónny stroj s menovitým fázovým napätím $U_n=3,6\text{kV}$, menovitým prúdom $I_n=27,6\text{A}$ a menovitým výkonom $P_n=200\text{kW}$. Hlavná izolácia cievok bola SAMICATHERM 366.28, čo predstavuje kalcínovo-sľudovú izoláciu spájanú epoxidovou živicom, navinutou diskontinuálne. Na elimináciu povrchových výbojov na výstupe cievok z drážky statora sa použila polovodivá ochrana CoronaShield P8001– fenolová živica s polovodivým plnivom, napäťovo nezávislá.

Cievka bola zavesená na izolačnom lane vo Faradayovej kletke. Konce cievky boli ošetrené proti korónovým výbojom homogenizačnou elektródou, na ktorú sa privádzalo vysoké napätie. Uzemnenie bolo realizované prostredníctvom štyroch bodov, t.j. pomocou pásov alobalu každej šírky 5cm ovinutých 1,5cm od koncov vodivej vrstvy drážkovej časti izolácie, spojené prístrojovými svorkami K201 vid' Obr.1. Jednotlivé svorky boli vzájomne spojené vodivou sieťovinou a cez uzemňovací vodič privedené na potenciál zeme vid' Obr.1. Toto usporiadanie modeluje uloženie cievky v drážke statora.



Obr. 1 Princíp uzemnenia meraného objektu

Na meranú vzorku sa priviedlo napätie 2kV a nechalo pôsobiť tri minúty. Ak sa za túto dobu objavili stabilné výboje, tak za hodnotu napätia počiatkových čiastkových výbojov $U_{poč}$ sa považovala práve táto hodnota. Ak sa neobjavili stabilné čiastkové výboje napätie sa postupne zvyšovalo o 200V s rovnakou dobou pôsobenia, až pokiaľ sa nezachytili stabilné výboje. Pri napätí počiatkových výbojov sa spustil program na vyhodnocovanie výbojovej činnosti. Ďalšie merania boli realizované pri zvyšovaní napätia krokom 200V do napäťovej hladiny 6kV (nominálna hodnota združeného napätia vysokonapäťovej cievky).



Obr. 2 Schéma zapojenia pre priame meranie čiastkových výbojov

Na detekciu čiastkových výbojov bola použitá priama galvanická metóda. Schéma zapojenia je na Obr.2. C_x predstavuje meranú vzorku, v našom prípade spomínanú vysokonapäťovú cievku. Regulovaným zdrojom U_{reg} sa nastavovalo napätie na meranej vzorke. Pomocou meracieho zariadenia MTE3 sa zachytávali a zosilňovali úbytky napätí na meracej impedancii Z_m , ktoré vznikali v dôsledku impulzov čiastkových výbojov. [2]

III. NAMERANÉ HODNOTY

Za účelom vyhodnotenia a vzájomného porovnania jednotlivých cievok boli stanovené tri napäťové hladiny a to:

- počiatkové napätie čiastkových výbojov $U_{poč}$,
- nominálne fázové napätie cievky $U_{nf}=3,6kV$,
- nominálne združené napätie cievky $U_{nz}=6kV$.

Pri jednotlivých napäťových hladinách boli zaznamenávané:

- maximálna hodnota zdanlivého náboja čiastkových výbojov q_{max} ,
- početnosť výbojov n_{str} ,
- fázový uhol vzniku čiastkových výbojov pre kladnú aj zápornú polvlnu priloženého napätia $\varphi+\varphi-$.

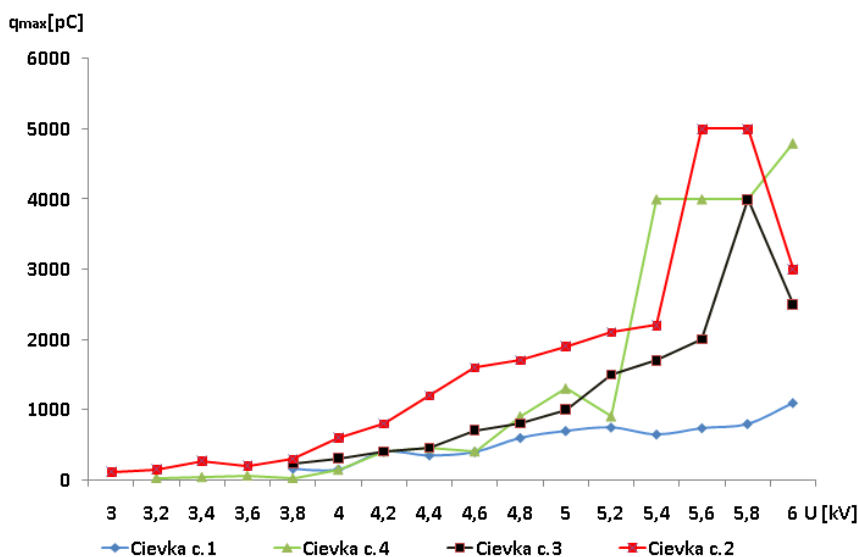
Namerané hodnoty zdanlivého náboja, početnosti a fázových uhlov vzniku čiastkových výbojov pre cievky č.1 až č.4 sú uvedené v Tabuľke 1.

Tabuľka 1
 NAMERANÉ HODNOTY CIEVKY Č.1

		U [kV]	$q_{\max}$ [pC]	n_{str} [-]	φ^+ [°]	φ^- [°]
Cievka č.1	$U_{\text{poč}}$	3,8	160	0,2	0-130	180-330
	U_{nz}	6	1100	1	0-150	170-360
Cievka č.2	$U_{\text{poč}}$	3	110	0,15	0-130	180-330
	U_{nf}	3,6	200	0,15	0-110	180-330
	U_{nz}	6	3000	0,15	0-150	180-330
Cievka č.3	$U_{\text{poč}}$	3,8	230	0,05	10-110	10-270
	U_{nz}	6	2500	0,6	0-150	180-330
Cievka č.4	$U_{\text{poč}}$	3,2	23	0,1	0-80	220-270
	U_{nf}	3,6	55	0,1	0-120	180-300
	U_{nz}	6	4800	0,1	10-150	180-330

IV. ZHODNOTENIE VÝSLEDKOV A ZÁVER

Namerané maximálne hodnoty zdanlivého náboja v závislosti od napätia pre jednotlivé cievky sú na Obr.2. Najvyššiu hodnotu zdanlivého náboja dosiahla cievka č.2, konkrétne 5000pC pri napätiach 5,6kV a 5,8kV. Následne pri napätí 6kV maximálna hodnota zdanlivého náboja klesla na hodnotu 3000pC. Táto cievka disponovala aj najnižšou hodnotou počiatočného napätia $U_{\text{poč}}=3\text{kV}$, pri ktorom maximálna hodnota zdanlivého náboja bola 110pC. Ako druhá v poradí z hľadiska najnižšej hodnoty počiatočného napätia bola cievka č.4, konkrétne $U_{\text{poč}}=3,2\text{kV}$. Charakteristika maximálnych hodnôt zdanlivého náboja na napätí pre túto cievku bola strmšia ako pri cievke č.2, pričom vyššie hodnoty zdanlivého náboja boli zaznamenané iba pri napätiach 5,4kV a 6kV. Počiatočné napätie cievky č.3 bolo $U_{\text{poč}}=3,8\text{kV}$, čo je v porovnaní s cievkou č.2 o 800V vyššia hodnota. Pribeh cievky č.3 je podobný s priebehom cievky č.4 do napätia 5,2kV. Maximálna hodnota zdanlivého náboja dosiahla 4000pC a to konkrétne pri napätí 5,8kV. V konečnom dôsledku cievky č.2, č.3 a č.4 vykazovali podobnú výbojovú činnosť. Najlepšie hodnoty zdanlivého náboja vykazovala cievka č.1. Stabilná výbojová činnosť bola zachytená pri $U_{\text{poč}}=3,8\text{kV}$, čo je rovnaká hodnota ako pri cievke č.4, pričom maximálna hodnota zdanlivého náboja nepresiahla 1100pC, čo je v porovnaní s ostatnými cievkami približne štvrtina.



Obr. 2 Grafická závislosť maximálnych hodnôt zdanlivého náboja od napätia

Nižšie napätie počiatkových výbojov hovorí o prítomnosti vzduchových dutiniek v izolácii, ktorých pôvod môže byť pri výrobnom procese. Vyššie maximálne hodnoty zdanlivého náboja sa týkajú nižšej kvality aplikovania polovodivej ochrany na výstupe cievky z drážky statora. Na základe týchto úvah je možná klasifikácia cievok. Konkrétne cievky č.2 a č.3 obsahujú väčší počet plyných dutiniek v izolácii, pričom majú aj horšiu kvalitu aplikovania polovodivej ochrany. Cievky č.4 a č.1 obsahujú nižší počet vzduchových dutiniek, pričom polovodivá ochrana cievky č.1 na základe výbojovej činnosti je najlepšie aplikovaná.

V konečnom dôsledku použitie polovodivej vrstvy na výstupe cievky z drážky má pozitívny vplyv. V najhoršom prípade maximálna hodnota zdanlivého náboja nepresiahla 5000pC. Namerané výsledky môžu byť neskôr použité pre neurónovú sieť ako vstupné dáta pri fáze učenia.

POĎAKOVANIE



Podporujeme výskumné aktivity na Slovensku/ Projekt je spolufinancovaný zo zdrojov EÚ. Tento článok bol vypracovaný v rámci projektu "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

Táto práca vznikla na základe podpory vedeckej grantovej agentúry VEGA MŠVVaŠ SR a SAV č. projektu 1/0368/2009.

LITERATÚRA

- [1] VEVERKA, Antonín: Technika vysokých napätí, SNTL/ALFA 1982, 04-527-82, 304 strán, 164 obrázkov, 23 tabuliek.
- [2] KOLCUNOVÁ, I., KVAKOVSKÝ, M., KURIMSKÝ, J.: Meranie povrchových výbojov na vn cievke točivého stroja. In: Elektroenergetika Journal. 4 (2009), ISSN 1337-6756.
- [3] MARTON, K., : Technika vysokých napätí 1, ALFA, Bratislava 1983.
- [4] KOLCUNOVÁ, Irida: Diagnostika elektrických strojov, Technická univerzita Košice 2006, ISBN 80-8073-550-6.

Použitie kapacitných senzorov pre snímanie čiastkových výbojov

Jaroslav PETRÁŠ, Jozef BALOGH, Jaroslav DŽMURA

Katedra elektroenergetiky, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

jaroslav.petras@tuke.sk, jozef.balogh@tuke.sk, jaroslav.dzmura@tuke.sk

Abstrakt — Na sledovanie aktivity čiastkových výbojov v izolačných systémoch elektroenergetických zariadení vysokého napätia sa používajú metódy priamej galvanickej metódy alebo alternatívne a doplnujúce metódy, ako sú nepriama induktívna metóda alebo akustická metóda. Meranie nepriamo metódou využíva senzory induktívne alebo kapacitné. V článku sú popísané kapacitné sondy využívané pre túto metódu.

Kľúčové slová — čiastkové výboje, kapacitné sondy

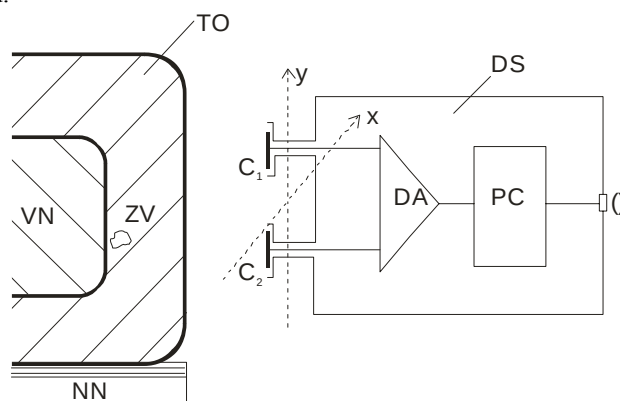
I. ÚVOD

Meranie čiastkových výbojov a výbojovej aktivity v izolačných systémoch patrí medzi dôležité metódy diagnostiky izolačných systémov. Táto metóda má už dnes miesto v súbore meracích diagnostických metód ako metóda merania čiastkových výbojov.

Jednou zo skupín snímačov čiastkových výbojov je skupina kapacitných sond. Signály emitované čiastkovými výbojmi v defektnom mieste izolačného systému vyvolávajú vznik rýchlych prechodných javov v elektrickom poli. Prechodné javy podobne, ako čiastkové výboje trvajú rádovo ns , a tieto musia byť spracované prístrojmi meracieho obvodu. Elektrické pole je snímané senzorom, ktorý tvorí kapacitná sonda zapojená na vstup rýchleho diferenciálneho zosilňovača. Vstup zosilňovača je navrhnutý takým spôsobom, aby boli v maximálnej možnej miere potlačené vplyvy cudzích polí (obr. 1) [17].

II. POUŽITIE KAPACITNÝCH SNÍMAČOV

Použitie kapacitných snímačov je analogické ako u induktívnych sond lineárneho typu. Kapacitné snímače umožňujú s vysokou presnosťou taktiež lokalizovať defektné miesto najmä v systémoch s epoxidovou izoláciou ako sú cievky vn strojov, suché transformátory, meniče napätia a prúdu apod.



Obr. 1 Princíp detekcie výbojov kapacitným snímačom

TO - testovaný objekt
DA - diferenciálny zosilňovač
VN - vysokonapäťová elektróda
PU - jednotka spracovania signálu

I - vysokonapäťová izolácia
ZV - zdroj čiastkových výbojov
DS - diferenciálna sonda
NN - zemná elektróda

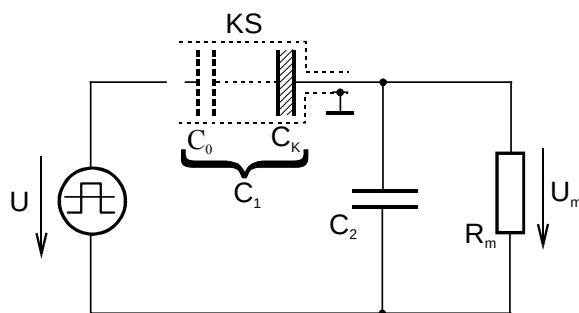
Naše výskumné experimenty boli orientované na porovnanie výsledkov meraní čiastkových výbojov priamou metódou, induktívnou metódou so snímačmi lineárneho a toroidálneho typu voči snímačom kapacitným a sledovalo sa správanie dvoch kapacitných sond:

1. Kapacitná sonda terčíková KST – kondenzátor je vytvorený terčíkom z (Cu) elektródy, ktorý so skúmanou izoláciou (kábel) a jej kovovým jadrom predstavuje kapacitu s prípadným defektným miestom.
2. Kapacitná sonda KS – miesto kovovej elektródy má osadený keramický kondenzátor diskového tvaru o kapacite $2000pF$. Pre tento účel sa hodia kondenzátory vyrobené z materiálu s vysokou relatívnou permitivitou (typu PERMITIT). Týmto sa upravujú elektrické vlastnosti vstupného pasívneho člena. Kondenzátorový disk obojstranne pokovovaný (Ag) umožňuje meranie priamym dotykom na uzemnenú časť vn zariadenia. Taktiež sa hodí aj na meranie vvn zariadení v epoxidovom prevedení.

III. ANALÝZA MODELU KAPACITNÉHO SNÍMAČA

Pri matematicko-fyzikálnej analýze snímačov čiastkových výbojov vychádzame z ich jednoduchých modelov, zostavených z diskretných RC prvkov. Predpokladáme pritom, že defektné miesto vyžaruje reálnu zložku intenzity elektrického poľa (E) ako aj poľa magnetického (H), úmerné veľkosti amplitúdy prúdového impulzu odpovedajúcemu čiastkovému výboju.

Náhradná schéma obvodu pozostáva z fiktívneho zdroja signálu U zo vstupnej RC časti nábojového zosilňovača, ako aj zo samotného nábojového zosilňovača. Jadro obvodu pritom tvorí kapacitná sonda C_1 (obr. 2).



Obr. 2 Náhradná schéma obvodu kapacitného snímača

kde: U - signál v tvare Diracovho impulzu vyvolaný defektným miestom,
 C_k - kapacita kondenzátorového snímača,
 C_0 - kapacita medzi snímačom a skúmaným objektom,
 C_2 - vstupná kapacita zosilňovača ($C_2 = 25 pF$),
 R_m - vstupný odpor zosilňovača ($R_m = 1M\Omega$).

Na obr. 2 je znázornený prípad, kedy je kondenzátorová sonda o kapacite C_k uložená nad meraným objektom čím sa vytvára kapacita C_0 (vzduchová medzera, čiastková kapacita izolácie a pod.). Výsledná kapacita C je daná sériovým zapojením oboch kondenzátorov.

Prenosová funkcia zjednodušeného usporiadania podľa obr. 1 v komplexnom tvare bude:

$$\frac{\overline{u_m}}{\overline{u}} = \frac{C_1}{C_1 + C_2} \frac{1}{1 + \frac{1}{j\omega R_m(C_1 + C_2)}} \quad (1)$$

Vstupné parametre zosilňovača sú spravidla: ($C_2 = 25 pF$) a ($R_m = 1M\Omega$). Hraničnú frekvenciu f_h obvodu vyjadríme vzťahom:

$$f_h = \frac{1}{2\pi R_m(C_1 + C_2)} \quad (2)$$

kde $R_m(C_1 + C_2) = \tau$.

Úpravou komplexného výrazu (1) získame rovnicu prenosovej funkcie pre pomer napätí:

$$\frac{U_m}{U} = \frac{C_1}{C_1 + C_2} \omega \tau \sqrt{\frac{1}{1 + (\omega \tau)^2}} \quad (3)$$

Odvođený výsledok poukazuje na to, že napätie na snímacom člene $R_m - C_2$ je určené okrem amplitúdy signálu U aj súčiniteľom $2\pi f R_m (C_1 + C_2)$.

Vychádzajúc z našich výpočtov a z experimentálnej práce, pracovná oblasť frekvencie leží v rozsahu $10^5 - 10^6$ Hz za podmienky, že volíme kombinácie kapacít $C_1 + C_2$ od $10^{-8} - 10^{11}$ F pri vstupnom odpore širokopásmového zosilňovača $R_m > 10^6 \Omega$.

IV. REAKCIA KAPACITNÉHO SENZORA NA KALIBROVANÉ IMPULZY

Experimentálne merania čiastkových výbojov si vyžadujú pred meraním na ľubovoľnom objekte vykonať jeho testovanie. Bol preto vytvorený vodič (anténa), do ktorého po impedančnom prispôbení bol injektovaný signál z profesionálneho kalibrátora.

Boli použité tri typy kalibrátorov:

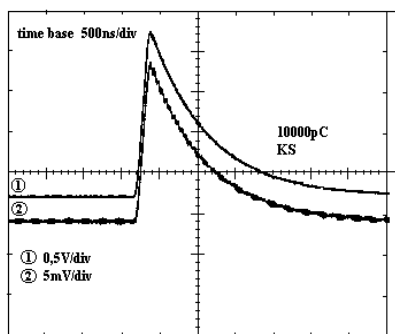
- TETTEX typ 0216 s rozsahom $10 - 100 - 1000 - 10000$ pC,
- kalibrátor TUR typ PET2 s rozsahom $5 - 25 - 50 - 250$ pC,
- kalibrátor, vyvinutý na pracovisku s rozsahom $25 - 50 - 75 - 100 - 250 - 500 - 1000 - 7500 - 10000$ pC.

Odozvy z tvarov impulzov kalibrátora TETTEX typ 0216 Zürich sú znázornené na obr. 3 a obr. 4.

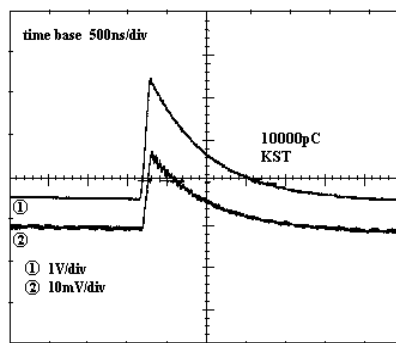
Testovanie bolo realizované na už popísaných sondách:

- kapacitnej sonde s dielektrikom o $\epsilon_r = 500$ (skratka KS),
- kapacitnej sonde s Cu terčíkom (skratka KST).

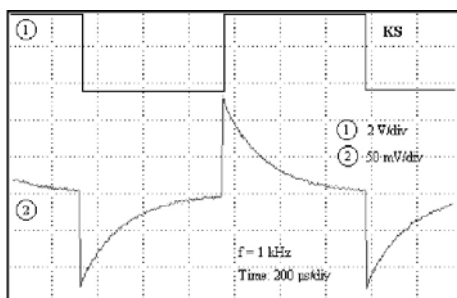
Ak porovnáme zachytený signál voči vnútenému, vidíme, že vnútený signál odpovedá zachytenému signálu, pričom hodnota výstupného signálu je vyššia pri použití kondenzátorového terčika (KS) voči terčíku s Cu (KST).



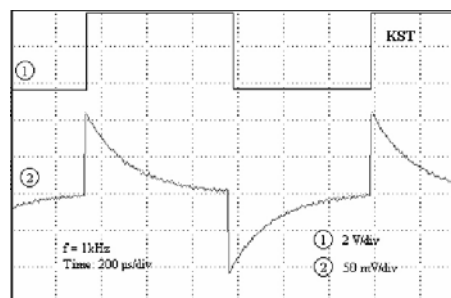
Obr. 3 Tvar impulzu z kalibrátora testovaný sondou KS, 1) injektovaný impulz 2) odozva injektovaného impulzu



Obr. 4 Tvar impulzu z kalibrátora testovaný sondou KST, 1) injektovaný impulz 2) odozva injektovaného impulzu



Obr. 5 Tvar a odozva signálu typu Diracovho impulzu na výstupe sondy KS

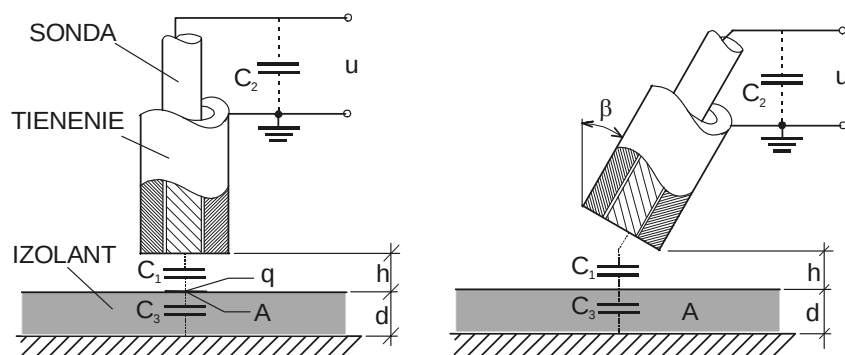


Obr. 6 Tvar a odozva signálu typu Diracovho impulzu na výstupe sondy KST

Tvar a odozva signálu typu Diracovho impulzu na výstupe kapacitnej sondy je znázornená obr. 5 a 6.

V. TESTOVANIE V ZÁVISLOSTI OD POLOHY SONDY

V praxi sa často stretávame s problémom na zakrivenom (nerovnom) povrchu. Vtedy sa nemôžeme vyhnúť natočeniu sondy do nejakého uhla β oproti kolmému uloženiu vzhľadom k povrch. Takáto situácia je zobrazená na obr. 7.



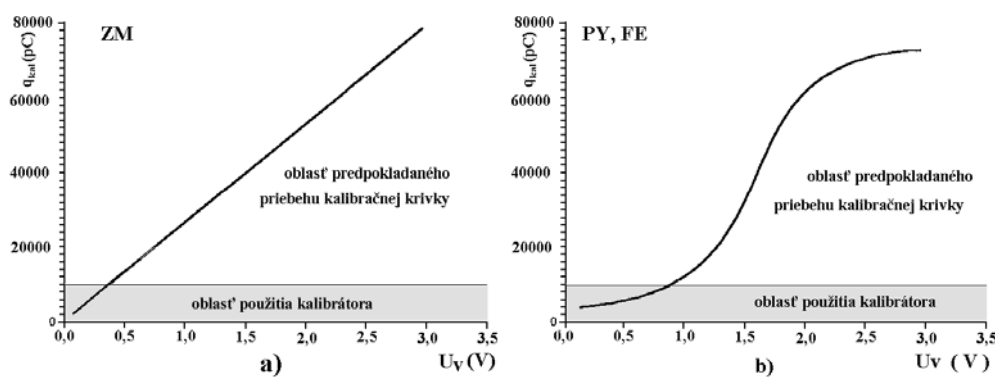
Obr. 7 Poloha sondy vzhľadom k povrchu

Ak porovnáme tieto dva obrázky, vidíme, ekvivalentné náhradné zapojenie ostáva rovnaké. S uhlom β by sa ale mohla meniť veľkosť meranej plochy A , a teda aj kapacity C_1 a C_3 . Na overenie týchto predpokladov boli vykonané testy na rovnakých vzorkách. Hodnoty C_1 , C_3 a A boli určené ako funkcia uhla β . Bola na to použitá dvojkroková metóda. Z týchto pozorovaní vyplynulo, že 30° je odchýlka menšia ako rozptyl vo výsledkoch. Do 45° , čo by bol extrémny prípad pri meraní, odchýlka nepresahuje 15%. S istotou môžeme teda povedať, že meranie pod uhlom nepredstavuje vážne problémy.

VI. KALIBRÁCIA MERACIEHO SYSTÉMU

Kalibrácia je neoddeliteľnou súčasťou merania. Určí sa ňou závislosť medzi zdánlivým nábojom (pC) čiastkových výbojov a veľkosťou amplitúdy nasnímaného napätia (V) jednotlivých nameraných výbojov. Kalibrácia je založená na injektovaní náboja známej veľkosti z kalibrátora čiastkových výbojov do meraného objektu a následnom zaznamenaní veľkosti amplitúdy snímacím obvodom. Objekt musí byť odpojený od testovacieho napätia. Z nameraných hodnôt je potrebné urobiť kalibračné krivky, čiže závislosť $q=f(U)$.

Pri aplikácii nepriamej metódy je zatiaľ možné merať s uspokojivou presnosťou len čiastkové výboje, ktorých hodnota nepresiahne $10000pC$, pretože nie sme schopní predvídať a matematicky popísať priebeh kalibračných kriviek pri vyšších hodnotách náboja. Predpokladaný priebeh kalibračných kriviek pri nepriamej metóde v porovnaní s priamou metódou je na (obr. 8).



Obr. 8 Predpokladané priebehy kalibračných kriviek (ak $q_{kal} > 10000pC$) pri a) priamom a b) nepriamom spôsobe snímania čiastkových výbojov [2]

Pri použití priamej (galvanickej) metódy majú kalibračné krivky lineárny priebeh. Pri meraní nepriamou (induktívnou) metódou majú nelineárny priebeh. Nelinearita je daná vlastnosťami použitých funkčných materiálov u jednotlivých sond [2].

Kalibrácia bola vykonaná pomocou kalibrátora firmy TETEX INSTRUMENTS typu 9216. Kalibrátor sme paralelne pripojili k meranému objektu. Do vzorky boli injektované tieto hodnoty náboja: $10pC$, $100pC$, $1000pC$, $1000pC$. Táto kalibrácia bola urobená pre induktívnu sondu typu PERMALOY pri maximálnom zosilnení 60dB. Impulzy pri injektovaní nábojov veľkosti $1000pC$ a $10000pC$ pri tomto zosilnení nedosiahli maximálnu hodnotu. Boli orezané, čo bolo spôsobené výstupom zosilňovača (max. 3,5V).

Novú kalibráciu je nutné urobiť pri každej zmene meraného objektu (aj polohy), snímačej sondy resp. meracej metódy, ako aj zosilnenia snímaného signálu. Je totiž závislá na kapacitných pomeroch celého zapojenia a umiestnenia. Pre správnosť vyhodnotenia je dôležité poznať tvar kalibračných kriviek. Ich tvar nie je pre jednotlivé sondy rovnaký.

Musíme však poznamenať, že z toho dôvodu kvôli väčšej presnosti je výhodnejšia kalibrácia pomocou kalibrátora s citlivejším rozsahom. Zo štyroch hodnôt kalibračných impulzov len veľmi ťažko dokážeme vystihnúť tvar jednotlivých kriviek.

VII. ZÁVER

Kapacitné senzory je možné využiť pri snímaní aktivity čiastkových výbojov nepriamou metódou ako alternatívu k indukčným senzorom. Popísali sme spôsob ich aplikácie, analyzovali sme model takejto sondy, aby sme mohli matematicky určiť charakteristiky sondy. Venovali sme pozornosť aj vplyvu polohy sondy k meranému objektu.

POĎAKOVANIE



Táto práca bola podporovaná Agentúrou Ministerstva školstva Slovenskej republiky pre štrukturálne fondy EÚ na základe projektu Vývoj unikátneho nízkoenergetického statického zdroja pre elektrosystémy (číslo projektu: 26220220029, prioritná os 2 Podpora výskumu a vývoja).

LITERATÚRA

- [1] Partial Discharge Measurements, IEC 60270, 1998.
- [2] Kršňák, I.: Detekcia čiastkových výbojov a analýza vypovedeschopných veličín, Doktorská dizertačná práca, Košice, 1996.
- [3] IEEE Guide for Partial Discharge Measurement in Liquid – Filled Power Transformers and Shunt Reactors, New York, July 2, 1992.
- [4] Kučinskij, G. S.: Čističnyje razrjady v vysokovoľtných konstrukciach, Izdatel'stvo Energia, Leningrad, 1979.
- [5] Balogh, J.: Snímanie čiastkových výbojov na rozšírenom RLC modeli, (Partial discharge scanning on extended RLC model), Zborník “ 9. MEDZINÁRODNÁ VEDECKÁ KONFERENCIA TECHNIKA VYSOKÝCH NAPÄTÍ “ Košice, september 1999 s. 81 – 87.
- [6] Böning, A.: Luftgehalt und Luftspalt verteilung geschichter Dielektrika I., Archiv für Elektrotechnik, 1963, č. 1 Springer Verlag.
- [7] Heller – Veverka, A.: Technika vysokých napětí, SNTL/ALFA Praha, 1978.
- [8] Alsheikhly, A., Kranz, H. G.: A new Approach for a Basic Understanding of PD – Phenomena, 7th ISH Dresden, 1991, Vol. 7, 71.04, p. 21.
- [9] Wlodek, R.: Mechanizmy działania wyładowan niezupełnych w dielektrykach, Wydawnictwo naukowe PWN, Warszawa, 1992.
- [10] Marton, K.: Elektrofizikálne základy snímania čiastkových výbojov senzormi kapacitnými a indukčnými, In: Proc. VI. SYMPOZIJUM, Problemy eksploatacji układów izolacyjnych wysokiego napięcia, AGH Krakow, 1997, str. 205 – 214.
- [11] Záliš, K.: Vývoj nového měřiče částečných výbojů, Zborník “ 9. MEDZINÁRODNÁ VEDECKÁ KONFERENCIA TECHNIKA VYSOKÝCH NAPÄTÍ “ Košice, september 1999 s. 26 - 31.
- [12] Valenta, L.: Měření částečných výbojů a revize norem IEC 60270 A IEC 60076 – 3, Zborník “ 9. MEDZINÁRODNÁ VEDECKÁ KONFERENCIA TECHNIKA VYSOKÝCH NAPÄTÍ, Košice, 20. – 21. September 1999.
- [13] Weber, H. J.: Technika merania čiastkových výbojov, Tettex Instrument – Information, č.407.
- [14] Marton, K., Balogh, J., Gargušová, T.: Diagnostika VN a VVN prvkov a zariadení. Zborník z konferencie “Elmat 96” TU FEI, Ferrost a.s., Košice 1996.
- [15] Dakin, T.W. et al.: IEEE Trans. on Power Apparatus and Systems, Vol. PAS-88, pp. 251-257, 1968.
- [16] Zaengel, W.: 6-th ISH, New Orleans 1989, Part 43. 07.
- [17] Lemke, E., Schmiegell, P.: Progres in PD-probe Measuring Technique, 7-th ISH, Dresden., 1991, Part 72.01.
- [18] Fuhr, J.: 7-th ISH, Dresden, 1991, Part 73.01.
- [19] Marton, K.: Teoretická analýza indukčného snímania čiastkových výbojov. KTVN, EF TU v Košiciach.
- [20] Marton, K. a kol.: Autorské osvedčenie č. 202 129, Úrad pro vynálezy a objevy, Praha, 1976.
- [21] Marton, K. a kol.: Výskumná správa GAV č.1/990345/92, EF TU Košice, 1992.
- [22] Marton, K., Balogh, J.: Elektrofizikálna analýza odozvy čiastkových výbojov modelovaných Diracovým impulzom. Zborník, Diagnostika 93, ZČU Plzeň, 1993.
- [23] Balogh, J.: Lineárne a toroidálne sondy a ich aplikácie v elektroenergetike, Pisomná práca k rigoróznjej skúške, Košice 1997.

Morphological Complexity of State Space Quaternary Memory

Milan GUZAN, Radoslav BUČKO

Dept. of Theoretical Electrotechnics and Electrical Measurement, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

milan.guzan@tuke.sk, radoslav.bucko@tuke.sk

Abstract — Contribution highlights the need to get to know the morphology of the state space for the simulation and implementation of new semiconductor memories. Graphic presentation of the boundary surface refers to the morphological complexity and complications in the state space, which generally can be expected in the analysis of multi-state autonomous circuits. Therefore every other newly designed structure is needed to analyze, but it is beyond the scope of this contribution.

Keywords — memory, boundary surface, stable limit cycle, singularity

I. INTRODUCTION

In the works [1] - [4] were published only in such cases of flip-flop circuits, when only one eigenvalue of the Jacobi matrix for the saddle was positive and the other had always negative real parts. This applied also to the of fourth-order system, which was published in [5]. Therefore the authors of work [5] concluded that there is a general feature of the saddle point also in the case of higher order. This conclusion was also reached also on the basis of the geometric interpretation of boundary surface (*BS*). Results reached from activities [6], [7] showed that unstable singularities in the case with the negative load need not have the character of the saddle, as was in analyzed structures with a positive load. In the works [6], [7] were presented ternary memories, where the element and also the load were characterized by negative differential resistance area (resonant tunnelling diodes - RTD) and unstable singularities become unstable nodes. Due to changes singularities from the saddle to unstable node could not be used graphical methods for calculation and displaying the morphology of *BS*. The answer brought only just computer simulation by using the grid technique [8], which indicated segmentation the state space into the regions of attraction for individual attractors. Therefore the question has arisen: What will the boundary surface look like for memory characterized by the four stable singularities, where the load will again be characterized by negative differential resistance area? Will there be present undesirable oscillations as in the case of a ternary memory, what will be make morphology of *BS* even more complicated?

II. MVL MEMORY CIRCUIT

Multiple valued logic (MVL) memory circuit is shown in Fig. 1. Symbols of non-linear elements correspond to RTDs.

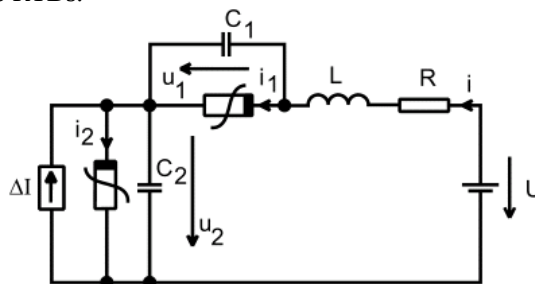


Fig. 1 MVL memory circuit

Circuit in Fig. 1 we can describe by the system (1) where the characteristics of nonlinear elements $f_k(u_k)$ according to [9] are defined by (2) where $^k g_i$ are conductivities of the k -th

element and kU_i are breakpoints of characteristics shown in Figure 2. If $k=1$ it is a load, if $k=2$ it is the element.

$$\begin{aligned} L \left(\frac{di}{dt} \right) &= U - Ri - (u_1 + u_2) \equiv Q_1 \\ C_1 \left(\frac{du_1}{dt} \right) &= i - f_1(u_1) \equiv Q_2 \\ C_2 \left(\frac{du_2}{dt} \right) &= i - f_2(u_2) + \Delta I \equiv Q_3 \end{aligned} \quad (1)$$

$$\begin{aligned} f_k(u_k) &= \frac{1}{2} ({}^k g_0 + {}^k g_3) u_k + \frac{1}{2} [({}^k g_1 - {}^k g_0) |u_k - {}^k U_1| + \\ &({}^k g_2 - {}^k g_1) |u_k - {}^k U_2| + ({}^k g_3 - {}^k g_2) |u_k - {}^k U_3|] - \\ &\frac{1}{2} [({}^k g_1 - {}^k g_0) {}^k U_1 + ({}^k g_2 - {}^k g_1) {}^k U_2 + ({}^k g_3 - {}^k g_2) {}^k U_3] \end{aligned} \quad (2)$$

Capacity C_1 , C_2 includes the capacity of the equivalent circuit of elements, or parasitic capacity on the chip. Inductance L is the inductance of inputs to the elements and resistor R expresses resistance of conductive connections on the chip. Supply voltage $U=440mV$ and for now we consider control pulse [10] $\Delta I=0$ and $R=0$.

IV characteristics of the element and the load in Figure 2 correspond to the parameters specified in Table 1a). Graphical representation of projection of the IV characteristics and singularities in the plane i, u_2 is shown in the Fig. 2b). Singularities $S1$, $S2$ and $S3$ are stable and $N1$, $N2$ are unstable.

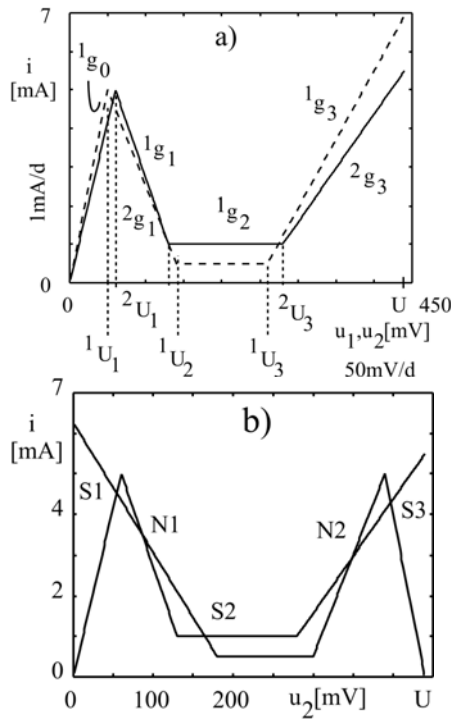


Fig. 2 a) IV characteristics of the RTD element (solid line) and load (dashed line), when $\Delta I = 0$. Parameters the both elements are listed in Table 1a
 b) Graphical interpretation of the circuit in Figure 1 in the projection into the plane i, u_2

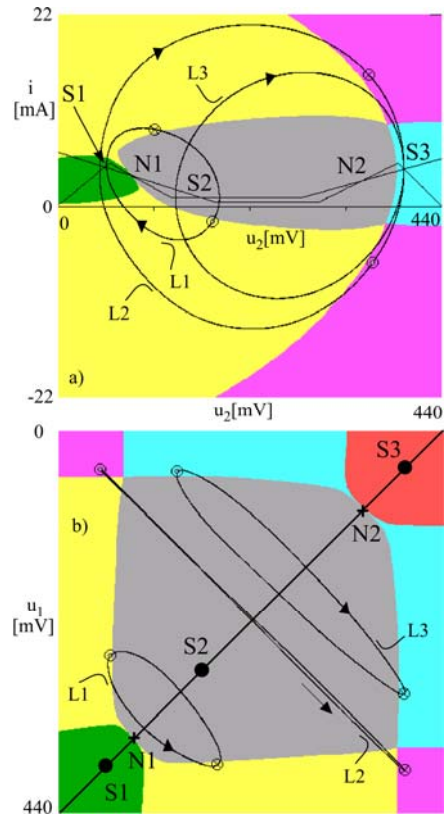


Fig. 3 Cross-sections of the BS through the saddle $N1$ in the Monge projections

Table 1
 COORDINATES OF SINGULARITIES AND PARAMETERS OF SEGMENTS, WHICH CREATE THE APPROPRIATE SINGULARITY

	g_0 [S]	g_1 [S]	g_2 [S]	g_3 [S]	U_1 [mV]	U_2 [mV]	U_3 [mV]
a) element	0,0833	-0,0571	0	0,0281	60	130	280
load	0,1	-0,05	0	0,032	50	140	260
b) element $\equiv$ load	0,0833	-0,0571	-0,012	0,0281	60	120	250

In work [7] was presented the case (Fig. 3) for $L = 1 \cdot 10^{-10}H$, $C1 = C2 = 5 \cdot 10^{-13}F$ and the parameters listed in Table 1a), when the circuit shown in Figure 1 was characterized not only by three colored regions of attraction for attractors $S1$, $S2$ and $S3$ (green, gray and red), but even the other three areas corresponding to the three stable limit cycles (SLC) $L1$, $L2$ and $L3$ (yellow, purple and light blue). For completeness, it should be added that in Fig. 3b the projection of stable singularities $S1$, $S2$ and $S3$ into the plane u_1, u_2 are indicated by symbol •

and unstable singularities $N1$, $N2$ are indicated by symbol $+$. As is evident from Figure 3, the presence of SLC morphologically considerably complicates the BS . Therefore questions mentioned above were raised.

III. QUATERNARY MEMORY

In the design of the quaternary memory was used the same relation (2), as well as for ternary memory, while IV characteristic of the element is identical to characteristic of the load. Parameters are given in Tab. 2b). IV characteristics of the element and the load together with created singularities in the projection into plane i, u_2 are shown in Fig. 4. Fig. 4b) shows cross-section of the BS at $i = 0$, where the symbols $\bullet$ or $+$ again correspond to stable ($S1$, $S2$, $S3$, $S4$), or unstable ($N1$, $N2$, $N3$) singularity. Color range of attraction regions corresponds to the location of descriptions the singularities $S1$ - $S4$ and $L1$ - $L4$ SLC in colored areas, so e.g. the region of attraction for $S3$ is red and the region of attraction for SLC $L4$ is white. As is evident from Fig.4b), without the presence of SLC , memory would be quaternary, but due to presence up to four SLC memory is finally characterised by eight steady state. Due to the symmetry of the circuit, BS is also symmetric with respect to projection of the IV characteristics into the plane u_1, u_2 .

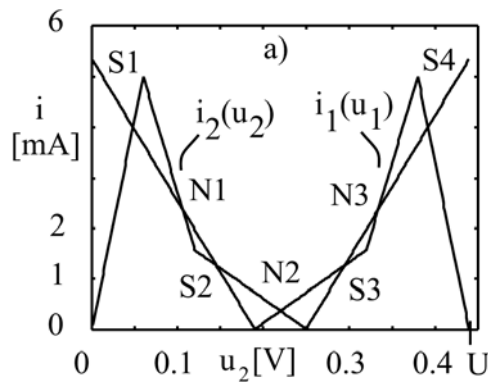


Fig. 4a) Projection of the IV characteristics and depiction the singular points into the plane i, u_2

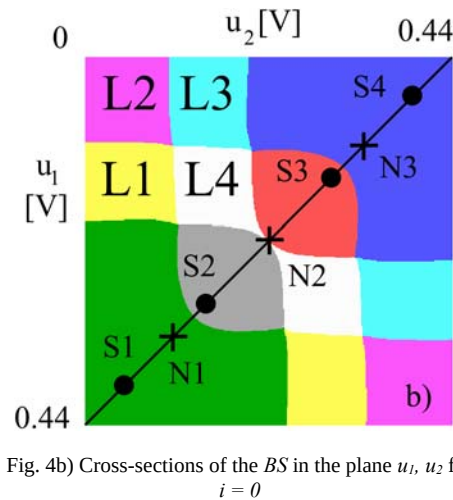


Fig. 4b) Cross-sections of the BS in the plane u_1, u_2 for $i = 0$

For a coherent picture how this BS in 3D space looks, there were made cross-sections through BS in the range from -14 mA to $+14$ mA with step 1 mA. Figure 5 shows some cross-sections through BS selected to provide a comprehensive picture of the BS morphology in 3D space. Denotation of singularities, description of axes and color scale in Fig. 5 is the same as in Fig.4b). From Fig. 5 it is clear that the regions of attraction for $S1$ - $S4$ as well as for $L4$, gradually vanish with increasing / decreasing current. It means that they are encapsulated in regions of attraction for the SLC . When $i = -14$ and $+14$ mA regions of attraction for the dynamic attractors $L1$, $L2$ and $L3$ are already unambiguously dominant regions of attraction.

IV. CONCLUSION

Increasing the number of singularities in the state space brings more and more regions of attraction. This is not negative fact. Negative aspect is only the existence of SLC , which prevents using the memory in circuits. The cross-sections of BS show the attractivity of particular stable singularities, but also the attractivity of dynamic attractors - SLC , whose regions of attraction cover almost half of the state space. Although here is presented quaternary memory, graphic presentation of the BS indicates possible complications in the state space, which generally can be expected at analysis of similar multi-state autonomous circuits. Therefore every other newly designed structure is needed to analyze, but it is beyond the scope of this contribution and it may be subject of further research.

ACKNOWLEDGMENT



We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centre of Excellence of the Integrated Research & Exploitation the Advanced Materials and Technologies in the Automotive Electronics", ITMS 26220120055.

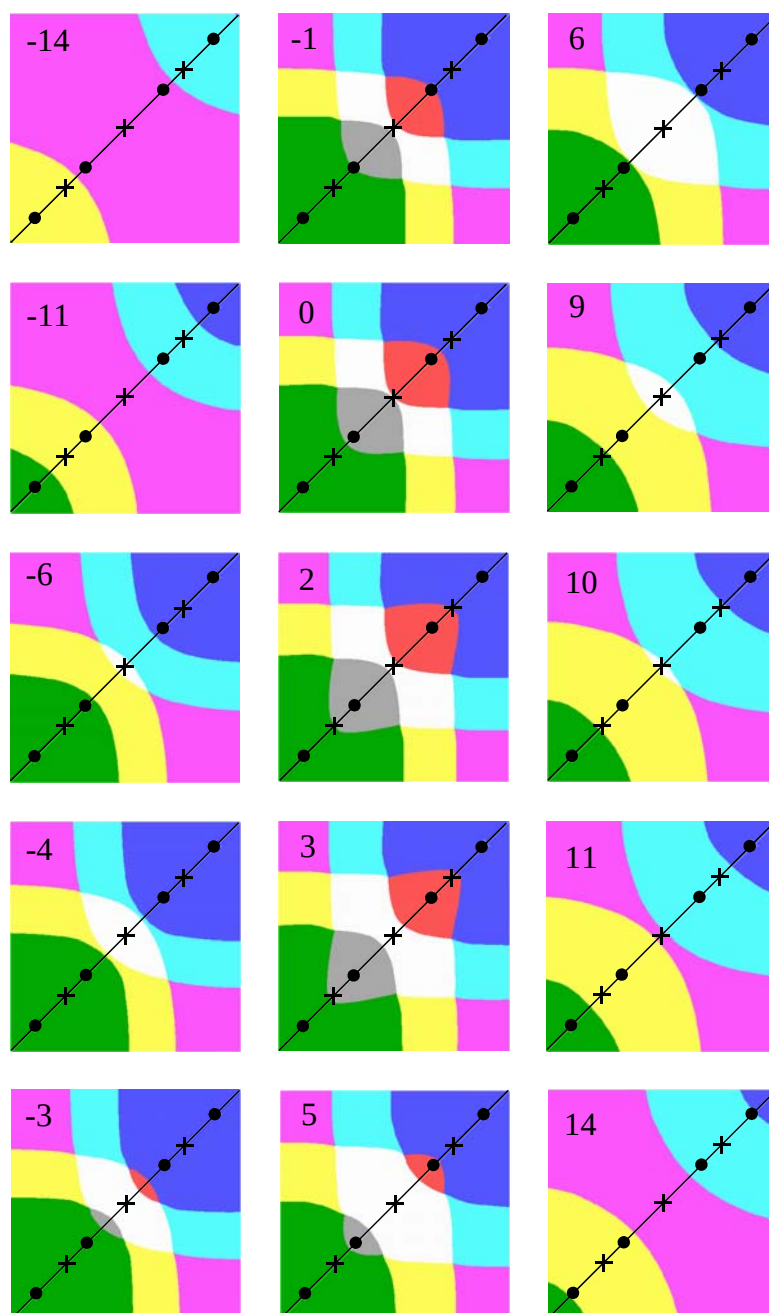


Fig. 5 The projections of cross-sections into the plane u_1, u_2 . Current (in mA) is denoted in the figure for the parameters listed in Tab. 1b).

REFERENCES

- [1] Špány, V., Pivka, L.: *Boundary Surfaces in Sequential Circuits*. *International Journal of Circuit Theory and Applications*, vol. 18, 1990, p. 349-360.
- [2] Špány, V., Pivka, L.: *Invariant Manifolds in sequential circuits*. *Elektrotechnický časopis*, roč.42, 1991, č.6, p. 281-293.
- [3] Špány, V., Pivka, L.: *2-Segment Bistability and Basin Structure in 3-Segment PWL Circuits*. *IEE Proceedings-G*, Vol. 140, 1993, No. 1, p. 61-67.
- [4] Pivka, L., Špány, V.: *Boundary Surfaces and Basin Bifurcations in Chua's Circuit*. *Journal of Circuits, Systems and Computers*, Vol. 3, 1993, No. 2, p. 441.
- [5] Špány, V., Pivka, L.: *Boundary surfaces in sequential circuits*. *Beitrage zur Theoretischen Elektrotechnik, Das Internationale Symposium in Ilmenau*, 1988, p.90-99.
- [6] Špány, V., Galajda, P., Guzan, M.: *Boundary Surfaces of One-port Memories*. *5th International Conference Tesla Millennium*, Beograd, October 1996, p.130-137.
- [7] Galajda, P., Guzan, M., Špány, V.: *The State Space mystery with Negative Load*. *Radioengineering*, Vol. 8, 1999, No. 2, p. 2-7.
- [8] M. Guzan, V. Špány, P. Galajda, R. Bučko: *Visualization of the Boundary Surface of Multi-value Cell Using* (In Slovak), <http://www.posterus.sk/?p=9487>
- [9] Špány, V.: *The Expression of the Non-Linear Characteristics by Means of Absolutes Values*, *Slaboproudý obzor*, Vol. 40, No. 7, 1988, pp. 354-356, (In Slovak).
- [10] Guzan, M., Galajda, P., Špány, V.: *The space mystery in multivalued logic circuits*. *Konf. DSP, Herľany*, 1999, p.151-154.

The Crossing Numbers of Join of Special Graphs

Marián KLEŠČ, Daniela KRAVECOVÁ, Jana PETRILLOVÁ

Dept. of Mathematics and Theoretical Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

marian.klesc@tuke.sk, daniela.kravecova@tuke.sk, jana.petrilova@tuke.sk

Abstract – The investigation on crossing numbers of graphs is a classical and however very difficult problem. It is well known that this problem is NP-complete. The problem of reducing the number of crossings in the graph is studied not only in the graph theory, but also by computer scientists. There are only few results concerning crossing numbers of join of some graphs. In this paper, we give the crossing numbers of join products of the special graph G on six vertices with n isolated vertices as well as with the path P_n on n vertices and with the cycle C_n .

Keywords – crossing number, cycle, drawing, graph, join product, path

I. INTRODUCTION

Introduction of the VLSI technology revolutionized circuit design and had a strong impact on parallel computing. A lot of research aiming at efficient use of the new technologies has been done and further investigations are in progress. As a crossing of two edges of the communication graph requires unit area in its VLSI-layout, the crossing number together with the number of vertices of the graph immediately provide a lower bound for the area of the VLSI-layout of the communication graph. The crossing numbers has been also studied to improve the readability of hierarchical structures. The visualized graph should be easy to read and understand. For the understandability of graph drawings, the reducing of crossings is by far the most important. The problem of reducing the number of crossings was therefore studied not only by the graph theory community, and by VLSI communities, but also by computer scientists.

Let G be a graph, whose vertex set and edge set are denoted by $V(G)$ and $E(G)$, respectively. A drawing of G is a representation of G in the plane such that its vertices are represented by distinct points and its edges by simple continuous arcs connecting the corresponding point pairs. For simplicity, we assume that in a drawing (a) no edge passes through any vertex other than its end-points, (b) no two edges touch each other (i.e., if two edges have a common interior point, then at this point they properly cross each other), and (c) no three edges cross at the same point. The *crossing number* $cr(G)$ is the smallest number of edge crossings in any drawing of G . It is easy to see that a drawing with minimum number of crossings (an *optimal* drawing) is always a *good* drawing, meaning that no edge crosses itself, no two edges cross more than once, and no two edges incident with the same vertex cross each other.

Crossing number problems were introduced by Turán [19], who first inquired about the crossing number of the complete bipartite graph $K_{m,n}$. Turán devised a natural drawing of $K_{m,n}$ with $\lfloor \frac{m}{2} \rfloor \lfloor \frac{m-1}{2} \rfloor \lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor$ crossings, but the conjecture of Zarankiewicz [19] that such a drawing is the best possible, is still open. Crossing number problems are, in general, very difficult. It was proved by Garey and Johnson [5] that computing the crossing number of a graph is an NP-complete problem. The exact values of crossing numbers are known only for few specific families of graphs. The Cartesian product is one of few graph classes, for which exact results concerning crossing numbers are known. Harary *et al.* [7] conjectured that the crossing number of $C_m \times C_n$ is $(m-2)n$, for all m, n satisfying $3 \leq m \leq n$. This has been proved only for m, n satisfying $n \geq m$, $m \leq 7$. It was recently proved by Glebsky and Salazar [6] that the crossing number of $C_m \times C_n$ equals its long-conjectured value at least for $n \geq m(m+1)$. Besides Cartesian product of two cycles, there are several other exact results. In [2] and [10], the crossing numbers of $G \times C_n$ for all graphs G of order four are given. Bokal in [3] confirmed the general conjecture for the crossing number of Cartesian product of path and star formulated in [10]. The table in [13] shows the summary of known crossing numbers for Cartesian products of path, cycle and star with connected graphs of order five.

Zarankiewicz's conjecture [20] has been verified by Kleitman for $\min\{m, n\} \leq 6$, see [11]. In 1986 K. Asano started to study crossing numbers of multipartite complete graphs. In [1] he established the crossing numbers of the tripartite graphs $K_{1,3,n}$ and $K_{2,3,n}$. For the graph $K_{1,4,n}$, the crossing number was given independently in [8] and [9]. The crossing number of the graph $K_{1,5,n}$ was established in [18].

The join product of two graphs G_1 and G_2 , denoted by $G_1 + G_2$, is obtained from vertex-disjoint copies of G_1 and G_2 by adding all edges between $V(G_1)$ and $V(G_2)$. For $|V(G_1)| = m$ and $|V(G_2)| = n$, the edge set of $G_1 + G_2$ is the union of disjoint edge sets of the graphs G_1 , G_2 , and the complete bipartite graph $K_{m,n}$. Kulli and Muddebihal [17] gave the characterization for all pairs of graphs which join product is planar graph. Using Kleitman's result [11], the crossing numbers for join of two paths, join of two cycles, and for join of path and cycle were studied in [14]. Moreover, the exact values for crossing numbers of $G + P_n$ and $G + C_n$ for all graphs G of order at most four are given in [16]. Very recently, the crossing numbers of join of the special graph on six vertices with path and cycle were collected in [15].

Let $D(D(G))$ be a good drawing of the graph G . We denote the number of crossings in D by $cr_D(G)$. Let G_i and G_j be edge-disjoint subgraphs of G . We denote the number of crossings between edges of G_i and edges of G_j by $cr_D(G_i, G_j)$, and the number of crossings among edges of G_i in D by $cr_D(G_i)$. It is easy to see that for three edge-disjoint graphs G_i , G_j , and G_k , the following equations hold:

$$\begin{aligned} cr_D(G_i \cup G_j) &= cr_D(G_i) + cr_D(G_j) + cr_D(G_i, G_j), \\ cr_D(G_i \cup G_j, G_k) &= cr_D(G_i, G_k) + cr_D(G_j, G_k). \end{aligned}$$

Let nK_1 denote the graph on n isolated vertices and let P_n and C_n be the path and the cycle on n vertices, respectively. In this paper, we determine the crossing number for the join of the graph nK_1 with the special graph G on six vertices shown in Figure 1. This result enables us, in Sections III and IV, to give the crossing numbers of the graphs $G + P_n$ and $G + C_n$.

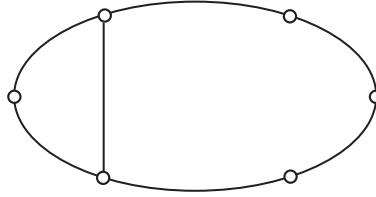


Fig. 1 The graph G on six vertices

In the paper, some proofs are based on Kleitman's result on crossing numbers of complete bipartite graphs. More precisely, he proved that

$$cr(K_{m,n}) = \left\lfloor \frac{m}{2} \right\rfloor \left\lfloor \frac{m-1}{2} \right\rfloor \left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor, \quad \text{if } m \leq 6.$$

In the proofs of the paper, we will often use the term "region" also in nonplanar drawings. In this case, crossings are considered to be vertices of the "map".

II. THE JOIN PRODUCT WITH DISCRETE GRAPHS

The graph G in Figure 1 consists of one 6-cycle $C_6(G)$ and one additional edge which together with the edges of $C_6(G)$ forms one 3-cycle and one 5-cycle. The graph $G + nK_1$ consists of one copy of the graph G and n vertices $f_1, f_2, \dots, f_n$, where every vertex f_i , $i = 1, 2, \dots, n$, is adjacent to every vertex of G . For $i = 1, 2, \dots, n$, let F^i denote the subgraph induced by six edges incident with the vertex f_i . In Figure 2 one can easily see that

$$G + nK_1 = G \cup K_{6,n} = G \cup \left(\bigcup_{i=1}^n F^i \right).$$

The graph $G + K_1$ is planar. The graph $G + 2K_1$ contains the complete graph K_5 as a subgraph. As $cr(K_5) = 1$, for $n = 1$ and 2 we have the next result.

Lemma 2.1: $cr(G + K_1) = 0$ and $cr(G + 2K_1) = 1$.

To prove the main result of this section, several partial results concerning various drawings of the graph $G + nK_1$ are needed.

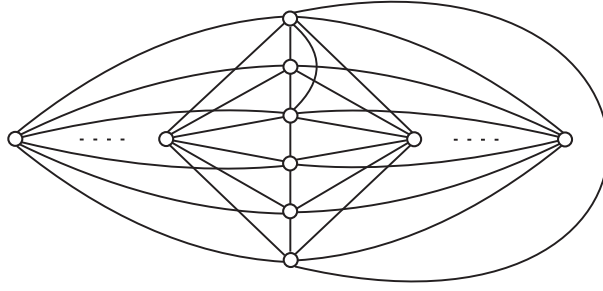


Fig. 2 The drawing of the graph $G + nK_1$

Lemma 2.2: Let D be a good drawing of the graph $G + nK_1$, $n \geq 3$, in which there are two different subgraphs F^i and F^j with $cr_D(F^i, F^j) = 0$. Then there are at least $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings in D .

Proof. The proof proceeds by induction on n . By Lemma 2.1, every good drawing of the graph $G + nK_1$ has at least $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings for $n = 1$, and $= 2$. In the rest of the proof, only good drawings of $G + nK_1$ in which there are two different subgraphs F^i and F^j with $cr_D(F^i, F^j) = 0$ are considered. Assume now a good drawing D of the graph $G + nK_1$, $n \geq 3$, with less than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings and let for every integer s , $s < n$, any good drawing of the graph $G + sK_1$ has at least $6\lfloor \frac{s}{2} \rfloor \lfloor \frac{s-1}{2} \rfloor + \lfloor \frac{s}{2} \rfloor$ crossings. The subdrawing of D induced by $F^i \cup F^j$ with $cr_D(F^i, F^j) = 0$ divides the plane in such a way that there are exactly two vertices of the graph G on the boundary of every region. Hence, one can easily verify that, in D , the edges of G cross the edges of $F^i \cup F^j$ at least once and that $cr_D(G, F^i \cup F^j) \geq 1$. Moreover, as $cr(K_{6,3}) = 6$, $cr_D(F^k, F^i \cup F^j) \geq 6$ for every $k = 1, 2, \dots, n$, $k \neq i, j$.

As $G + nK_1 = (G + (n-2)K_1) \cup (F^i \cup F^j)$ and $G + (n-2)K_1 = K_{6,n-2} \cup G$, for the number of crossings in D we have

$$\begin{aligned} cr_D(G + nK_1) &= cr_D(G + (n-2)K_1) + cr_D(F^i \cup F^j) + cr_D(K_{6,n-2}, F^i \cup F^j) + cr_D(G, F^i \cup F^j) \\ &\geq 6\lfloor \frac{n-2}{2} \rfloor \lfloor \frac{n-3}{2} \rfloor + \lfloor \frac{n-2}{2} \rfloor + 6(n-2) + 1 \geq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor. \end{aligned}$$

This contradiction with the assumption that D has less than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings completes the proof. $\square$

Lemma 2.3: Let D be a good drawing of the graph $G + nK_1$, $n \geq 3$, in which some subgraph $G \cup F^i$ is crossed at least four times by every subgraph F^j , $j = 1, 2, \dots, n$, $j \neq i$. Then there are at least $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + 2\lfloor \frac{n}{2} \rfloor$ crossings in D .

Proof. Assume a good drawing D of the graph $G + nK_1$ in which for some $i \in \{1, 2, \dots, n\}$ every subgraph F^j , $j = 1, 2, \dots, n$, $j \neq i$, crosses the subgraph $G \cup F^i$ at least four times. We know that $G + nK_1 = K_{6,n-1} \cup (G \cup F^i)$ and that the graph $K_{6,n-1}$ contains $n-1$ subgraphs F^j , $j \neq i$. Regardless of the number of crossings among the edges of $G \cup F^i$, for the number of crossings in D we have the following.

$$\begin{aligned} cr_D(G + nK_1) &= cr_D(K_{6,n-1}) + cr_D(G \cup F^i) + cr_D(K_{6,n-1}, G \cup F^i) \\ &\geq 6\lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n-2}{2} \rfloor + 4(n-1) \geq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + 2\lfloor \frac{n}{2} \rfloor. \end{aligned}$$

This completes the proof. $\square$

The next Theorem 2.1 is the main result of this section.

Theorem 2.1: $cr(G + nK_1) = 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ for $n \geq 1$.

Proof. By Lemma 2.1, Theorem 2.1 is true for $n = 1$ and 2 . Assume $n \geq 3$ and consider that there is a good drawing D of the graph $G + nK_1$ with less than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings. As

$$\begin{aligned} cr_D(G + nK_1) &= cr_D(K_{6,n}) + cr_D(G) + cr_D(K_{6,n}, G) \\ &\geq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + cr_D(G) + cr_D(K_{6,n}, G), \end{aligned}$$

our assumption enforces that

$$cr_D(G) + cr_D(K_{6,n}, G) < \lfloor \frac{n}{2} \rfloor$$

and therefore, the edges of G are crossed less than $\lfloor \frac{n}{2} \rfloor$ times in D . This implies that, in D , there is at least one subgraph F^i which does not cross G .

Without loss of generality, let $cr_D(G, F^n) = 0$. By Lemma 2.3, $cr_D(G \cup F^n, F^i) \leq 3$ for at least one subgraph F^i , $i \in \{1, 2, \dots, n-1\}$. Moreover, by Lemma 2.2, $cr_D(F^i, F^j) \neq 0$ for all $i, j = 1, 2, \dots, n$, $i \neq j$. Consider now the subdrawing D^* of $G \cup F^n$ induced by D . Our next analysis depends on whether or not the edges the 6-cycle $C_6(G)$ cross each other in D^* .

Assume first, that the edges of the 6-cycle $C_6(G)$ do not cross each other. Since $cr_D(G, F^n) = 0$, all edges of F^n are placed in D^* in one of two regions, say outside, in the view of the subdrawing of the 6-cycle $C_6(G)$ and the unique other edge of G is placed inside this 6-cycle $C_6(G)$ without crossings with the edges of F^n . The unique such drawing D^* is shown in Figure 3(a). It is easy to see that, in D , if some vertex f_i , $i \in \{1, 2, \dots, n-1\}$, is placed inside G in the 3-cycle, then G is crossed by at least three edges joining f_i with the vertices of G . Moreover, as F^i crosses F^n , $cr_D(G \cup F^n, F^i) \geq 4$. In D , if a vertex f_i is placed inside G in the 5-cycle, then G is crossed by at least one edge joining f_i with the vertices of G and $cr_D(F^n, F^i) \neq 0$ too. The reader can easy to verify that if a vertex f_i is placed in a region outside G in D , then F^i crosses the edges $G \cup F^n$ at least four times. But, in this case, some edge of F^i crosses G . If none of the edges of F^i crosses G , then $cr_D(G \cup F^n, F^i) \geq 6$. Let r be the number of vertices f_i , $i \in \{1, 2, \dots, n-1\}$, for which the subgraph F^i crosses G in D . Certainly, $r < \lfloor \frac{n}{2} \rfloor$, otherwise the edges of G are crossed at least $\lfloor \frac{n}{2} \rfloor$ times. The calculating of the necessary crossings in D gives

$$\begin{aligned} cr_D(G + nK_1) &= cr_D(K_{6,n-1}) + cr_D(G \cup F^n) + cr_D(K_{6,n-1}, G \cup F^n) \\ &\geq 6\lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n-2}{2} \rfloor + 2r + 6(n-r-1) \\ &= 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor \\ &\quad + (6n-6-4r-6\lfloor \frac{n-1}{2} \rfloor - \lfloor \frac{n}{2} \rfloor). \end{aligned}$$

This, together with the assumption that D has less than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings, implies that

$$4r > 6n-6-6\lfloor \frac{n-1}{2} \rfloor - \lfloor \frac{n}{2} \rfloor = 5\lfloor \frac{n}{2} \rfloor.$$

This contradict the fact that $r < \lfloor \frac{n}{2} \rfloor$.

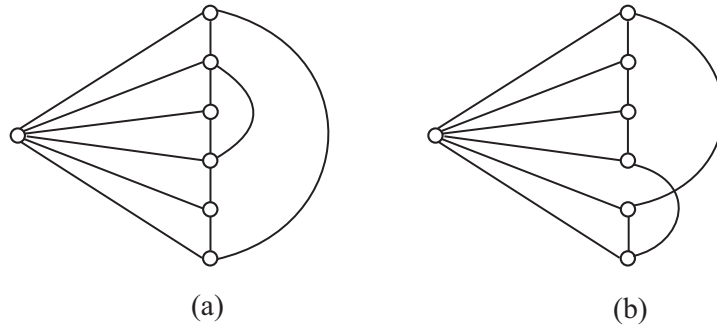


Fig. 3 The unique subdrawings of $G \cup F^n$ and $C_6(G) \cup F^n$

Hence, the edges of the 6-cycle $C_6(G)$ cross each other in D . By Lemma 2.3, in D , there is F^i for which $cr_D(G \cup F^n, F^i) \leq 3$ and, by Lemma 2.2, $cr_D(F^i, F^j) \neq 0$ for all $i \neq j$. Thus, in D , there is F^i for which $cr_D(G, F^i) \leq 2$. This implies that the subdrawing of the 6-cycle $C_6(G)$ divides the plane in such a way that on the boundary of one region, say unbounded, there are all six vertices of G and at least four vertices of G are placed on some other boundary inside the 6-cycle $C_6(G)$. This enforces that the 6-cycle $C_6(G)$ has exactly one internal crossing and the unique subdrawing D^{**} of $C_6(G) \cup F^n$ induced from D is shown in Figure 3(b). One can easy to verify that $cr_D(G \cup F^n, F^i) \geq 3$ for all $i = 1, 2, \dots, n-1$. Moreover, if some F^i crosses in D the edges of $G \cup F^n$ less than six times, then this F^i crosses G . Assume again that r subgraphs F^i cross in D the edges of G . Thus, $r < \lfloor \frac{n}{2} \rfloor$. Now the same reasoning as in the previous paragraph gives $4r > 5\lfloor \frac{n}{2} \rfloor$. This contradiction with $r < \lfloor \frac{n}{2} \rfloor$ completes the proof. $\square$

III. THE JOIN PRODUCT WITH PATHS

The graph $G + P_n$ contains $G + nK_1$ as a subgraph. For the subgraphs of the graph $G + P_n$ which are also subgraphs of the graph $G + nK_1$ we will use the same notations as above. Let P_n^* denote the path induced on n vertices of $G + P_n$ not belonging to the subgraph G . Hence, P_n^* consists of the vertices $f_1, f_2, \dots, f_n$ and of the edges $\{f_i, f_{i+1}\}$ for $i = 1, 2, \dots, n-1$. One can easily see that

$$G + P_n = G \cup K_{6,n} \cup P_n^* = G \cup \left(\bigcup_{i=1}^n F^i \right) \cup P_n^*.$$

As the graph $G + P_1$ is isomorphic with the graph $G + K_1$, the graph $G + P_1$ is planar. For $n \geq 2$ we have the next result.

Theorem 3.1: $cr(G + P_n) = 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ for $n \geq 2$.

Proof. In Figure 2 there is the drawing of the graph $G + nK_1$ with $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ crossings. In this drawing it is possible to add $n-1$ edges which form the path P_n^* on the vertices of nK_1 in such a way that only one edge of P_n^* is crossed by an edge of G . Hence, $cr(G + P_n) \leq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$. To prove the reverse inequality we assume that there is a drawing of the graph $G + P_n$ with fewer than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ crossings and let D be such a drawing. As the graph $G + P_n$ contains $G + nK_1$ as a subgraph, by Theorem 2.1, $cr(G + P_n) \geq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor$ and therefore, no edge of the path P_n^* is crossed in D . Moreover, as $G + P_n$ contains $K_{6,n}$ as a subgraph and $cr(K_{6,n}) = 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor$, in D there are at most $\lfloor \frac{n}{2} \rfloor$ crossings on the edges of G . This implies that in D there are at least $n - \lfloor \frac{n}{2} \rfloor$ subgraphs F^i for which $cr_D(G, F^i) = 0$.

Hence, for $n \geq 2$ at least one subgraph F^i does not cross the edges of G . Assume a subgraph F^i for which $cr_D(G, F^i) = 0$. This vertex f_i is placed in a region with all six vertices of G on its boundary, say unbounded region. Regardless of the number of crossings among the edges of G , the subdrawing D^* of $G \cup F^i$ induced from D divides the plane in such a way that every region outside G contains at most two vertices of G on its boundary. As, in D , none of the edges of P_n^* is crossed, all vertices f_j , $j = 1, 2, \dots, n$, $j \neq i$, are placed outside G . But, in this case, $cr_D(G \cup F^i, F^j) \geq 4$ for all $j \neq i$. Thus, for $n = 2$ there are at least four crossings in D and we are done. For $n \geq 3$, Lemma 2.3 implies that among the edges of the subgraph $G + nK_1$ there are at least $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + 2\lfloor \frac{n}{2} \rfloor$ crossings. This contradiction with the assumption of less than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ crossings in D completes the proof. $\square$

IV. THE JOIN PRODUCT WITH CYCLES

The graph $H + C_n$ contains both $G + nK_1$ and $G + P_n$ as subgraphs. Let C_n^* denote the subgraph of $G + C_n$ induced on the vertices $f_1, f_2, \dots, f_n$. Hence, we have the next result concerning the crossing number of the join of G with the cycle C_n

Theorem 4.1: $cr(G + C_n) = 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 2$ for $n \geq 3$.

Proof. In the drawing in Figure 2 it is possible to add n edges in such a way that these edges, together with the vertices of nK_1 , form the cycle C_n^* and that the edges of C_n^* are crossed only two times. Hence, $cr(G + C_n) \leq 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 2$. To prove the reverse inequality assume that there is a drawing of the graph $G + C_n$ with at most $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ crossings and let D be such a drawing. By Theorem 3.1, none edge of C_n^* is crossed, otherwise deleting the crossed edge from C_n^* results in the drawing of the graph $G + P_n$ with fewer than $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ crossings. Moreover, as $G + C_n = G \cup K_{6,n} \cup C_n^*$ and $cr(K_{6,n}) = 6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor$, in D , there are at most $\lfloor \frac{n}{2} \rfloor + 1$ crossings on the edges of G .

So, in D there are at least $n - \lfloor \frac{n}{2} \rfloor - 1$ subgraph F^i for which $cr_D(G, F^i) = 0$. Since $n \geq 3$, in D , there is at least one subgraph F^i which does not cross the edges of G . Now, the same consideration as in the previous section gives that, in D , every subgraph F^j , $j \neq i$, crosses the edges of $G \cup F^i$ at least four times. Thus, for $n = 3$, in D , there are at least eight crossings between the edges of $G \cup F^i$ and the edges of $F^j \cup F^k$, $j, k \in \{1, 2, 3\}$, $j, k \neq i$. If some other crossing appears in D , we are done. Otherwise the edges of G do not cross each other and the subdrawing of $G \cup F^i$ divides the plane as shown in Figure 3(a). Moreover, none crossing appears on the edges of C_3^* and therefore, both vertices f_j and f_k , $j, k \in \{1, 2, 3\}$, $j, k \neq i$, are placed in D in the same region outside G . Since $cr_D(G, F^j \cup F^k) \leq 2$, one can easily verify that $cr_D(F^j, F^k) \neq 0$. This additional crossing confirms that, in D , there are at least nine crossings. So, Theorem 4.1 is true for $n = 3$. Let us consider now $n \geq 4$. Lemma 2.3 implies that in D there are at least $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + 2\lfloor \frac{n}{2} \rfloor$ crossings among the edges of $G + nK_1$. This contradiction with the assumption of at most $6\lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor + \lfloor \frac{n}{2} \rfloor + 1$ crossings completes the proof. $\square$

ACKNOWLEDGMENT

This work was supported by the Slovak VEGA grant No. 1/0309/11 (50%). This work is also the result of the project implementation Development of Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

REFERENCES

- [1] K. Asano, "The crossing number of $K_{1,3,n}$ and $K_{2,3,n}$," *J. Graph Theory* 10 (1986), 1-8.
- [2] L. W. Beineke and R. D. Ringeisen, "On the crossing numbers of products of cycles and graphs of order four," *J. Graph Theory* 4 (1980), 145-155.
- [3] D. Bokal, "On the crossing numbers of Cartesian products with paths," *J. Combin. Theory B* 97 (2007), 381-384.
- [4] E. Draženská and M. Klešč, "The crossing numbers of products of cycles with 6-vertex trees," *Tatra Mt. Math. Publ.* 36 (2007), 109-119.
- [5] M. R. Garey and D. S. Johnson, "Crossing number is NP-complete," *SIAM J. Algebraic. Discrete Methods* 4 (1983) 312-316.
- [6] L. Y. Glebsky and G. Salazar, "The crossing number of $C_m \times C_n$ is as conjectured for $n \geq m(m+1)$," *J. Graph Theory* 47 (2004), 53-72.
- [7] F. Harary, P. C. Kainen and A. J. Schwenk, "Toroidal graphs with arbitrarily high crossing numbers," *Nanta Math.* 6 (1973), 5867.
- [8] P.T. Ho, "The crossing number of $K_{1,m,n}$," *Discrete Math.* 308 (2008), 5996-6002.
- [9] Y. Huang and T. Zhao, "The crossing number of $K_{1,4,n}$," *Discrete Math.* 308 (2008), 1634-1638.
- [10] S. Jendroľ and M. Ščerbová, "On the crossing numbers of $S_m \times P_n$ and $S_m \times C_n$," *Časopis pro pěstování matematiky* 107 (1982), 225-230.
- [11] D. J. Kleitman, "The crossing number of $K_{5,n}$," *J. Combin. Theory* 9 (1970), 315-323.
- [12] M. Klešč, "The crossing numbers of products of paths and stars with 4-vertex graphs," *J. Graph Theory* 18 (1994), 605-614.
- [13] M. Klešč, "The crossing numbers of Cartesian products of paths with 5-vertex graphs," *Discrete Mathematics* 233 (2001), 353-359.
- [14] M. Klešč, "The join of graphs and crossing numbers," *Electronic Notes in Discrete Mathematics* 28 (2007), 349-355.
- [15] M. Klešč, "The crossing numbers of join of the special graph on six vertices with path and cycle," *Discrete Math.* 310 (2010), 1475-1481.
- [16] M. Klešč and Š. Schrötter, The crossing numbers of join products of paths with graphs of order four, *Discussiones Mathematicae - Graph Theory*, (to be published).
- [17] V. R. Kulli and M. H. Muddebihal, "Characterization of join graphs with crossing number zero," *Far East J. Appl. Math.* 5 (2001), 87-97.
- [18] H. Mei and Y. Huang, "The crossing number of $K_{1,5,n}$," *International J. Math. Combin.* 1 (2007), 33-44.
- [19] P. Turán, "A note of welcome," *J. Graph Theory* 1 (1977), 7-9.
- [20] K. Zarankiewicz, "On a problem of P. Turán concerning graphs," *Fund. Math.* 41 (1954), 137-145.

Kombinovaná segmentácia — inicializácia aktívnej kontúry pomocou Graph cut metódy

Peter KARCH, Marián BAKOŠ, Iveta ZOLOTOVÁ

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

peter.karch@tuke.sk, marian.bakos@tuke.sk, iveta.zolotova@tuke.sk

Abstrakt — Článok popisuje metódu implementácie kombinovanej metódy segmentácie pri zjednodušenej inicializácii Aktívnej kontúry pomocou Graph cut metódy. To umožňuje dosiahnuť rýchlu evolúciu kontúry pri metóde Aktívnej kontúry pri minimálnej interakcii používateľa v rámci tejto navrhutej kombinovanej metódy.

Kľúčové slová — segmentácia, Graph cut, Aktívna kontúra

I. ÚVOD

Pokročilé metódy segmentácie [1], [2], [3] a [5] sú po ich precíznej inicializácii v podobe inicializačnej kontúry, inicializačných terminálov a dobrom nastavení ich parametrov dostatočne presné, ale časovo náročné. Pokiaľ je potrebné výsledok dosiahnuť v relatívne krátkom čase, bez presných nastavovaní parametrov a bez precíznej inicializácie, výsledky nie sú až také presvedčivé. V nami použitej metóde segmentácie sa skombinujú dve metódy segmentácie, a to *Graph cut* a *Aktívna kontúra*. Pre zrýchlenie výsledku segmentácie pomocou metódy *Aktívnej kontúry* sa pre jej inicializáciu použije výsledná segmentácia z metódy *Graph cut*. Hlavnou myšlienkou vzniku tejto kombinovanej metódy bolo zrýchlenie pomalej evolúcie krivky pri metóde *Aktívnej kontúry* dosiahnutím malého počtu krokov evolúcie krivky. Pri zjednodušenej inicializácii *Aktívnej kontúry* je počet jej evolúcií pre dosiahnutie dobrého výsledku vysoký, čo je časovo náročné. Pre zrýchlenie evolúcie by ju používateľ musel časovo náročnejšie inicializovať pomocou presnejšej kontúry segmentovaného „objektu“. Túto presnejšiu kontúru je možné dosiahnuť aj pomocou *Graph cut* metódy pri jednoduchšej časovo nenáročnej inicializácii.

II. POUŽITÉ METÓDY SEGMENTÁCIE

A. *Graph cut* segmentácia

Táto metóda patrí medzi metódy založené na grafoch, kde hlavnou myšlienkou je vytvorenie grafu $G = (V, E)$, ktorý predstavuje samotný segmentovaný vstupný obraz I . Tento graf pozostáva z vrcholov

$$V = P \cup \{S, T\}, \quad (1)$$

kde P je množina všetkých pixelov obrazu I a S, T sú množiny pixelov predstavujúce dodatočné terminály identifikujúce pixely patriace objektu S a pixely patriace pozadiu T . Ďalej tento graf pozostáva z hrán

$$E = N \bigcup_{p \in P} \{\{S, p\}, \{p, T\}\}, \quad (2)$$

kde N je množina susednosti medzi jednotlivými pixelmi v obraze a tieto hrany sú označované ako *n-linky*. Hrany reprezentujúce vzťah medzi pixelmi v obraze a terminálmi sa nazývajú *t-linky*.

Rez v grafe, ktorý vo výsledku oddeľuje objekt od jeho pozadia a nachádza sa na hranici medzi objektom a pozadím v mieste, kde je maximálny tok v grafe. Tento tok je určený pomocou algoritmu pre určenie maximálneho toku v grafe [1].

Segmentácia je určená pomocou značkovacieho vektora $A = \{A_1, \dots, A_p, \dots, A_{|P|}\}$. Segmentácia objektu od jeho pozadia je definovaná ako problém binárneho značenia, kde každý pixel je označený ako $A_k = \{O, B\}$, pričom O predstavuje pixel patriaci „objektu“ a B predstavuje pixel patriaci „pozadiu“. Tieto dve úrovne sú identifikované na základe terminálových uzlov v obraze a určujú pixely, ktoré určite patria „objektu“ alebo „pozadiu“ a predstavujú silné obmedzenie segmentácie. Ostatné pixely v obraze predstavujú slabé obmedzenia segmentácie a sú určené na základe hraničných a regionálnych vlastností A popísaných hodnotovou funkciou:

$$E(A) = \lambda \cdot R(A) + B(A), \quad (3)$$

kde $R(A)$ predstavuje regionálne pravidlo [5], ktoré určuje jednotlivé penalizácie za priradenie pixela p k „objektu“ alebo „pozadiu“. Pravidlo hraníc je určené pomocou $B(A)$ [5] a môže byť interpretované ako penalizácia za nesúvislosť medzi dvoma susednými pixelmi p a q . Koeficient λ určuje dôležitosť regionálneho pravidla voči hraničnému.

B. Aktívna kontúra založená na regiónoch

Základnou myšlienkou modelu *Aktívnej kontúry* je evolúcia krivky a pomocou jej obmedzení detekovať objekty v obraze. Je závislá na používateľskej interakcii, ktorá pozostáva zo stanovenia krivky okolo detekovaného objektu. Táto metóda *Aktívnej kontúry* je implementáciou modelu aktívnej kontúry popísanej [3] založenej na *Mumford-Shahovej* segmentačnej technike a metóde *Level set*, ktorá pre zastavenie vývoja krivky na požadovanej hranici segmentovaného „objektu“ nevyužíva hranovú funkciu. Je vhodná pre segmentáciu obrazov, kde „objekt“ a „pozadie“ sú rozdielne a homogénne. Pozostáva z nasledujúcich krokov algoritmu: inicializácia ϕ^0 pomocou ϕ_0 pre $n = 0$, určenie $c_1(\phi^n)$ [3] a $c_1(\phi^n)$ [3], určenie $\partial\phi/\partial t$ [3] pre získanie ϕ^{n+1} , reinitializácia ϕ [4] následne sa tieto kroky opakujú $n = n + 1$, pokiaľ riešenie nie je statické, alebo nedosiahne predpísaný počet iterácií.

C. Navrhnutá kombinovaná metóda segmentácie

Pre zjednodušenie inicializácie sa prvá metóda v rámci kombinovanej metódy použije *Graph cut*. Týmto spôsobom je potrebné určiť len niekoľko pixelov v obraze patriacich „objektu“ a niekoľko pixelov patriacich „pozadiu“, namiesto určenia presnejšej inicializačnej kontúry objektu, ak by sme chceli dosiahnuť čo najmenší počet krokov evolúcie krivky *Aktívnej kontúry*. Zjednodušenie *Graph cut* metódy dosiahneme zmenšením počtu určovaných hrán vo vytvorenom grafe pre túto metódu tým, že z *t-líní* určíme len tie, ktoré spájajú terminály „objektu“ a „pozadia“ s ich inicializačnými pixelmi. Hrany *n-líní*, v grafe, určíme len s ohľadom na 4-susedstvo. Pri takto zjednodušenej *Graph cut* metóde dosiahneme len približnú výslednú segmentáciu vybraného objektu v obraze, ale za krátky čas. Presnosť dosiahneme použitím druhej metódy, metódy *Aktívnej kontúry*. Táto metóda využije pre svoju inicializáciu približný výsledok z prvej metódy a pre dosiahnutie presnejšej výslednej segmentácie požadovaného „objektu“ tak bude potrebovať menší počet krokov evolúcie krivky.

III. EXPERIMENTY

Experiment bol vykonaný na vstupnom obraze Obr. 1 a porovnaný po časovej a kvalitatívnej stránke segmentácie. Táto navrhnutá metóda kombinovania dvoch metód bola následne porovnaná s výsledkom dosiahnutým jednotlivými typmi segmentácií.

Metóda *Graph cut* bola inicializovaná pomocou určenia niekoľkých pixelov identifikujúcich pixely patriace „objektu“ a „pozadiu“ Obr. 2a. Konštanta je nastavená na hodnotu $\lambda = 0$.

Hodnota každej *t-linky* spájajúcej každý neidentifikovaný pixel v obraze s terminálom „objektu“ a „pozadia“ je určená hodnotou $R_p = 0$.

Hodnota každej *n-linky* spájajúcej každé dva susedné pixely v obraze je určená hodnotou $B_{\{p,q\}}$. Táto hodnota je určená $B_{\{p,q\}} = \exp\left(-\left(I_p - I_q\right)^2 / 2\sigma^2\right)$. Metóda *Aktívnej kontúry* je inicializovaná pomocou Obr. 3a výslednej segmentácie dosiahnutej použitím *Graph cut* segmentácie Obr. 2b a počet jej iterácií, potrebných na dokončenie segmentácie Obr. 3b, je nastavený na hodnotu 100.



Obr. 1 Vstupný obraz



Obr. 2 a) Vľavo: Inicializácia pre Graph cut, b) vpravo: Samotná Graph cut segmentácia pre $\sigma=0,5$ $t=0,32885$

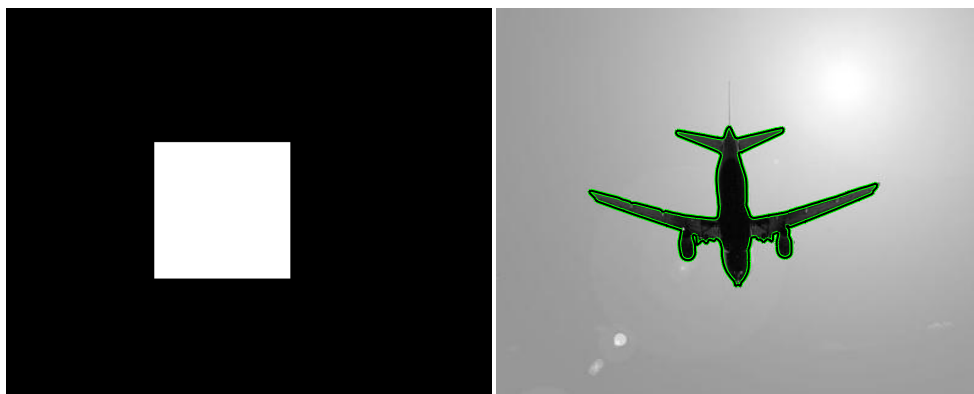


Obr. 3 a) Vľavo: Inicializačná kontúra získaná pomocou metódy Graph cut pre Aktívnu kontúru, b) Vpravo: Výsledná kombinovaná segmentácia po 100 iteráciách za čas $t=5.2751s$



Obr. 4 Porovnanie čiastkových segmentácií v detaile

Pre porovnanie na to istom obraze Obr. 1 bola vykonaná aj segmentácia pomocou samostatnej *Aktívnej kontúry*. Tá bola inicializovaná čo najjednoduchším spôsobom, a to maskou v tvare štvorca Obr. 5a. Evolúcia krivky aktívnej kontúry Obr. 5b si pre dosiahnutie dobrého výsledku vyžadovala 500 iterácií pri čase $t = 21.544s$, zatiaľ čo Kombinovaná metóda, potrebovala na dosiahnutie takého istého výsledku len $t = 5.2751s$ a počet potrebných iterácií *Aktívnej kontúry* bol 100. Tento čas by mohol byť lepší, ale inicializácia by bola oproti *Graph cut* metóde časovo náročnejšia.



Obr. 5 Aktívna kontúra. a) Vľavo: Inicializácia Aktívnej kontúry,
b) Vpravo: Samotná evolúcia krivky po 500 iteráciách za čas $t=21.544s$

IV. ZÁVER

Navrhnutá kombinovaná metóda umožňuje využiť výhody *Graph cut* metódy, ako aj metódy *Aktívnej kontúry*, ktoré využíva pri dosiahnutí dobrého výsledku segmentácie vstupného obrazu za čo najkratší možný čas. Táto kombinovaná metóda využíva možnosť jednoduchšej inicializácie *Graph cut* metódy Obr. 2a len pomocou približného určenia inicializačných terminálov. Takto dosiahnutá segmentácia nie je síce presná, ale je rýchla. Tento výsledok Obr. 3a je následne použitý, ako automatická inicializácia *Aktívnej kontúry*. Tým odbremenuje používateľa použiť zdĺhavejšiu inicializáciu pomocou vytvorenia inicializačnej kontúry v prípade metódy *Aktívnej kontúry*. Naproti tomu metóda *Aktívnej kontúry* presnejšie dokončí segmentáciu Obr. 4 dosiahnutú pomocou, čo najviac zjednodušenej metódy *Graph cut*.

POĎAKOVANIE

Táto publikácia bola vytvorená realizáciou projektu Rozvoj centra informačných a komunikačných technológií pre znalostné systémy (číslo projektu: 26220120030) na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja (100%).

LITERATÚRA

- [1] Y. Boykov, V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization In Vision", *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no.9, pp. 1124-1137, September 2004.
- [2] S. Lankton, A. Tannenbaum, "Localizing Region-Based Active Contours", *IEEE Transactions on Image Processing*, vol.17, no.11, pp.2029-2039, November 2008.
- [3] T. F. Chan, L. A. Vese, "Active Contours Without Edges", *IEEE Transactions on Image Processing*, Vol. 10, No. 2, pp 266-277, February 2001.
- [4] M. Sussman, P. Smereka, S. Osher, "A level set approach for computing solutions to incompressible two-phase flow", *J. Comput. Phys.*, vol. 119, pp. 146-159, 1994.
- [5] P. Karch, I. Zolotová, "Interactive Contour Deformation of an Object Using Graph Cut", *INES 2010: 14th International Conference on Intelligent Engineering Systems*, pp 227-230, May 2010.

Vzdialené monitorovanie a supervízne riadenie na báze aplikačného a webového servera

Iveta ZOLOTOVÁ, Miloš PAVLÍK, Roman MIHAL, Miroslav KOLLÁR

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

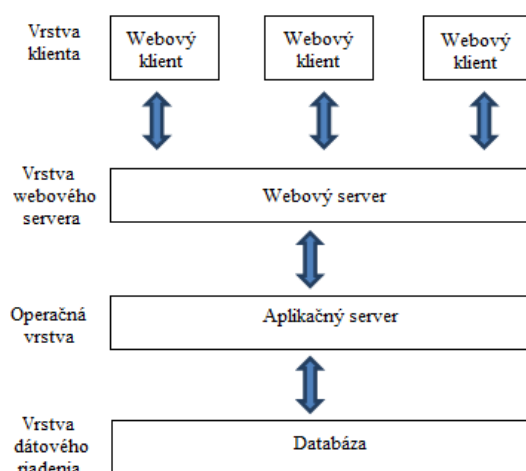
iveta.zolotova@tuke.sk, milos.pavlik@tuke.sk, roman.mihal@tuke.sk,
miroslav.kollar@student.tuke.sk,

Abstrakt — Článok sa zameriava na vzdialené monitorovanie a supervízne riadenie reálnych a virtuálnych riadených systémov na báze viacvrstvovej softvérovej klient – server architektúry. Sústreďuje sa na aplikačný a webový server a ich klientov. Demonštruje ju na reálnej sieťovej aplikácii s použitím profesionálnych produktov spoločnosti Rockwell Automation, Inc. v rámci laboratórneho distribuovaného systému riadenia. Monitorované a supervízne riadené systémy môžu byť reálne alebo virtuálne.

Kľúčové slová — softvérová architektúra, hrubý klient, tenký klient, vzdialené monitorovanie a supervízne riadenie Rockwell Automation, Inc.

I. ÚVOD

Viacvrstvomá architektúra klient-server je založená na myšlienke oddeliť subjekty, ktoré poskytujú služby (servery) od subjektov (klientov), ktoré využívajú ich služby. Logicky súvisiace časti (napr. prezentačná, webová, operačná a dátová časť) sú od seba oddelené do vrstiev. V rámci tejto architektúry sa využíva synchrónna komunikácia, čiže klient a server si vymieňajú údaje vo forme otázok a odpovedí [1],[2].



Obr. 1 4-vrstvomá klient-server architektúra

Údaje z technológie a z procesnej úrovne riadenia slúžia aj pre monitorovanie a supervízne riadenie. Supervízne riadenie je riadenie, pri ktorom sa používa vizualizácia procesov pre grafické znázornenie technologických činností pomocou HMI (Human Machine Interface)

rozhrania. Pri vzdialenom supervíznom riadení sa využíva hrubý klient aplikačného servera alebo tenký klient webového servera, ktorým môže byť ľubovoľný webový prehliadač [3].

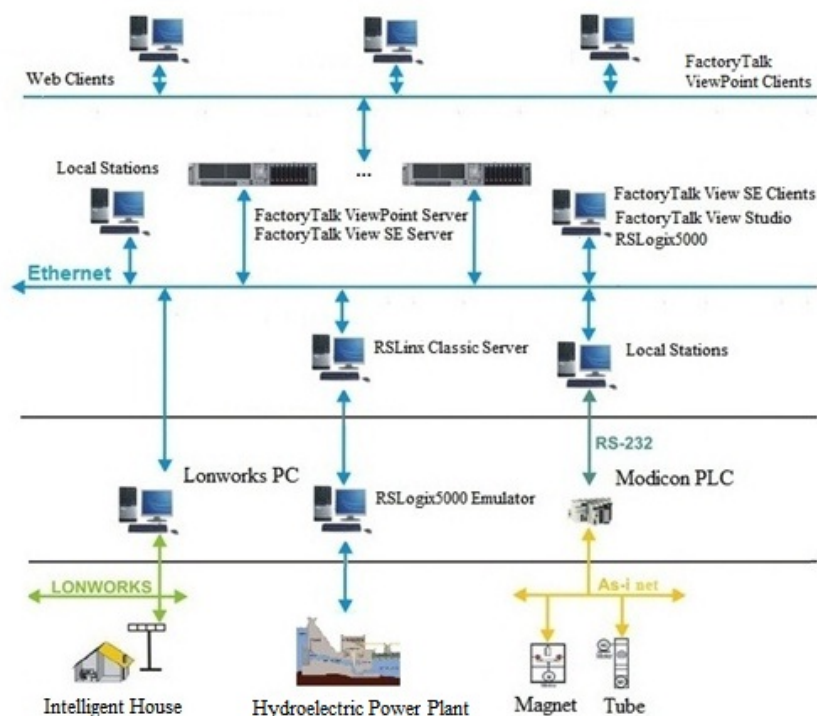
Článok popisuje problematiku analýzy, návrhu a realizácie viacvrstvovej klient-server architektúry v aplikácii pre vzdialené monitorovanie a supervízne riadenie v rámci pedagogického a výskumného laboratórneho distribuovaného systému riadenia [4].

II. ANALÝZA, NÁVRH A REALIZÁCIA APLIKÁCIE VZDIALENÉHO MONITOROVANIA A SUPERVÍZNEHO RIADENIA

Hlavným cieľom prezentovanej aplikácie je modifikovať a doplniť laboratórny distribuovaný systém riadenia o ďalší virtuálny model technológie a jeho monitorovanie a riadenie tak, aby bolo možné skúmať a vyučovať základné vlastnosti použitia aplikačných a webových serverov vo vzdialenom monitorovaní a supervíznom riadení a to s využitím profesionálnych nástrojov spoločnosti *Rockwell Automation, Inc.*

Riadeným systémom - technológiou je nový virtuálny model vodnej elektrárne, ktorého správanie a procesné riadenie je zadefinované v emulovanom PLC (*RSLogix5000 Emulator*). Zber údajov z technológie pre toto PLC je zabezpečené pomocou OPC servera (*RSLink Classic*) [5].

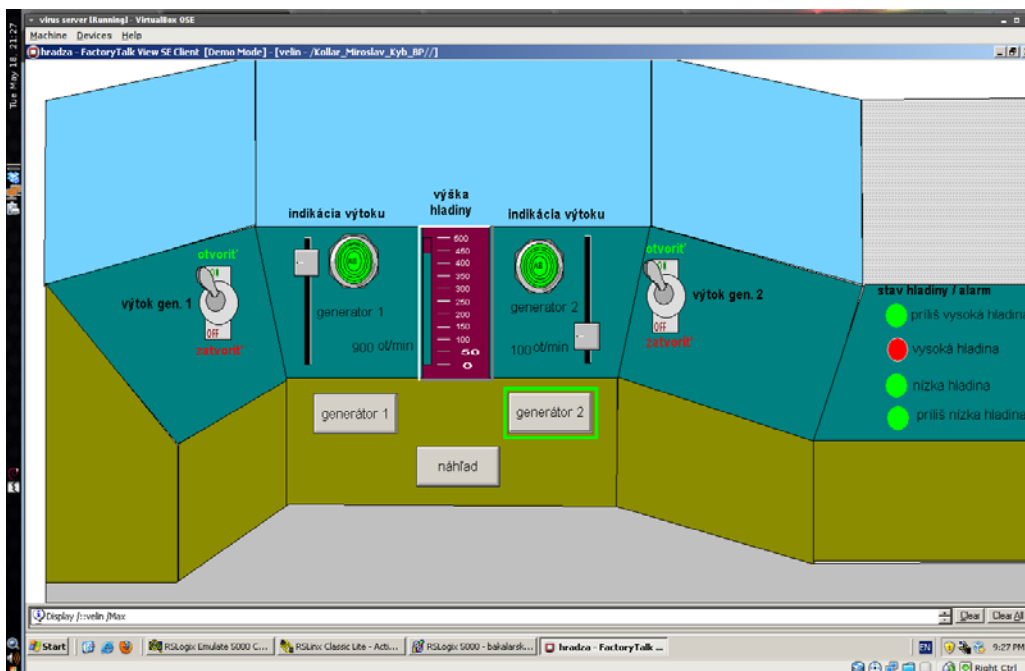
Aplikácia vzdialeného monitorovania a riadenia sa vyvíja vo vývojovom prostredí *FactoryTalk View Studio* a realizuje sa na báze softvéru *FactoryTalk View Site Edition* a jeho doplnku *FactoryTalk ViewPoint*. HMI aplikačný server (*FactoryTalk View SE Server*) obsahuje aplikáciu s grafickými pohľadmi a ovládacími panelmi a zabezpečuje aj zber údajov z technológie. Pre užívateľov v real-time režime je potrebné použiť hrubého klienta na báze *FactoryTalk View SE Client*. Pre vytvorenie aplikácie vzdialeného webového monitorovania a supervízneho riadenia sa používa webová vrstva na báze webového servera. Poskytuje real-time prístup k aplikácii prostredníctvom tenkého klienta, ktorým môže byť ľubovoľný internetový prehliadač nainštalovaný na ľubovoľnom operačnom systéme. V rámci tejto verzie aplikácie sa údaje neukladajú do databázy, t.j. nie je použitá databázová vrstva [6].



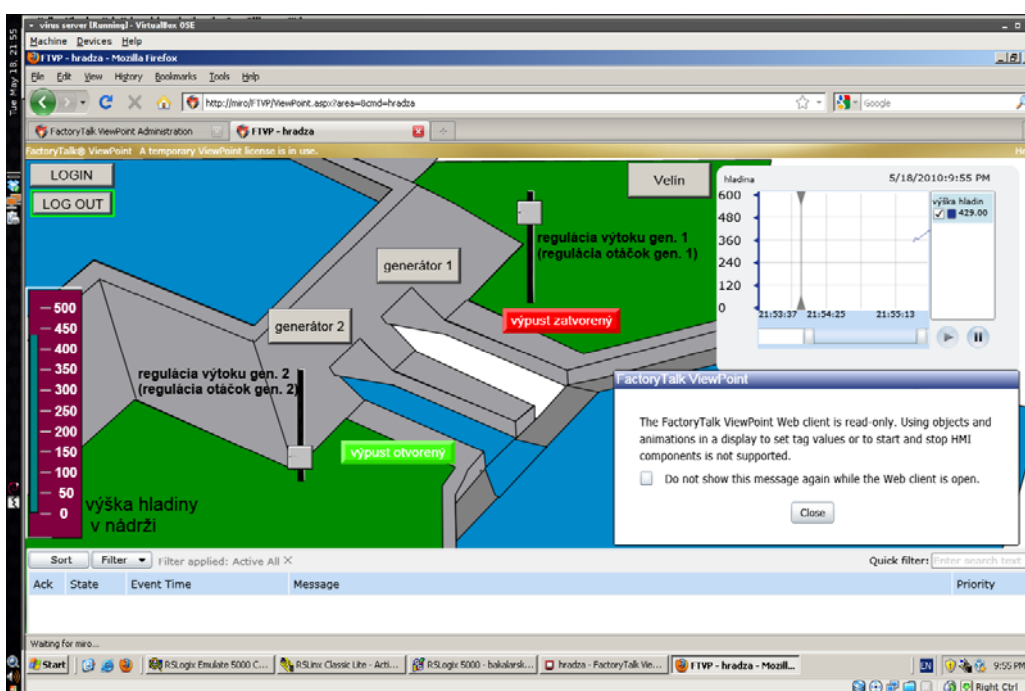
Obr. 2 Viacvrstvová architektúra pre vzdialené monitorovanie a riadenie

K aplikácii sa pristupuje buď pomocou hrubého klienta (*FactoryTalk View SE Client*) alebo pomocou tenkého klienta, čo môže byť internetový prehliadač. Pre prístup cez *FactoryTalk View SE Client* je nutná inštalácia *FactoryTalk Service Platform* a *FactoryTalk Directory*.

Pre prístup do aplikácie pomocou doplnku *FactoryTalk ViewPoint* je potrebné najprv aplikáciu publikovať v administrácii *FactoryTalk ViewPoint*, nastaviť prístupové práva a následne je možné aplikáciu zobraziť pomocou internetového prehliadača zadaním adresy https://nazov_pocitaca/FTVP. Na obr. 4 je úvodná obrazovka zobrazená pomocou tenkého klienta webového servera [7].



Obr. 3 Obrazovka Velín vo FactoryTalk View SE Client



Obr. 4 Úvodná obrazovka zobrazená pomocou tenkého klienta webového servera

III. ZÁVER

Navrhnutý a realizovaný distribuovaný systém s viacvrstvovou (v prípade hrubého klienta dvojvrstvovou a v prípade tenkého klienta trojvrstvovou) klient-server architektúrou je príkladom ako sofistikovane využiť technológie pre vzdialené monitorovanie a supervízne riadenie systémov, ktoré je možné riadiť vzdialene pomocou hrubého alebo tenkého klienta. Oddelenie jednotlivých súčastí systému, v rámci viacvrstvej klient-server architektúry a riadenia systému, do vrstiev zvyšuje jeho modifikovateľnosť a výkonnosť, keďže je tu možnosť paralelných spracovaní podporovaných servermi a dobrá rýchlosť komunikácie medzi vrstvami. Ďalšie rozšírenia napr. o databázovú vrstvu systému sú možné napríklad použitím nástroja *FactoryTalk Historian SE*.

POĎAKOVANIE

Táto publikácia je podporená grantom KEGA No. 037-011TUKE-4/2010 (50%) a projektom Rozvoj Centra informačných a komunikačných technológií pre znalostné systémy (projekt číslo: 26220120030) podporované Výskumom a rozvojom Operačného programu financovaného z ERDF (50%).

LITERATÚRA

- [1] M. Franeková, P. Peniak, "Ukážky použitia Matlab OPC Toolboxu s technológiou Rockwell Automation" in *Konferencia Technical Computing Prague 2009*. Humusoft. Praha: november 2009. ISBN 978-80-733-0. p. 28-29.
- [2] D. Garlan, M. Shaw, *An Introduction to Software Architecture, Advances in Software Engineering and Knowledge Engineering*, Vol. 1, ad. Ambriola and Tortora, World Scientific Publishing Co., 1993.
- [3] J. Liguš, I. Zolotová, P. Karch, "Information and control system of traverse and its integration into cybernetic centre", in *Electronics and electrical engineering*. No. 6 (2010), p. 147-152. - ISSN 1392-1215.
- [4] M. Kolenčík, K. Žáková, "A Contribution to Remote Control of Inverted Pendulum", in *17th Mediterranean Conference on Control and Automation* : Thessaloniki, Greece, 24.-26.6.2009. - Piscataway : IEEE, 2009. - ISBN 978-1-4244-4685-8. - S. 1433-1438.
- [5] D. Mudrončík, I. Zolotová, *Priemyselné programovateľné regulátory*. Elfas.r.o., Košice, 2000.
- [6] <http://www.rockwellautomation.com/rockwellsoftware/factorytalk/services.html>
- [7] M. Kollár, "Softvérové architektúry supervízneho vzdialeného riadenia", Bakalárska práca, KKUI FEI TU, 2010.
- [8] D. Garlan, M. Shaw, C. Okasaki, C. Scott, R. Swonger, "Experience with a course on architectures for software systems," in *Proceedings of the Sixth SEI Conference on Software Engineering Education*, Springer-Verlag, LNCS 376, October 1992.
- [9] L. Landryová, P. Osadník, M. Valas, O. Winkler, "Integrating tools and methodologies in FutureSME delivery platform", *Seminar ASR '2010 "Instruments and Control"* © 2010, VŠB-TUO, Ostrava, ISBN 978-80-248-2191-7.
- [10] M. Pavlík, "Riadenie modelov Lanovka, Zvony a Trubica a ich zaradenie do DSR na KKUI", Diplomová práca, KKUI FEI TU, 2009.
- [11] D.E. Perry, A.L. Wolf, *Foundations for the Study of Software Architecture*, Software Engineering Notes, Vol. 17, No. 4, pp. 40-52, October 1992.
- [12] D. Popescu, Q. Meng, A.A. Ionete, C. Ionete, R. Aguridan "Remote vs. Simulated, Virtual or Real-Time Automation Laboratory" in *International Conference on Automation and Logistics*, Shenyang, China, August 2009, ISBN 978-1-4244-4794-7, s. 1410 - 1415.
- [13] O. Winkler, M. Valas, P. Osadník, L. Landryová, "Development of Communication Model for Data Transfer Between MES and ERP Systems for SMEs", Published by: Poliscrypt - Politecnico di Milano, Piazza Leonardo da Vinci, 32 - 20133 Milano, ISBN 9788864930077, Editors: Marco Garetti, Marco Taisch, Sergio Cavalieri, Sergio.
- [14] P. Božek, *Complex Control and Metrology Security of Automatized Trial System*, Manufacturing Engineering, number 2, volume IV, 2005, p. 31-35, Prešov, ISSN 1335-7972.

Title:

Electrical Engineering and Informatics II

Proceeding of the Faculty of Electrical Engineering and Informatics
of the Technical University of Košice

Editorial Board:
Chairman:
Dušan Kocur

Members:
Ján Kollár
Dušan Levický
Alena Pietriková
Ján Sarnovský
Liberios Vokorokos

Technical Editorial Board:
Katarína Kubišová
Ján Papaj

Published by: **Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9,
042 00 Košice, Slovak Republic**

Year of publication: **March, 2011**

Design: **FEI TU of Košice**

Pages: **543**

Edition: **First edition**

Printing: **50 pcs CD**

Language: **English, Slovak**

Copyright © 2011 FEI TUKE

ISBN 978-80-553-0611-7