Acta Universitatis Sapientiae

Electrical and Mechanical Engineering

Volume 6, 2014

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 5–19

Development and Embedded Implementations of a Hardware-Efficient Optical Flow Detection Method

László BAKÓ,[1] Fearghal MORGAN,[2] Szabolcs HAJDÚ,[3]
Sándor-Tihamér BRASSAI,[3] Róbert MONI,[3] Călin ENĂCHESCU[4]

[1]Research Department, Petru Maior University of Tîrgu Mureş, Romania,
e-mail: lbako@ms.sapientia.ro
[2]Bio-Inspired Electronics and Reconfigurable Computing (BIRC) Research Group,
National University Ireland Galway, Ireland,
e-mail: fearghal.morgan@nuigalway.ie
[3]Electrical Engineering Department, Faculty of Technical and Human Sciences,
Sapientia Hungarian University of Transylvania, Romania,
e-mail: tiha@ms.sapientia.ro
[4] Department of Informatics, Faculty of Sciences and Letters,
Petru Maior University of Tîrgu Mureş, Romania,
e-mail: ecalin@upm.ro

Manuscript received September 29, 2015; revised November 25, 2015.

Abstract: The main goal of the proposed project is to enhance the capabilities of a wheeled or flying mobile robot with features like egomotion estimation and/or obstacle avoidance. This implies the implementation of vision-based navigation of robots using artificial vision, computed with on-board embedded hardware. The current paper aims to contribute on the implementation of a real-time motion extraction from a video feed using embedded FPGA circuits. An alternative implementation using a Raspberry Pi is also presented. A performance analysis is given with references to other works.

Keywords: motion extraction, optical flow, embedded implementation, real-time, FPGA circuit, VHDL, low-resource-cost, Raspberry Pi.

1. Scientific background

The optical flow calculation involves extracting a dense velocity field of an image sequence assuming that the intensity is preserved during the motion. This result may then be used for other applications, such as three-dimensional (3-D) reconstruction, time interpolation of image sequences, video compression, motion segmentation, tracking, robot navigation, and time to collision estimation. There are several ways to recover 3-D information from two-

The results of this study were partially presented at the 5th International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics 2015.

dimensional images (2-D) using various signals. In this article we will describe implementation of a motion flow system in real time, with low resource cost properties. Optical flow algorithms are widely covered in the specific literature. Some authors have undertaken a comparative study of the accuracy of different approaches with synthetic sequences [1]. We have focused on a model of classical gradient based method of Lucas & Kanade (L & K) [2]. Several authors have emphasized satisfactory balance between precision and efficiency in this model, which is an important factor in deciding which model is best suited for use as a real-time processing system.

One of the most important choices at the design level of a vision system is the selection of the image acquisition hardware. For instance, there are alternatives to the cameras similar to vertebrate-like single-lens eyes, such as insect-like compound eyes [3, 4] that are developed by prestigious research groups. These offer a dynamically adaptable structure with panoramic field of view, low distortion and aberration, and good temporal resolution while yielding high spatial resolution alongside a reduced size. These properties are highly useful for visually-controlled navigation, specifically for tasks like take-off, landing, collision avoidance and other optically driven responses, which do not require a high resolution image acquisition. The local sensory adaptation capabilities of insect compound eyes can compensate for significant changes in light intensity at the photoreceptor level and distribute information in a neuronal circuitry, resulting in fast and low-power integrated signal processing.

The processing hardware support [5, 6, 7, 8, 9, 10] selection for implementing these artificial vision systems can be critical for a high value outcome [10, 11, 12, 13]. Being at the center of group's research activity, the new generations of SRAM-based FPGA devices are a proper choice for the implementation of reconfigurable computing platforms that need accelerated processing in real-time systems. On the other hand, the hardware-software co-design problem is more complex in system development because the components need to be more advanced. The requirements for runtime partial reconfiguration capability in embedded applications can be sustained by storing multiple bit-stream generation choices, including direct bit-stream manipulation for logic blocks and hybrid one-dimensional and two-dimensional physical area relocation control modules.

The main goal of the proposed project is to enhance a mobile robot with evolutionary optimization capabilities for tasks like ego motion estimation and/or obstacle avoidance. The robot will learn to navigate different environments and will adapt to changing conditions. Using the run-time reconfiguration properties of modern digital reconfigurable hardware-based (FPGA) platforms [12, 13, 14, 15], an otherwise days-long evolutionary cycle of a physical robot can be slashed to a matter of milliseconds. By implementing this technique, the most common issues that emerge when using evolutionary

simulation – modeling the real world environment [15, 16, 17] as accurately as possible and modeling only those characteristics of the robot that are relevant for achieving the desired behavior – are avoided.

2. The developed method for resource-efficient optical flow extraction

The first studies on optical flow computation date back to 1980 and there are many alternative methods offered. They can be based on gradient, correlation, energy and phase methods, creating well-defined groups [4]. Gradient methods are based on the evaluation of spatial-temporal derivatives. The first such methods are presented by Horn and Schunck [1], respectively Lucas and Kanade [2]. All these methods are difficult to implement in digital hardware, due to their high resource-cost.

If we represent the image with a matrix A, its values will represent the gray level of a point in the image. When representing a grayscale pixel on 8 bits, these values will vary between 0 and 255. In *Fig. 1* we can see a frame of test video sequence named GRID. The images corresponding to this and other sequences were used as inputs to the algorithm for calculating the Optical Flow (OF), developed using Visual C++.

The gradient in an image is a vector indicating the direction of variation of image intensity (grayscale variation direction). This can be determined by calculating the value difference of adjacent image points. Consider a new matrix B which contains the gradient values of the matrix A. Using the values adjacent to the pixel *p* in the image in the calculation of the gradient, will result in a properly aligned gradient. Detection of outliers in this gradient will then lead to the detection of edges in images. This method, however, is sensitive to noise and luminance variations. The effect of noise can be reduced by calculating the average values of the gradient in the orthogonal direction, too. A horizontal gradient used so far is made by calculating the difference between values of two columns.

$$B(j, k)=A(j, k+1) - A(j, k-1) \tag{1}$$

This can be represented as a filter matrix of the form

–1	0	1

where the values multiplied with the pixel gray-level values will determine the locations of sharp tone differences in the image.

In order to reduce noise sensitivity of the method we have studied the possibilities of determining the average value of the gradient calculated from the video images.

Vertical edges are obtained by averaging the three rows in this matrix:

–1	0	1
–1	0	1
–1	0	1

Similarly, averaged horizontal edge values may be obtained using a vertical mask of the form:

1	1	1
0	0	0
–1	–1	–1

The result of these operations will be placed at the location indexed by the central element of these matrices. In fact, these 3×3 mask matrices are a basic form in these types of applications, but can have many variations by changing the weighting of the cells. We have experimented two of the well known mask matrices in the literature, with which we run experiments. The first option is the one developed by Roberts and Sobel's is the second. These methods are effective, as demonstrated by the abundance of their applications in the literature. It is also important to mention, that these methods require fewer resources for implementation in digital hardware than other methods such as Canny, LoG (Laplacian of Gaussian), Prewitt, Frei-Chen.

A. Optical flow computation experiments with different video sequences as input data

In order to test and validate the algorithm developed and implemented at first in software, we chose three different video sequences. Two of them are real and the third video is an animation.

In the top left corner of *Fig. 1* we can see a frame from one of the video sequences used as test data, called Grid (31 frames with a resolution of 320×240, with 8 bits/pixel). These images were used as inputs to the algorithm for calculating the optical flow (Optical Flow - OF).

Examples of calculating the horizontal (top right) and vertical (bottom left) gradient of the Grid sequence of video frames can be seen in *Fig. 1*. The detection of horizontal and vertical edges is the next step performed, as the bottom right section of *Fig. 1* shows.

In *Fig. 2* we can see one frame of the test video sequence called *Anim* (51 frames with a resolution of 200×200, with 8 bits/pixel).

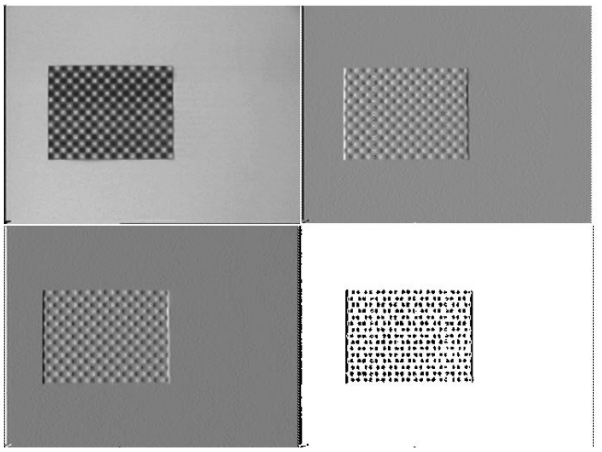

Figure 1: The GRID video sequence's frame, with computed horizontal, vertical gradients and detected combined edges.

We also have implemented a feature of the program to calculate the gradient direction obtained with the following trigonometric relationship:

$$\phi = \arctan\left(\frac{B_v(j,k)}{B_h(j,k)}\right) \tag{2}$$

After reading frames of the video sequence files, the first operation performed by the method's testing program is a Gaussian filtering with a filter matrix of 5×5 pixels. This first step is followed by the calculation of vertical, horizontal and combined gradients, with results stored separately. The algorithm continues with the positive and negative edge detection based on the frame intercorrelations, than it comes to determining the optical flow.

The effort invested in writing this software without the use of existing function libraries for image processing, has paid off in the next phase of the project – presented in this paper – the FPGA hardware implementation of the method using hardware description language (VHDL) and Xilinx ISE development environment (Design Suite 14.7).

B. Description of the system designed and built for parallelized implementation on FPGA

In *Fig. 3* polygons with green background symbolize BRAM modules (Block RAM) of the FPGA circuit. These were configured using IP Core Generator tool from Xilinx ISE development system Design Suite to store a selected video frame sequence (image grayscale, 8-bit resolution of 200×200 pixels), using 10 of 38 Kbits of BRAM memory.

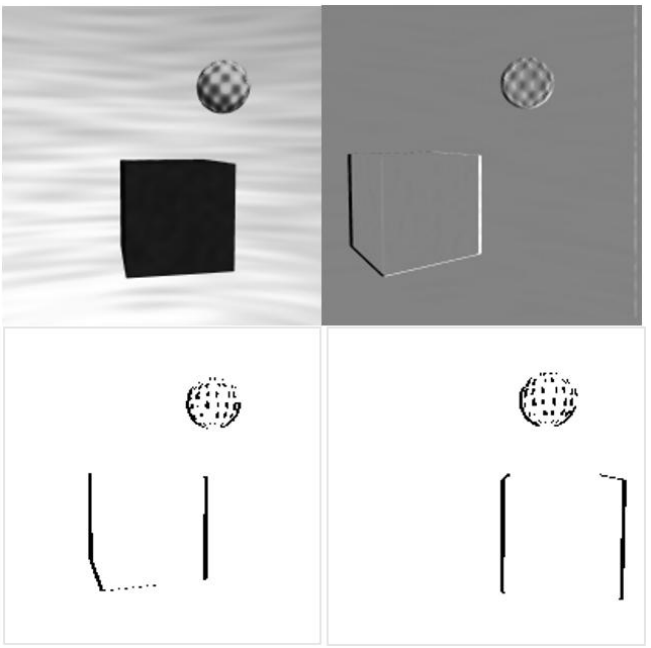

Figure 2: The ANIM video sequence's frame, with computed horizontal, vertical gradients and detected combined edges.

We developed a double pipe-line structure to parallelize execution of operations. The calculation steps determined in the C++ program were implemented here in separate modules that are synchronized by a finite state machine (FSM). Observe the two parallel pipe-lines, processing data from two consecutive frames of video.

Each of these performs the following steps:

- Scanning the image to determine the minimum, maximum and average values, data needed for subsequent calculations, scaling, etc.
- It runs matrix Gaussian filtering algorithm.
- Reading consecutive pixel values, that are inserted into the pipe-line which runs several phases:
  - ✓ vertical and horizontal gradient computation,
  - ✓ positive and negative edge detection,
  - ✓ determining gradient direction.
- After completing these calculations, the results are saved in separate BRAM modules.
- Based on these partial results, which can be computed from two consecutive images in a synchronized manner, the method calculates their intercorrelation.

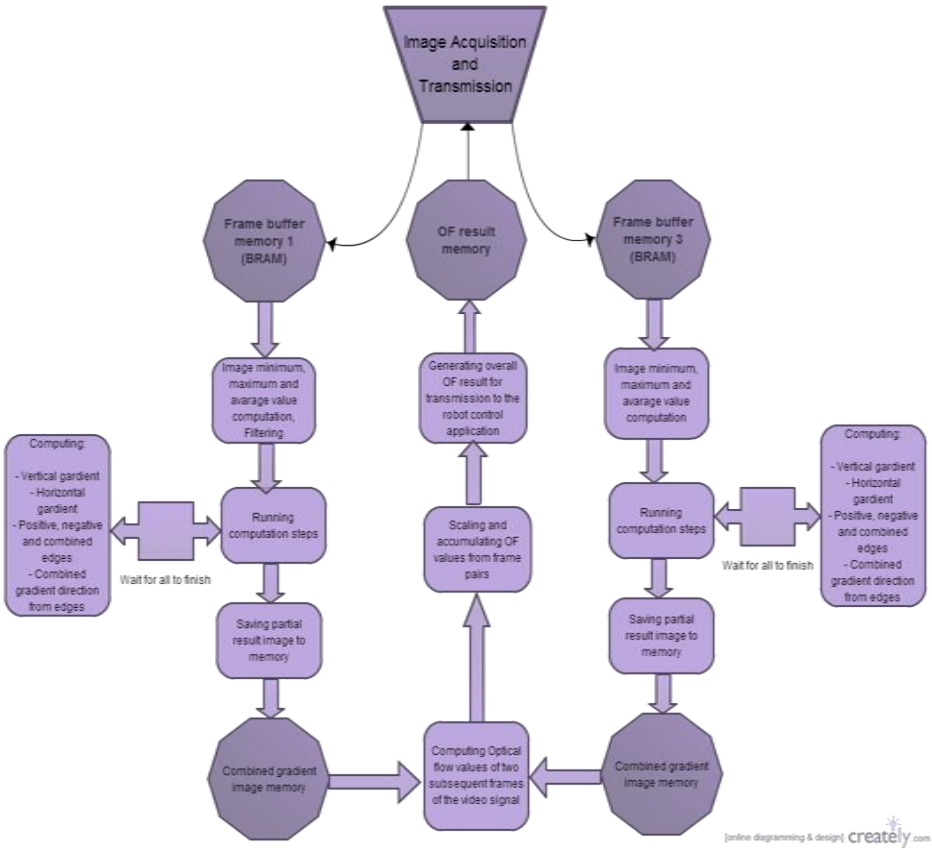

Figure 3: Block diagram and sequence of operations implemented on the FPGA.

- This yields the corresponding OF values.
- The OF values will be scaled and accumulated from several pairs of images in the sequence.
- The end result is saved in the dedicated OF BRAM memory, from where it can be passed on to an application that will use it.

The state-diagram of the finite-state-machine (FSM) controlling one thread of the pipe-line structure is shown in *Fig. 4*. Note the loop formed by the states 1, 2, 3 and 5 corresponding to the data input phase from the BRAM memory (Frame Buffer in *Fig. 4*) and the image parameters computation. It then passes to the second loop (states 4, 5, 6, 7 and 8) where it performs the calculations of the gradient, edge detection, OF, etc. The last state saves the results. The novelty consists in a method able to detect the image parameters while running the filter algorithm, thus saving an entire image scanning cycle.

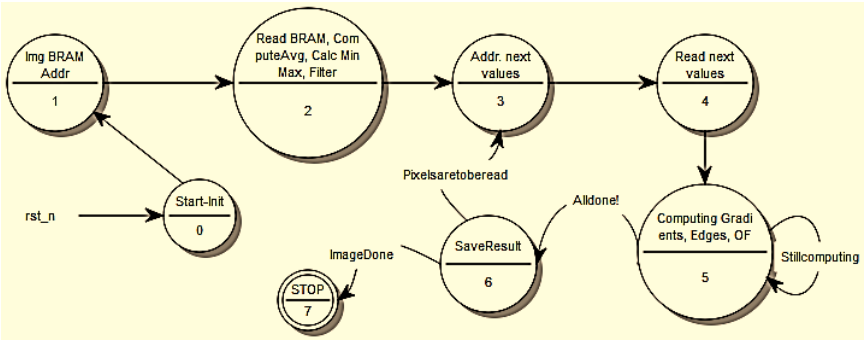

Figure 4: State-diagram of the FSM controlling one thread of the pipe-line structure.

The new, resource-efficient motion estimation method developed by our team, uses an OF extraction algorithm consisting of the following steps:

a) Based on the detections results of the previous stages, from each frame of the video sequence we have generated a flag matrix signaling the edge positions in the image. A flag value (logical 1 bit value) is placed on the *x, y* coordinates of the generated matrix in the vicinity of the locations where an edge is detected. While scanning the input images with the Sobel filter matrixes, for each output value a single bit of the flag matrix is generated, therefore reducing the size of the data to be processed in the next step.

b) The next step consists in scanning the flag matrix pairs generated from two consecutive frames with a 5×5 pixel window to determine the local direction of travel (motion) of the existing edges.

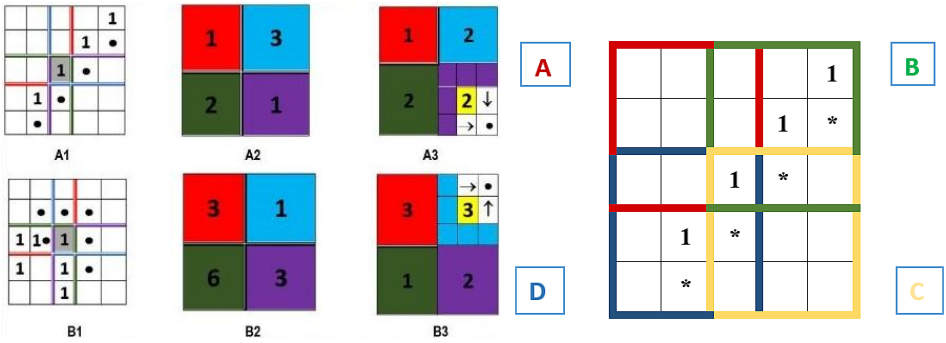

Figure 5: The introduced method for determining the local displacement of edges.

As can be seen in the examples in *Fig. 5*, the evaluation windows are divided into four quadrants, and the resulting value of the direction of movement will be saved to coordinates that are at the intersection of these quadrants of 3×3 pixels. *Fig. 5* - A1 shows a local displacement of the edges formed by 4 points. The number of points in each quadrant is then calculated (*Fig. 1* - A2) for two

consecutive evaluated frames (*i* and *i+1*). An increase in the number of bits in a quadrant shows the direction of movement of the edge. The chosen quadrants will have an increased number of flag bits (*Fig. 1* - A3).

3. Test results of the developed embedded OF extraction system

Translation (synthesizing) programs of functional hardware description languages like VHDL to Verilog do not result in a series of instructions executed sequentially but in a draft of a digital logic circuit required to perform the algorithm described.

In this respect, the test - debug - of these programs is achievable through circuit simulation techniques. However, in order to simulate a digital circuit, implemented using a hardware description language, we need a testbench module (also developed in VHDL or Verilog) that generates input signals for the unit under test (UUT). These will yield time-varying output signals of the UUT, that will reflect the behavior of the designed circuit, thus aiding the debug process.

These VHDL simulation codes can check the outputs of the module under test (UUT - Unit Under Test) for the generated inputs, and returns status messages or error signals if detected. The Xilinx environment provides the ISIM simulation compiler that generates the graphical representation of the input signals, internal signals of the UUT and outputs in the form of timing diagrams.

In this section of the paper we present a few of these diagrams, for the implemented OF extraction project.

In *Fig. 6* one can follow the partial simulation of one thread of the pipe-line structure in *Fig. 3*. Note the double addressing of the dual-port BRAM memory to get two values simultaneously in the same clock cycle. The finite state automaton executes the first loop (states s1, s2, s3 and s5) to control the sequential reading of BRAM and calculation of the image parameters. After reaching the highest memory address (0 ... 39 999, for an image of 200×200 pixels) the FSM transitions to state s4, where the final values of the calculated parameters are available. It is important to note the time required for these operations, which is 800μs in accordance with the same timing diagram.

One of the steps difficult to implement in hardware was the calculation of the image mean values, because it requires at least one division operation, which is only possible in a digital circuit to values which are equal to 2^n.

To solve this problem and minimize the error introduced with divisions by 2^n values closest to the current divider values, we used the following method: division was achieved by using shift registers, and the error was reduced by averaging two consecutive displacement values.

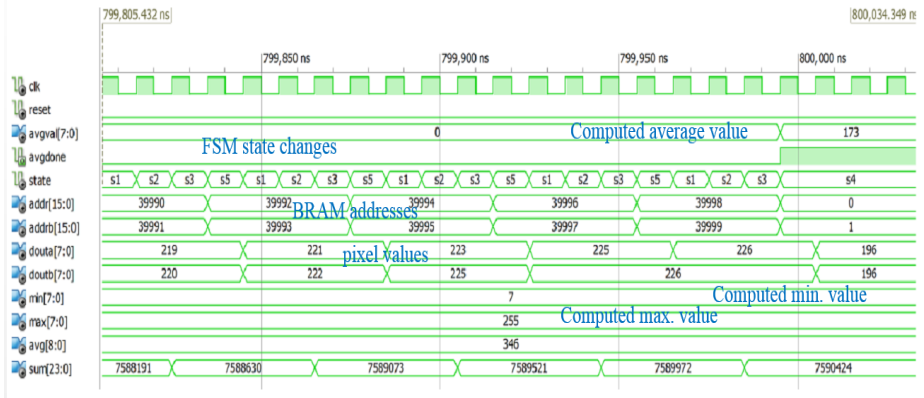

Figure 6: Simulation of reading the input values from BRAM memory (frame buffer).

The image in Fig. 7 shows a complete execution cycle of the developed algorithm by the FSM pipe-line control structure. One can observe in *Fig. 7* the evolution of calculating the image minimum and maximum values, followed by the second loop, with the gradient computations and scaling. It should be noted in this case, that the total execution time is approximately 1.6 ms. As it results from the analysis of the time diagram in *Fig. 7*, each partial result obtained in state s8 is saved and sent to the next component, namely at the end is placed in a BRAM memory called *Combined gradient memory image* in *Fig. 3*.

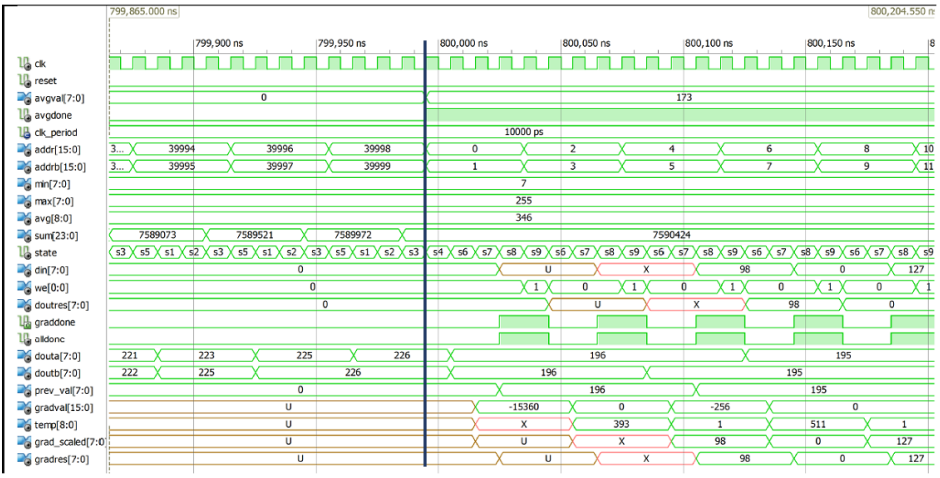

Figure 7: Simulation result showing the passage (vertical marking line) from the parameters calculation loop to calculating image gradients.

Since the hardware resources required to implement this computation flow occupies only about 1-2% of the capacity of the FPGA circuit used (a Xilinx

Virtex 5 FX30T) (Table 1), this structure should be instantiated more than twice, thus leading to a more efficient parallelized structure with the possibility of processing multiple frames of the video sequence simultaneously. This extension, however, is restricted by the number of available FPGA BRAM memories (68). In its current form, with only two parallel pipe-line structures (two frames processed simultaneously) the project uses at least 40 BRAM modules.

Table 1: *Device utilization summary for one thread of the implemented OF extraction pipe-line structure*

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice Registers	108	20480	0%
Number of Slice LUTs	206	20480	1%
Number of fully used LUT-FF pairs	79	235	33%
Number of bonded IOBs	20	360	5%
Number of Block RAM/FIFO	20	68	29%
Number of BUFG/BUFGCTRLs	3	32	9%

4. Validating the FPGA implementation of the new method using viciLAB

viciLab [8], [9] is a remote/local FPGA prototyping platform, with GUI console toolsuite support. It enables the user to create and implement a digital logic component application and GUI console. The viciLab tools perform automated creation of the remote/local design FPGA bitstream from a VHDL model description, and perform remote or local Digilent Nexys3 module Xilinx Spartan-6 FPGA configuration. The system also permits user-specific real-time FPGA application development with interactive control/visualization console. The viciLogic wrapper integrates the user design with the FPGA hardware core, and generates design metadata to aid automation and faster and easier GUI prototype development. The HDL parsing process also produces a machine readable description of the HDL design structure, which is used during course building and client GUI application creation to automate the creation of interactive animations. The wrapper auto-detects and connects the SDRAM interface, and clock and reset signals, and provides a user menu for defining signal connections to FPGA module display devices (LEDS and 7-segment displays). The DSPModule is the area where the application's main processing elements are placed. The GUI written in Python retrieves the video feed from a PC's webcam and saves the image frames into the cellular RAM memory of the Nexys 3 development board. The wrapper extracts the image data from the external RAM and drives the signals necessary for the DSPModule (dspBlock)

to perform the designed computation steps. The implemented computation processes of the dspBlock performs the following steps, also shown in *Fig. 8*:

- The video feed from the webcam is converted by the Python GUI into the format with a resolution of 100×100 and 32 bits per pixel.
- The video data is stored in the FPGA board's SRAM in a grayscale format, but still in 32 bits/pixel. In the processing phase, though, only 8 bit / pixel are used as input values.
- Each frame of the webcam video signal is scanned with a 5×5 window in order to perform a Sobel edge detection, using a filter matrix.
- The edges are stored in a 20×20 bit matrix, according to the local edge values found by the previous step.
- This matrix is then processed using the previously presented method with 5×5 local OF detection windows. The 16 resulting windows are processed in parallel by the dspBlock using as many separate VHDL processes.

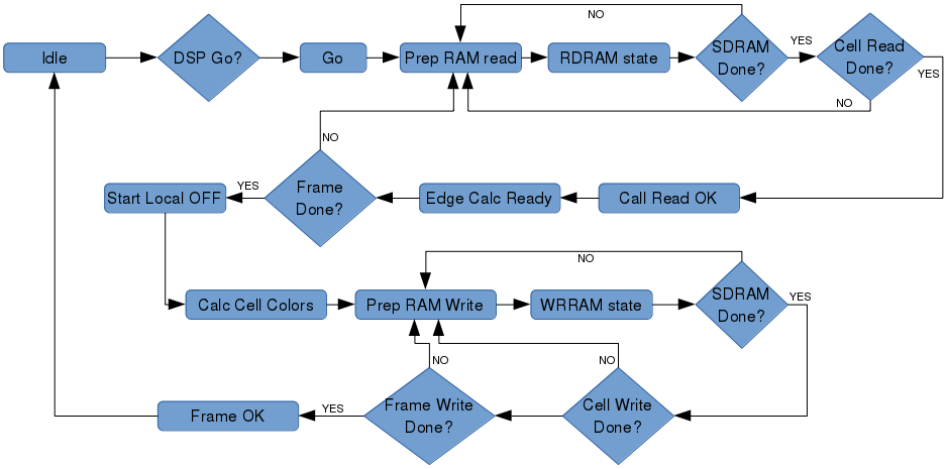

Figure 8: State diagram of the finite state machine controlling the image processing circuit.

As *Fig. 8* shows, the finite state machine (FSM) controlling the OF computation circuits contains a module that is responsible for the display of the OF result values. In order to ease the evaluation of the outputs, a color code has been assigned to each of the eight possible (45, 90, 135, 180, 225, 270, 315 or 360 degrees) optical flow direction values yielded by the circuit.

These colors will fill a square, as *Fig. 9* presents, (overlaying the displayed edges) corresponding to the 5×5 bit windows of the edge bit matrix used as inputs to the displacement computation phase. These will show the OF direction

in the respective areas of the image, yielding the local OF values. Averaging these gives the overall OF value.

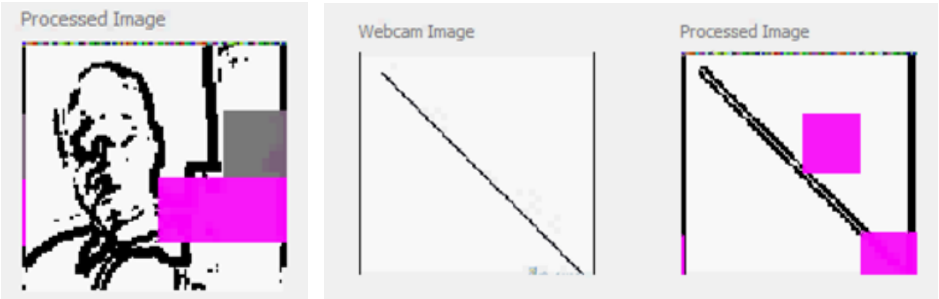

Figure 9: State diagram of the finite state machine controlling the image processing circuit.

Measurements have shown that one complete cycle for local OF and overall OF determination computes in under 25 microseconds, well within real-time requirements.

5. Raspberry Pi implementation of the developed OF method

In order to test the performance of the method on a different embedded platform we have implemented it on a Raspberry Pi compact computer.

To achieve real-time video processing of captured frames from a 30 fps 640×480 PX resolution camera, the platform was chosen to be a Raspberry PI 2 model B. The credit card sized computer contains a Broadcom SoC consisting of a 900 MHz Quad-core ARM CortexA7 CPU, a 250MHz Broadcom Video Core IV GPU, with 1 GB memory. It is also capable of sending data with a speed of 2 Gbps through the CSI-2 connector from a dedicated 5 megapixel camera directly to the GPU.

RGB frames from the camera module can be received with chosen resolution and speed defined in software. The maximum video recording features are 1920×1080 pixel on 30 fps, 1280×720 pixel on 60 fps and 640×480 pixel on 60/90 fps. For the application, 640×480 pixel sized frames where received on 30 fps.

Using OpenCv API to easily process images and to show the results, the procedure begins with transforming the three channel RGB frames received from the camera to one channel grayscale frames. The transformed pixel values were stored in the memory with 8 bit unsigned char values, varying the intensity of the grey value from 0 to 255. Edge detection in the frames is done by the Sobel operator, with 5×5 pixel kernels.

6. Results and conclusions

The designed OF extraction system implemented on a FPGA circuit is functional, operating in real-time. The precision of the OF computation is influenced by the sensitivity of the edge detection phase to the lighting conditions of the environment. One way to overcome this issue is to increase the framerate of the input video signal by using a dedicated camera directly attached to the FPGA development board. We have experimented with the use of an Omniview OF7670 sensor to replace the PC webcam, and found that the framerate would be increased tenfold. This is currently at about 3-5 fps due to the latency of the data communication via the wrapper core between the dspBlk and the webcam. By using the dedicated camera we can reach up to 30 fps with the same dspBlk structure. The limitation in this case proved to be the resources of the Spartan 6 FPGA on the Digilent Nexys 3 board supported by viciLab.

There is room for expansion in this type of project, but with certain limitations. The alternative is, however, the use of the dedicated processor module (PowerPC440 core) of the FPGA used for the execution of those tasks that require sequential steps. On the other hand, by introducing this component into the system, other problems can be solved, such as accessing the external DDR-2 RAM modules of the used OPUS FPGA development platform, as well as real-time image acquisition as input, using peripheral interfaces attached to it.

All these avenues of development will be studied and, if favorable feasibility is found, will be exploited in later stages of the research project.

The viability of the implementation results will need to be validated by demonstrating the method with a mobile robot. The final demonstration will show the collision-free, (semi-)autonomous drive of a mobile robot or even of a group of collaborating robots in a highly-cluttered environment. The implemented systems will yield a new class of artificially intelligent robots that can adapt their hardware structure in order to behave better in a changing environment. Individual or collaborating groups of robots with these abilities could be used in a variety of reconnaissance or monitoring tasks. For instance the capability to assimilate and share acquired knowledge about its environment can be useful in scenarios where hazardous spaces need to be explored and mapped fast (ex. search in earthquake-damaged buildings).

Acknowledgements

The research presented in this paper was supported by the European Social Fund under the responsibility of the Managing Authority for the Sectoral Operational Programme for Human Resources Development, as part of the grant POSDRU/159/1.5/S/133652.

References

[1] Horn, B. K. P., Schunck, B. G. "Determining Optical Flow", Technical Report. Massachusetts Institute of Technology, Cambridge, MA, USA, 1980.

[2] Lucas, B., Kanade, T., "An iterative image registration technique with an application to stereo vision," In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[3] Floreano, D., Pericet-Camara, R., Viollet, S., Ruffier, F., Brückner, A. et al. "Miniature curved artificial compound eyes," in *Proceedings of the National Academy of Sciences*, vol. 110, num. 23, p. 9332–9337, 2013.

[4] Duhamel, P.E. et al., "Hardware in the Loop for Optical Flow Sensing in a Robotic Bee," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, 2011, pp. 1099–1106.

[5] Ruffier, F., Franceschini, N., "Optic flow regulation: the key to aircraft automatic guidance," *Robotics and Autonomous Systems*, vol. 50, no. 4, pp. 177–194, March 2005.

[6] Serrão, M., Rodrigues, J. M. F., du Buf, J.M.H., "Navigation Framework Using Visual Landmarks and a GIS," *Procedia Computer Science*, vol. 27, pp. 28-37, 2014.

[7] Xiuqing, W., Zeng-Guang, H., Feng, L., Min, T., Yongji, W., "Mobile robots' modular navigation controller using spiking neural networks," *Neurocomputing*, vol. 134, pp. 230–238, June 2014.

[8] Tomasi, M., Barranco, F., Vanegas, M., Díaz, J., Ros, E., "Fine grain pipeline architecture for high performance phase-based optical flow computation," *Journal of Systems Architecture*, vol. 56, no. 11, pp. 577–587, November 2010.

[9] Botella, G., Ros, E., Rodriguez, M., Garcia, A., Romero, S., "Pre-processor for bioinspired optical flow models: a customizable hardware implementation," in *Electrotechnical Conference MELECON 2006*, IEEE Mediterranean, Limassol, 2006, pp. 93–96.

[10] Zhaoyi, W., Dah-Jye, L., Brent E., N., "FPGA-based Real-time Optical Flow Algorithm," *Journal of Multimedia*, vol. 2, no. 5, pp. 38-45, September 2007.

[11] Díaz, J., Ros, E., Pelayo, F., M. Ortigosa, E., Mota, S., "FPGA-Based Real-Time Optical-Flow System," IEEE *Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 274–279, February 2006.

[12] Schlessman, J. et al., "Hardware/Software Co-Design of an FPGA-based Embedded Tracking System," in *Conference on Computer Vision and Pattern Recognition Workshop CVPRW '06*, 2006, pp. 123–123.

[13] Chase, J., Nelson, B., Bodily, J., Zhaoyi W., Dah-Jye, L., "Real-Time Optical Flow Calculations on FPGA and GPU Architectures: A Comparison Study," in *16th International Symposium on Field-Programmable Custom Computing Machines*, 2008, pp. 173–182.

[14] Barranco, F., Tomasi, M., Diaz, J., Vanegas, M., Ros E., "Parallel Architecture for Hierarchical Optical Flow Estimation Based on FPGA," *IEEE Transactions On Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 6, pp. 1058–1067, June 2012.

[15] Jin, S., Kim, D., Nguyen, D. D., Jeon, J. W., "Pipelined Hardware Architecture for High-Speed Optical Flow Estimation using FPGA," in *18th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, Washington DC, 2010, pp. 33–36.

[16] Grabe, V., Bulthoff, H. H., Robuffo G. P., "On-board velocity estimation and closed-loop control of a quadrotor UAV based on optical flow," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, 2012, pp. 491–497

[17] Vatavu, A., Danescu, R., Nedevschi, S., "Real-Time Dynamic Environment Perception in Driving Scenarios Using Difference Fronts," in *Proceedings of the 2012 IEEE Intelligent Vehicle Symposium*, June 2012, Alcala de Henares, Spain, pp. 717–722.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 20–34

Open-Source Research Platforms and System Integration in Modern Surgical Robotics

Árpád TAKÁCS,[1] Imre J. RUDAS,[1] Tamás HAIDEGGER[1,2]

[1]Antal Bejczy Center for Intelligent Robotics,
Óbuda University, Budapest, Hungary,
e-mail: {arpad.takacs; imre.rudas; tamas.haidegger}@irob.uni-obuda.hu
[2]Austrian Center for Medical Innovation and Technology,
Wiener Neustadt, Austria

Manuscript received June 27, 2015; revised November 16, 2015.

Abstract: In modern medical research and development, the variety of research tools has grown in the previous years significantly. It is crucial to exploit the benefits of shared hardware platforms and software frameworks in order to keep up with the technological development rate. Sharing knowledge in terms of algorithms, applications and instruments allows researchers to help each other's work effectively. This is a relatively new trend in the traditionally closed domain of Computer-Integrated Surgery, where community workshops and publications are now providing a thorough overview of system design, capabilities, know-how sharing and limitations. This paper overviews the emerging collaborative platforms, focusing on available open-source research kits, software frameworks, cloud applications, teleoperation training environments and shared databases that will support the synergies of the diverse research efforts in this area.

Keywords: surgical robotics, shared hardware platforms, software frameworks, cloud applications, teleoperation training.

1. Introduction

Medical robotics is one of the most rapidly developing fields of modern robotics, which is partly due to its competitiveness. The surgical robotics market is estimated to grow at an annual rate of 12% through 2018, reaching a size of $18 billion [1]. The manufacturers and developers consider these high-end hardware platforms and software programs the key assets of the research, protecting them in various ways (patents, industrial secrets etc.). Nevertheless,

The results of this study were partially presented at the 5th International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics 2015.

there is a growing need for open-source and easily accessible platforms and software/hardware solutions to facilitate future development in all fields of robotics. Various high-end robot controllers have already been available for such purpose, e.g. the Real-Time Application Interface (RTAI) for Linux platforms [2]. There has been a significant rise of open-source efforts in the field of Computer-Integrated Surgery (CIS), encouraging numerous key industrial stakeholders to support these efforts.

In medical robotics, due to the uniqueness and physical dimensions of hardware platforms, there is a lack of mobility and accessibility for most of the developers in the community. The sharing of program codes, toolkits and frameworks are usually carried out through online databases, granting access to the hardware through cloud-based control platforms. In the past decades, the concept of medical robotics has never been separated from the terms of telerobotics and teleoperation [3]. With the recent rise of cloud robotics, a promising perspective has appeared where not only the development and research, but the process of testing and operation could also be applied through cloud-based platforms. The aim of this paper is to provide a thorough overview into the emerging collaborative platforms, focusing on available open-source research kits, software frameworks, cloud applications, teleoperation training environments and shared domain ontologies for surgical robotics.

The paper is organized as follows: in Section 2, the most relevant medical robotics software platforms are presented. In Section 3, certain issues are discussed with the system-related application programming interfaces, addressing the research hardware environment in Section 4. Section 5 is dedicated to the da Vinci Research Kit, followed by a review of current community efforts and system integration in Section 6. The paper is concluded with a discussion and the projection of a future roadmap.

2. Software in medical robotics

In this paper, some of the open-source and free-to-use platforms and software solutions are discussed, where one has complete control over the software components, allowing program code customization and re-implementation. In most cases, there exists a wide community of developers, which continuously maintains, develops and updates the software. The most important of these open-source platforms are listed below.

2.1. 3D Slicer

3D Slicer[1] is the most popular and most widely used open-source, free software package that can be used for visualization and image analysis, particularly for medical imaging [4]. The software was designed to be natively available for multiple operation system platforms, including Windows, Linux and Mac OS X. 3D Slicer is operated based on the NA-MIC kit and other software components [5]. These include the Visualization Toolkit (VTK) and the Insight Segmentation and Registration Toolkit (ITK), which will be discussed later in this paper. The modularity of the Slicer 3D is shown in *Fig. 1*. Both research and clinical projects employ the 3D Slicer for applications, such as brain tumor removal [5] or prostate biopsy, using the OpenIGTLink robotic platform [7]. The 3D Slicer has a remarkable flexibility and connectivity to other software platforms, such as the Open Core Control software for surgical robotics [8]. Besides visualizing the actuator positions in a given application, *virtual fixtures* (control boundaries for safety that should not be crossed during an intervention) can also be specified.

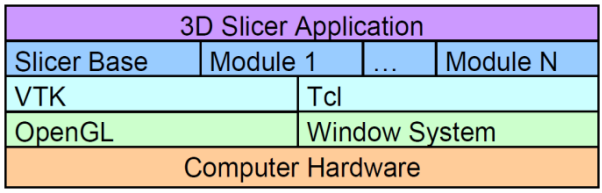

Figure 1: The modularity of the 3D Slicer [4].

2.2. Visualization Toolkit

The Visualization Toolkit[2] (VTK) is a free toolkit, its primary use includes image processing, visualization, 3D volume rendering and scientific visualization [9]. Due to the object oriented design, other modules can be integrated into the software for modification and expansion purposes [10]. VTK includes a C++ class library and other interpreted interface layers, such as Tcl/Tk, Java and Python.

[1] http://www.slicer.org/
[2] http://www.vtk.org/

2.3. Insight Segmentation and Registration Toolkit

Similarly to the VTK, the Insight Segmentation and Registration Toolkit[3] (ITK) is an open-source, cross-platform system, mostly used in segmentation and image registration problems [11]. ITK includes a vast collection of biomedical image analysis that was created within the framework of Visible Human Project [12]. ITK is also used by 3D Slicer, as a component.

2.4. Computer Integrated Surgical Systems and Technology

Computer Integrated Surgical Systems and Technology[4] (CISST) is an extended collection of libraries, a useful tool for many CIS and medical robotics applications. The function and libraries are used by e.g. the Surgical Assistant Workstation[5] (SAW), a cross-platform framework based on C++ for device integration in computer assisted intervention applications. CISST supports interchangeability, therefore all the devices that meet the basic requirements are interoperable with each other. The requirements for interchangeability are based on two main restrictions imposed by the use of *commands*: 1) parameters must be derived from a base type, and 2) a finite number of signatures are supported [13].

2.5. Image-Guided Surgery Toolkit

The Image-Guided Surgery Toolkit[6] (IGSTK) was created to support the development of image-guided applications, where intra-operative tracking is also possible [14]. The main features of the IGSTK toolkit include [15]:

- reading and display of medical images,
- interface to common tracking,
- GUI and visualization capabilities,
- multi-scale axial view,
- four-quadrant view (axial, sagittal, coronal or 3D),
- point-based registration,
- robust common internal services for logging, exception-handling and problem resolution.

[3] http://www.itk.org/
[4] https://www.cisst.org/
[5] https://www.cisst.org/Saw
[6] http://www.igstk.org/

2.6. Medical Imaging Interaction Toolkit

The Medical Imaging Interaction Toolkit[7] (MITK) is an open-source software system for development of interactive medical image processing software. MITK combines the VTK and ITK toolkits with several customized interactive components [16]. The versatility of the hardware platforms is increased due to the combination of these elements. The software system can be extended with additional modules e.g. the built-in interactive image segmentation.

2.7. Public Software Library for UltraSound

The Public Software Library for UltraSound[8] (PLUS) is a software platform written in C++ and built on the NA-MIC Kit [17], [18]. PLUS contains library functions and applications, supporting tracked ultrasound image acquisition, calibration and processing. The software package is equipped with numerous tools that are related to ultrasound data processing, extended with the support of optical and electro-magnetic trackers or other imaging devices [19].

2.8. National Alliance for Medical Image Computing

The National Alliance for Medical Image Computing[9] (NA-MIC) is an interdisciplinary team of medical experts, software engineers and computer scientists, developing new computational tools for medical image data visualization and analysis. Therefore, the NA-MIC kit is not standalone software, but rather a collection of methodologies and tools [20]. Numerous software packages are integrated in this kit, such as the 3D Slicer, VTK and ITK.

2.9. Surgical Assistant Workstation

The main purpose of the Surgical Assistant Workstation (SAW) is to integrate different components of a robotic surgical system, using and reusing the elements in the system structure, as shown in *Fig. 2*. Developed by the Johns Hopkins University, SAW supports the most common tools of CIS, such as tracking systems, stereo viewers, haptic devices, and other common hardware platforms, including 3D Slicer, various medical research robots and the da Vinci master console and robotic arms, created by Intuitive Surgical Inc. [21]. SAW is written in C++ and the research was founded by the National Science

[7] http://www.mitk.org
[8] https://www.assembla.com/spaces/plus/wiki
[9] http://www.na-mic.org/

Foundation[10] (NSF). Thanks to the high level of modularity, SAW can be extended with new components in many research systems. The easy connectivity among multiple devices allows one to integrate them into a sophisticated surgical system. An example was demonstrated by JHU by integrating a snake robot with the da Vinci console for laryngeal surgery [22].

2.10. The Common Toolkit

Common Toolkit[11] (CTK) supports biomedical image computing, licensed under Apache 2.0. The toolkit can be used for academic, commercial and other purposes free of any restrictions. The main scope of the current CTK development efforts includes the *DICOM*, *DICOM Application Hosting*, *Widgets* and *Plugin Framework* [23].

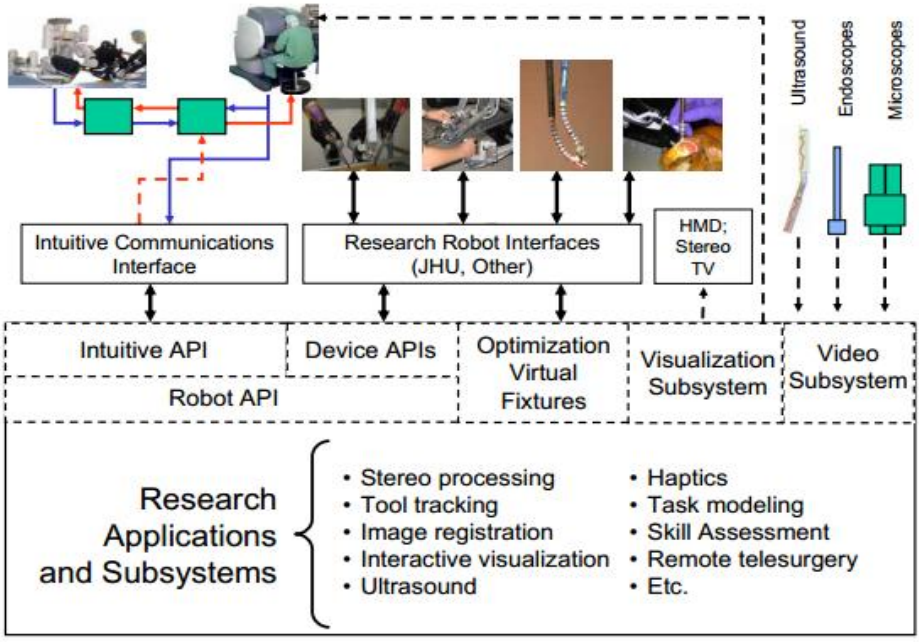

Figure 2: Architecture and capabilities of the Surgical Assistant Workstation [22].

[10] http://www.nsf.gov/
[11] http://www.commontk.org/

2.11. The NifTK

NifTK[12] is a translational imaging platform combining various toolkits developed at the Centre for Medical Image Computing (CMIC), University College London (UCL). These toolkits include the NiftyReg, a collection of programs to perform rigid, affine and non-linear registration for medical images; the NiftySim, a solver for non-linear elastic or viscoelastic deformation; the NiftyRec, a software package for fully 3D Stochastic Emission Tomographic Reconstruction; the NiftySeg for image segmentation; and NiftyView, a cross-platform graphical user interface providing an entry point to the above mentioned packages. The NifTK is widely used in rigid instrument tracking and computer aided surgery planning [24].

3. System-related APIs

In general, the academic community aims for the development of generic development bits, supporting a wide range of components and devices, such as the examples listed in Section 2. However, some manufacturers have recently developed particular Application Programming Interfaces (APIs) for various systems, such as Medtronics' intra-operative navigation platforms, which features a StealthLink research interface [25], or the OpenIGT Link connection with the KUKA Sunrise controller and 3D Slicer [26].

Today's most deployed surgical system, the da Vinci Surgical System[13] (Intuitive Surgical Inc., Sunnyvale, CA) is not provided with open access by default. Data cannot be retrieved from the robot, programs and components are not subject to change and one cannot extract any information about the basic operation principle, mostly due to liability issues. Limited amount of information can be recorded using various data collection tools [27]. In some cases, the manufacturer allows access to previous generations of their systems, which become transparent by the provided open-source software [28].

4. Research hardware environment

On closed systems, it is fairly difficult to conduct fundamental research, for obvious reasons. Therefore, in order to achieve technological development, some of the manufacturers grant partial accessibility to their closed systems. In the case of the da Vinci, there exists a real-time stream of kinematic and user event data from the robot that can be read, provided by the de Vinci Application

[12] http://cmic.cs.ucl.ac.uk/home/software/
[13] http://www.intuitivesurgical.com/

Programming Interface. It is important to mention that the total replacement of certain components, such as the controller body, can transform the da Vinci system into and open-source platform.

Raven II is one of the most successful open-source robotic platforms. Developed at the University of Washington and supported by DARPA[14], the Raven II became the greatest competitor of the da Vinci system [28]. Furthermore, with the help of the National Institutes of Health[15] (NIH), 8 robots have been created and distributed to European and North-American locations. Currently, the Raven II research platform can be purchased from Applied Dexterity Inc.[16] The platform operates based on the Robot Operating System (ROS) architecture.

5. The da Vinci Research Kit

The da Vinci Research Kit (dVRK) is one of the most capable research platforms in surgical robotics. In fact, the kit is a collection of retired, first-generation da Vinci robot components and tools, provided with additional open-source control electronics and software.

5.1. Hardware components

The dVRK contains the components listed below:

- Two da Vinci Master Tool Manipulators (MTMs),
- Two da Vinci Patient Side Manipulators (PSMs),
- A stereo viewer,
- A foot pedal tray,
- Manipulator Interface Boards (dMIBs),
- Basic accessory kit.

The research kit contains the original, unmodified mechanical components, therefore it is possible to transform a da Vinci Classic system into a research kit, although some of the components are not available for researchers due to their commercial use. In the dVRK hardware set, the Endoscopic Camera Manipulator (ECM) is not included along with several other components from the original system, but the lack of these elements is not a major issue from the development point of view. In general, for research purposes, the control electronics and control software are the most essential parts of the system.

[14] http://www.darpa.mil
[15] http://www.nih.gov/
[16] http://applieddexterity.com/

Recently, a novel, open controller platform was created by JHU, Worchester Polytechnic Institute (WPI) and their partners [30]. The source files of the control electronics were also published online. The research platform is equipped with an IEEE 1394a Firewire interface, capable of maintaining a communication speed of 400 Mbit/sec. In order to achieve a satisfactory degree of security and reliability, it is crucial to create real-time communication between the devices in the system. The control box includes two FPGA modules and two Quad Linear Amplifiers (QLA), as shown in *Fig. 3*. The assembly described above is capable of driving and controlling a single robotic tool. Two da Vinci Master Tool Manipulators (MTMs) and two da Vinci Patient Side Manipulators (PSMs) can be controlled using four sets of control electronics, requiring a total of 8 pieces of FPGAs and QLAs. The integration of the dVRK to a retired, fully operational da Vinci robot is shown in *Fig. 4.*

The da Vinci Research Kit is based on the centralized computation and distributed I/O architecture [31]. The main advantage of this structure is that there is only one control electronics that maintains contact with the peripheral inputs and outputs, allowing the central computer unit to perform the calculations, located at the control units. In general, the central unit is a Linux-based computer with some real-time component expansion.

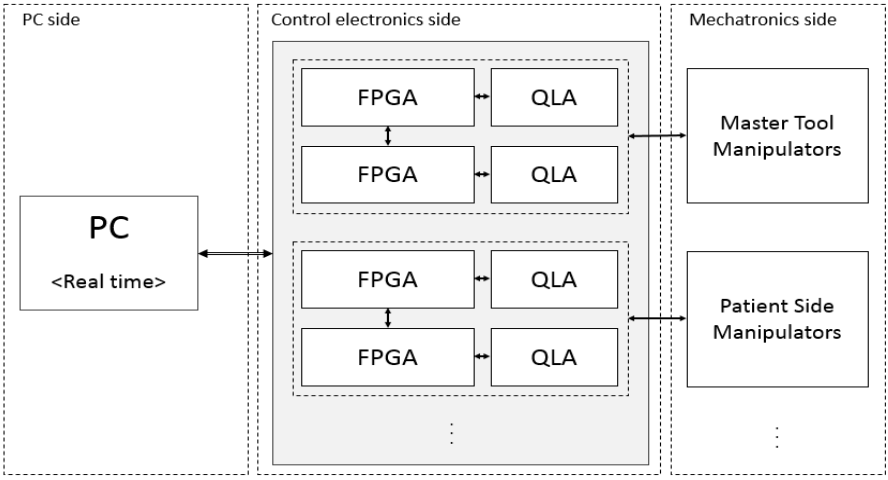

Figure 3: Schematic representation of the hardware structure Workstation [32].

5.2. Low level software architecture

The FPGA module firmware is available online[17] and published under a BSD license, therefore it can be freely modified. The RT-FireWire is one of the best approaches to solve the real-time communication between the subsystems over Firewire, while the communication implementation is achieved through standard Linux C++ libraries [33].
The PC-side operating system is preferably Linux-based, as there exists a real-time extension (RTLinux), a Linux OS that runs under the supervision of a hard real-time microkernel [34]. The software architecture, as a whole, can be divided into five *functional layers* (I-V) and three *development layers* (A-C) [35]. The functional layers, implemented on the PC side, are stratified by the complexity of their function, while the development layers are sorted by the programming language complexity they use. The open-source property is extensively supported by the previously described SAW and CISST libraries, allowing the system to be used as a completely open research platform.

Figure 4: The da Vinci Surgical System and the da Vinci Research kit. System components: Patient Side Manipulators (left), the dVRK controller (middle) and the Master Tool Manipulators (right).

[17] https://github.com/jhu-cisst/mechatronics-firmware/wiki/FPGA-Program

6. System integration and current community efforts

The integration of the above mentioned systems to real hardware applications has become a widely researched and published topic recently. Beside the development and testing of the available systems, new interface standards are also created, in order to achieve a reliable closed-loop control and synchronization. The OpenIGTLink is an open-source network communication protocol that defines a messaging protocol for data transfer, primarily focusing on images, commands and joint positions [36]. Since its introduction in 2008, several use cases, actual and potential roles of the OpenIGTLink have been presented [37]. One of the use cases focuses on an MRI-compatible manipulator for prostate biopsy, where three major components were equipped with OpenIGTLink interfaces: a 3T MRI scanner, a planning and navigation software and an MRI-compatible needle placement robot [38]. Other use cases include an MRI-compatible manipulator for stereotactic neurosurgery, an open interface of lightweight robot system and a robot for transrectal ultrasound guided brachytherapy. One of the greatest advantages of the OpenIGTLink platform is the simple design and the multi-platform C++ software library, allowing the users to minimize the engineering effort, facilitating multi-site and academic–industrial collaborations while enabling transition from research to clinical use.

6.1. Use cases

There are a great number of ongoing research projects involving the dVRK, mostly focusing on the master and slave side manipulators and the vision system. One of the limitations of the da Vinci Surgical System is the difficulty of the simultaneous operation of the patient side manipulators and the stereo vision system, decreasing the efficiency of the surgical interventions. To overcome this issue, the Novel Master Interface (NMI) and the Controllable Vision System (CVS) were developed at the Seoul National University, Korea [39]. The evaluation and validation of the system's clinical applicability, peg task experiments are in progress. The NMI is a wireless communication interface including a multidirectional switch, a Bluetooth module, an encoder and a button cell battery. The multidirectional switch is mounted on the patient side manipulator, as shown in *Fig. 5*. The multidirectional switch sends data to the CVS through the Bluetooth module in real-time, controlling the camera position of a novel camera holder hardware interface.

Figure 5: The NMI structure: (a) front side, (b) back side, (c) NMI attached on the MTM.

Another setup for the da Vinci laparoscopic camera handling was developed at Óbuda University [40]. A low-cost, lightweight Computer Assisted Laparoscopic robot arm (CALap) was primarily created for surgical training purposes, which was extended with the Apollo classicalbox trainer [41]. The CALap hardware is based on a mechanical structure created using aluminum profiles, extended with additional gears and 3D-printed parts. The software design follows the master–slave concept, having high level programming and low level electronic handling approach. This allows one to integrate the setup into the dVRK system.

In the past years, the Robot Operating System (ROS) has become an essential part of robotic research and industry. The large set of libraries and tools facilitates the integration of different robot components on a single platform. To interface the dVRK with ROS, a CISST–ROS stack was developed, which allows the dVRK to communicate using ROS messages [42]. The interface has been tested through several use cases, such as an augmented reality-based 3D measuring application, motion planning framework for surgical assistance, learning by observation for surgical subtasks and satellite servicing. The latter is addressing a non-medical robotics application, such as refueling spacecrafts in on-orbit scenarios [43].

7. Conclusion and future roadmap

The da Vinci Research Kit is one of the greatest breakthroughs in the field of open-source surgical robotic research and development, which is mostly due to the direct access to an actual clinical system, even though these systems are retired and out-of-date. As of May, 2015, there are 17 dVRK research teams operating around the world, maintaining an active community through meetings and workshops. Particularly in Europe, several actions and projects (EuroSurge[18], I-SUR[19], ACTIVE[20]) have given a boost to synchronized robotics

[18] http://www.eurosurge.eu/eurosurge/

research, mostly founded by the EU Commission. IEEE Robotics and Automation Society[21] has also contributed to the generalization of surgical robotics through study groups, and there are initiatives for forming workgroups for surgical robotics ontologies. A great impact on the entire research field is expected, where more and more attention is to be given to open-source research instead of strictly commercial development.

The effectiveness of surgical robotics will evidently become higher with the use of open-source platforms and software. This paper reviewed the most widely-used, currently available software and hardware research platforms, aided with some highlights to the features and recently realized projects they supported.

Acknowledgement

T. Haidegger is a Bolyai Fellow of the Hungarian Academy of Sciences.

References

[1] ALPHANOW. Could Titan Medical Storm The Robotic Surgery Market? Available: http://alphanow.thomsonreuters.com/2014/03/titan-storm-robotic-surgery-market/

[2] Dozio, L., Mantegazza, P., "Real-time distributed control systems using RTAI," in Proc. of the *6th IEEE Intl. Symp. on Object-Oriented Real-Time Distributed Computing*, Hakodate, 2003. pp. 11–18.

[3] Takács, Á, Kovács, L., Rudas, I., Precup, R. E., Haidegger, T., "Models for Force Control in Telesurgical Robot Systems," *Acta Polytechnica Hungarica*. Forthcoming 2016.

[4] Pieper, S., Halle, M., Kikinis, R. "3D Slicer," in proceedings of the *1st IEEE International Symposium on Biomedical Imaging: Nano to Macro*, Arlington, VA, 2004. pp. 632–635.

[5] Arata, J., Tada, Y., Kozuka, H., Wada, T., Saito, Y., Ikedo, N., Fujimoto, H., "Neurosurgical robotic system for brain tumor removal," *International journal of computer assisted radiology and surgery*, vol. 6, issue 3, 2011. pp 375–385.

[6] Haidegger, T., Kovacs, L., Fordos, G., Benyo, Z., Kazanzides, P., "Future trends in robotic neurosurgery," in proceedings of the *14th Nordic–Baltic Conference on Biomedical Engineering and Medical Physics*, Riga, Latvia, 2008. pp. 229–233.

[7] Lasso, A., Tokuda, J., Hata, N., Fichtinger, G., "Robot-assisted MRI-guided prostate biopsy using 3D Slicer. NA-MIC Tutorial Contest, 2010. Available: http://www.na-mic.org/Wiki/images/4/43/ DBP2JohnsHopkinsTransRectalProstateBiopsy.pdf

[8] Arata, J., Kozuka, H., Kim, H. W., Takesue, N., Vladimirov, B., Sakaguchi, M. Fujimoto, H., "Open core control software for surgical robots." *International journal of computer assisted radiology and surgery*, vol 5, issue 3, 2010. pp. 211–220.

[9] Kitware. About Visualization Toolkit, http://www.vtk.org/VTK/project/about.html, 2014.

[19] http://www.isur.eu/isur/

[20] http://www.active-project.eu/

[21] http://www.ieee-ras.org/

[10] Schroeder, W. J., Avila, L. S., Hoffman, W. "Visualizing with VTK: a tutorial." *Computer Graphics and Applications, IEEE*, vol 20, issue 5, 2000. pp. 20–27.
[11] Kitware. About Insight Segmentation and Registration Toolkit, http://www.itk.org/ITK/project/about.html, 2014.
[12] U.S. National Library of Medicine. The Visible Human Project—Applications, http://www.nlm.nih.gov/pubs/factsheets/visible_human.html, 2014.
[13] Deguet, A., Kumar, R., Taylor, R., Kazanzides, P., "The cisst libraries for computer assisted intervention systems," in *MICCAI Workshop on Systems and Arch. for Computer Assisted Interventions*, Midas Journal, 2008.
[14] Gary, K., Ibanez, L., Aylward, S., Gobbi, D., Blake, M. B., Cleary, K., "IGSTK: an open source software toolkit for image-guided surgery," *Computer*, vol 39, issue 4, 2006. pp. 46–53.
[15] IGSTK. What is IGSTK? Available: http://public.kitware.com/IGSTKWIKI/index.php/What_is_IGSTK, 2014
[16] ITK. The Medical Imaging Interaction Toolkit (MITK), Available: http://www.mitk.org/MITK, 2014.
[17] Lasso, A., Heffter, T., Pinter, C., Ungi, T., Chen, T. K., Boucharin, A., Fichtinger, G., "PLUS: An open-source toolkit for developing ultrasound-guided intervention systems," in proceedings of the *4th Image Guided Therapy Workshop*, Boston, MA, 2011. p. 103.
[18] Lasso, A., Heffter, T., Rankin, A., Pinter, C., Ungi, T., Fichtinger, G. "PLUS: open-source toolkit for ultrasound-guided intervention systems," *IEEE Trans Biomed Eng*, vol. 61, no. 10, 2014. pp. 2527–2537.
[19] Lasso, A., Heffter, T., Pinter, C., Ungi, T., Fichtinger, G., "Implementation of the PLUS open-source toolkit for translational research of ultrasound-guided intervention systems." *MICCAI-Systems and Architectures for Computer Assisted Interventions*, 2012. pp. 1–12.
[20] Pieper, S., Lorensen, B., Schroeder, W., Kikinis, R., "The NA-MIC Kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community," in proceedings of the *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*, Arlington, VA, 2006. pp. 698–701.
[21] CISST. System Requirements for Surgical Assistant Workstation. Available: https://www.cisst.org/main/images/c/cd/SAW-SystemRequirements-Rev2.pdf, 2007.
[22] Vagvolgyi, B., DiMaio, S., Deguet, A., Kazanzides, P., Kumar, R., Hasser, C., Taylor, R., "The surgical assistant workstation," in proceedings of the *MICCAI Workshop: Systems and Architectures for Computer Assisted Interventions*. 2008.
[23] Wolf, I., "Toolkits and Software for Developing Biomedical Image Processing and Analysis Applications," in: *Biomedical Image Processing*, T. M. Deserno, Ed. Springer Berlin Heidelberg, 2011, pp. 521–544.
[24] Zombori, G., Rodionov, R., Nowell, M., Zuluaga, M. A., Clarkson, M. J., Micallef, C., Diehl, B., Wehner, T., Miserochi, A., McEvoy, A. W., Duncan, J. S., Ourselin, S., "A Computer Assisted Planning System for the Placement of sEEG Electrodes in the Treatment of Epilepsy," in *Information Processing in Computer-Assisted Interventions*, Stoyanov, D., Collins, D. L., Sakuma, I., Abolmaesumi, P., Jannin, P. Eds. Springer International Publishing, 2014, pp. 118–127.
[25] StealthLink Manual v09, Medtronic, 2005.
[26] SurgRob. OpenIGT Link connects KUKA Sunrise controller and 3D Slicer. Available: http://surgrob.blogspot.hu/2015/03/openigt-link-connects-kuka-sunrise.html.
[27] Reiley, C. E., Lin, H. C., Yuh, D. D., Hager, G. D., "Review of methods for objective surgical skill evaluation." *Surgical endoscopy*, vol 25, issue 2, 2011. pp. 356–366.
[28] Leven, J., Burschka, D., Kumar, R., Zhang, G., Blumenkranz, S., Dai, X. D., Taylor, R. H., "DaVinci canvas: a telerobotic surgical system with integrated, robot-assisted, laparoscopic

ultrasound capability," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*. pp. 811–818).

[29] IEEE Pulse. Surgical Robots in Space: Sci-fi and Reality Intersect, Available: http://pulse.embs.org/july-2014/surgical-robots-space-sci-fi-reality-intersect/, 2014.

[30] AIM WPI. daVinci Surgical Robot Robot Research System (dVRK), Available: http://aimlab.wpi.edu/research/projects/ daVinci_Robot_Research_System, 2014.

[31] Chen, Z., Deguet, A., Taylor, R. H., DiMaio, S., Fischer, G. S., Kazanzides, P., "An open-source hardware and software platform for telesurgical robotics research," in *MICCAI Workshop on Systems and Arch. for Computer Assisted Interventions*, Midas Journal: http://hdl. handle. net/10380/3419, 2013.

[32] Jordán, S. , Takács, Á., Tar, J., Rudas, I., Haidegger, T., "Towards Open Source Surgical Robotics," in proceeding of the *4th Joint Workshop on Computer/Robot Assisted Surgery*, Genoa, Italy, 2014.

[33] Zhang, Y., Orlic, B., Visser, P., Broenink, J., "Hard real-time networking on FireWire." 2005

[34] Barabanov, M., & Yodaiken, V. (1996). Real-time linux. Linux journal, 23.

[35] Kazanzides, P., Chen, Z., Deguet, A., Fischer, G. S., Taylor, R. H., DiMaio, S. P., "An Open-Source Research Kit for the da Vinci R Surgical System," in proceedings of the *IEEE International Conference on Robotics & Automation (ICRA)*, Hong Kong, pp. 6434–6439, 2014.

[36] Tokuda, J., Fischer, G. S., Papademetris, X., Yaniv, Z., Ibanez, L., Cheng, P., Liu, H., Blevins, J., Arata, J., Golby, A. J., Kapur, T., Pieper, S., Burdette, E. C., Fichtinger, G., Tempany, C. M., Hata, N., "OpenIGTLink: an open network protocol for image-guided therapy environment*," Int J Med Robot*, vol 5, issue 4, 2009. pp. 423–434.

[37] Tokuda, J., Fischer, G. S., Iordachita, I., Tauscher, S., Kazanzides, P. von Tiesenhausen, C., Burdette E. C., Tempany, C. M., "Roles of OpenIGTLink in Medical Robotics Research", in proc. of the *IEEE International Conference on Robotics & Automation (ICRA)*, Seattle, 2015.

[38] Eslami, S., Shang, W., Li, G., Patel, N., Fischer, G. S., Tokuda., J., Hata, N., Tempany C. M., Iordachita, I., "In-Bore Prostate Transperineal Interventions with an MRI-guided Parallel Manipulator: System Development and Preliminary Evaluation." *Int. J. Med. Robot. Comput. Assist. Surg.* (Accepted), 2015.

[39] Kim, M., Park, W. J., Lee, C., Kim, Y. O., Choi, S., Suh, Y. S., Yang, H. K., Kim, H. J., Kim, S., "A Development of Novel Vision System for Laparoscopic Surgical Robot System," in proceedings of the *IEEE International Conference on Robotics & Automation* (*ICRA*), Seattle, 2015.

[40] Nagy, D. Á., Takács, Á. , Haidegger T., Rudas, I. J., "The CALap System—A Low-Cost Lightweight Robotic Arm for Laparosopic Camera Handling," in proceedings of the *IEEE International Conference on Robotics & Automation (ICRA)*, Seattle, 2015.

[41] Lengyel, B., "Education of minimally invasive surgery in the digital age," Ph.D Dissertation, Semmelweis University, Budapest, Hungary, 2009.

[42] Chen, Z., Deguet, A., Vozar, S., Munawary, A., Fischery G., Kazanzides, P., "Interfacing the da Vinci Research Kit (dVRK) with the Robot Operating System (ROS)," in proceedings of the *IEEE International Conference on Robotics & Automation (ICRA)*, Seattle, 2015.

[43] National Aeronautics and Space Administration Goddard Space Flight Center, "Robotic refueling mission," March 2015. [Online]. Available: http://ssco.gsfc.nasa.gov/robotic refueling mission.html.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 35–46

Vertical Handover and Load Balancing Decision Algorithms for Heterogeneous Cellular-WLAN Networks

Zsolt Alfred POLGAR, Andrei Ciprian HOSU,
Zsuzsanna Ilona KISS, Mihaly VARGA

Department of Communications,
Technical University of Cluj Napoca, Cluj Napoca, Romania
e-mail: {Zsolt.Polgar; Andrei.Hosu; Zsuzsanna.Kiss; Mihaly.Varga}@com.utcluj.ro

Manuscript received July 17, 2015; revised November 18, 2015.

Abstract: Multi-access and heterogeneous wireless communications are one of the solutions for providing generalized mobility and improved user experience. This paper proposes Vertical Handover (VHO) and Load Balancing (LB) decision algorithms for heterogeneous network architectures which integrate cellular networks and Wireless Local Area Networks (WLANs). The cellular-WLAN VHO and LB decisions are taken based on parameters which characterize both the coverage and traffic load. Computer simulations performed in realistic scenarios show that the proposed VHO algorithm ensures better performance compared to "classical" ones and that the LB mechanism can significantly offload the congested cellular networks when WLAN connectivity is available.

Keywords: ubiquitous connectivity, vertical handover, heterogeneous networks, decision algorithm, network state information, load balancing.

1. Introduction

An important characteristic of Next Generation Networks will be the integration of heterogeneous wireless access technologies, which will lead to increased overall system efficiency and improved user experience. Several issues concerning service continuity and resource management are still unsolved and require further research.

The authors of [1] present a survey of existing technologies that support multimedia communications in a heterogeneous wireless network and the main requirements and solutions for mobility management are discussed. In [2] the

The results of this study were partially presented at the 5th International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics 2015.

authors review the emerging protocols and architectures aiming to support intersystem handovers and present an optimized handover framework built around the functionality introduced by the IEEE 802.21 standard.

In [3] a new architecture and a new network selection scheme that takes into account the resource usage and the user's preferences are proposed. The solution presented ensures the selection of the most suitable network for each flow while taking into account the QoS requirements of the services. In [4] the authors propose another solution for handover management which answers the user's requirements and ensures service continuity in 3G-WLAN, 3G-WMAN and WMAN-WLAN networks.

In [5] a novel MIHF (Media Independent Handover Function) based seamless inter-RAT (Radio Access Technology) handover algorithm is proposed for UMTS and WiMAX networks. This solution uses cross-layer techniques for providing lossless handover while keeping acceptable delays. Improvement possibilities of the inter-system handover mechanisms in the 3GPP Evolved Packet Core environment are studied in [6] while other VHO optimization mechanisms for 4G networks are proposed in [7].

Resource sharing and management in heterogeneous wireless networks involve complex operations with contradictory requirements, but in the same time can offer more efficient usage of the limited frequency bands. Many papers studied various aspects related to cooperation in heterogeneous networks. In [8] the authors propose a Cooperative Radio Resource Management (CRRM) solution between heterogeneous air-interfaces. Strategies for CRRM in coexisting WiMAX and HSDPA networks are developed in [9]. Scenarios with or without inter-system VHO were considered, showing that RRM combined with VHO maximizes the throughput.

In [10] the authors investigate the issue of parallel transmissions over multiple RATs, focusing their attention on the QoS perceived by the final users. A simple but effective CRRM algorithm is proposed and evaluated in 802.11a-UMTS and 802.16e systems scenario. In [11] it is proved that load balancing is a significant method to achieve resource sharing over heterogeneous wireless networks and to provide better services.

The concept of soft load balancing mechanisms was presented in [12], while in [13] the authors propose a Flow Diversion-based Vertical Handoff Algorithm relying on soft load balancing. The authors adopt a Fuzzy Neural Network to determine the optimal flow-dividing ratio in order to balance the network load.

Other interesting cooperation solutions over heterogeneous networks proposed in the technical literature are the following: in [14] the authors propose a Multi-Radio Cooperative Automatic Retransmission Request scheme, which combines long-range and short-range communications for retransmission of lost packets; in [15] the authors illustrate a way of implementing cooperation

mechanisms at IP Multimedia Subsystem level between networks that share the same IP core; in [16] the authors present a hierarchical architecture for load balancing based on the idea of grid in computer networks.

This paper proposes an improved VHO decision algorithm and a LB algorithm for 3G/4G – WLAN heterogeneous network architectures which offer support for advanced MIH mechanisms. The structure of the paper is the following: Section 2 introduces the system model, Section 3 presents the proposed VHO and LB algorithms, while Section 4 presents the evaluation methodology of the proposed algorithms. Section 5 presents the results obtained by computer simulations using the Network State Information (NSI) acquired by field measurements and Section 6 presents the main conclusions the paper.

2. System model

The system model considered, presented in *Fig. 1*, can be described as follows: it is given a heterogeneous network composed of 3G/4G cellular networks, providing large coverage, and public WLAN/WiFi networks having the role to offload the traffic passing through the 3G/4G networks. The coupling between the heterogeneous networks is implemented by specially designed gateways, the Service Continuity Gateways (SCG). The coupling infrastructure includes also a Central Database (CD) which stores link state and traffic related information for each wireless network. A specially designed server, the Connectivity Support Server (CSS), controls the access of the users to CD.

Two categories of users are considered:

- Individual users with mobile terminals (MT) equipped with one cellular and one WiFi interface. The MT runs the algorithms which implement the VHO between cellular and WiFi networks, aiming to maintain the service continuity.
- Mobile Routers (MR) installed in transportation vehicles and equipped with one or several cellular and WiFi interfaces. These MRs run the algorithms which can implement not only VHO between the cellular and WiFi networks but also load balancing operations which allow the joint usage of the transmission resources available in several networks. In an urban environment the transportation vehicles have low speed and frequent stops which makes possible the usage of WLANs for providing Internet access to the passengers.

It is supposed also that the mobile terminals and the MRs are equipped with GPS receivers, being capable to establish their geographical position and speed.

This paper proposes VHO and LB decision algorithms which allow selecting the best WLANs for transmission respectively distributing the data flows of the users on several WLANs, when the speed of the mobile terminal is low. The

NSI and traffic related information necessary for selecting the target networks for VHO and/or LB can be acquired from the CD of the coupling infrastructure. The algorithms implementing the routing in the heterogeneous network and the access to the CD are beyond the scope of this paper. See for details [17] [18].

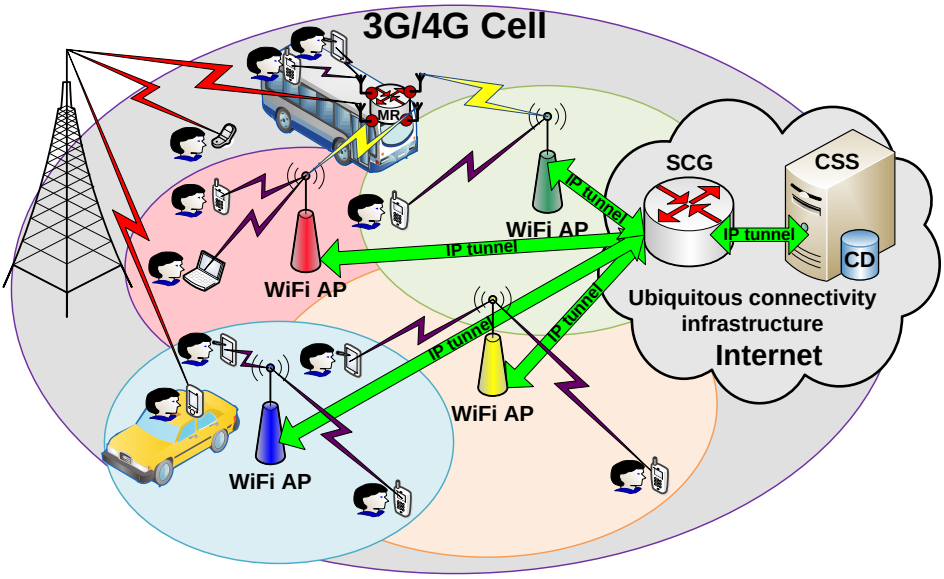

Figure 1: The system model of the considered heterogeneous network.

3. Vertical handover and load balancing algorithm for cellular-WLAN heterogeneous networks

The proposed algorithm which performs both cellular-WLAN and WLAN-WLAN VHO processes is based on the flowchart presented in *Fig. 2*. The algorithm is intended for individual users who do not perform load balancing between the cellular and WiFi networks. The users travelling with vehicular speed are connected to cellular networks while pedestrian users try to connect, if possible, to WLANs (see *Fig. 1*), the cost of the Mbyte transmitted in WLANs being significantly lower compared to 3G/4G networks.

The selection of the target WLAN is performed, by the user terminal, based on the Received Signal Strength (RSS) and Available Transmission Rate (ATR) parameters of the wireless link. The cost of the transmission taking place in the WLAN also can be considered.

The estimated ATR of the wireless link can be computed as:

$$ATR_{est} = f(RSS) \cdot (1 - BLER) \cdot (1 - ChBPF) \quad (1)$$

where $f(\text{RSS})$ gives the average bit rate of the WiFi connection as a function of the RSS; ChBPF – Channel Busy Period Fraction represents the fraction of the active time when the WiFi channel is used for transmission.

The WiFi link's Block Error Rate (BLER) can be computed based on the evaluation of the Signal to Interference and Noise Ratio (SINR) of the WiFi link or by counting the ACK/NACK messages received by the MAC layer.

The parameters used in target network selection have different measurement units, so normalization is a necessary step. The used max-min normalization is described by the following relation:

$$v_{ij} = (\mathrm{x}_{ij} - \min_i(\mathrm{x}_{ij})) / (\max_i(\mathrm{x}_{ij}) - \min_i(\mathrm{x}_{ij})) \quad (2)$$

where x_{ij} is the value of the j-th parameter in the i-th network and v_{ij} is the normalized value of x_{ij}.

In order to compare the different networks a utility function is defined:

$$C_i = \sum_{j=1}^{M} w_j v_{ij} \quad (3)$$

where M is the number of parameters and w_j is the weight of parameter j.

The VHO target network is the one with the highest value of the utility function. In order to compute the weights w_j a pair-wise comparison of all parameters should be performed using a pairwise comparison matrix **B**, with dimension MxM, whose elements are the b_{ij} comparisons between the i-th and j-th parameter. In order to build this matrix it is needed to indicate how many times more important or dominant one element is over another [19].

Finally the weight vector **w** can be computed by solving the equation [20]:

$$(\mathbf{B} - \lambda \mathbf{I}) \cdot \mathbf{w} = 0 \quad (4)$$

where λ is the eigenvalue of **B** and **I** is the identity matrix. The weight vector $\mathbf{w} = [w_1, ..., w_M]^T$ is the eigenvector of **B** corresponding to eigenvalue λ_{max}.

The LB algorithm was designed for Mobile Routers equipped with 3G/4G interfaces and several WiFi interfaces. This algorithm is an extension of the VHO algorithm presented in *Fig. 2*, more exactly the 3G-WLAN and WLAN-WLAN VHO process (see the shaded box in *Fig. 2*) is replaced by a LB process, performed according to the flowchart presented in *Fig. 3*. The LB process starts when the speed of the vehicle drops below an imposed limit (close to pedestrian speed) and it distributes some of the flows passing through the 3G/4G link (or links) on several WiFi links. The selection of the target WLANs involved in the LB process is performed based on the utility functions

and if significant changes of the RSS or of the ATR parameter are detected on any WiFi link, then the LB process is restarted and all flows are reassigned.

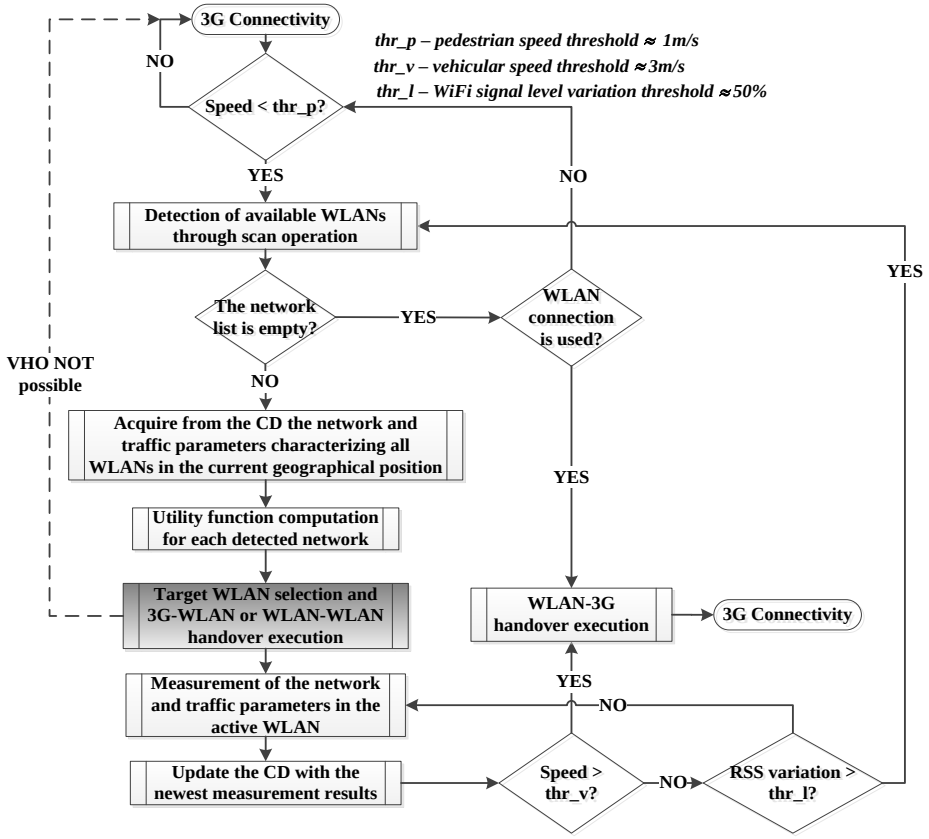

Figure 2: The proposed cellular-WLAN and WLAN-WLAN VHO algorithm.

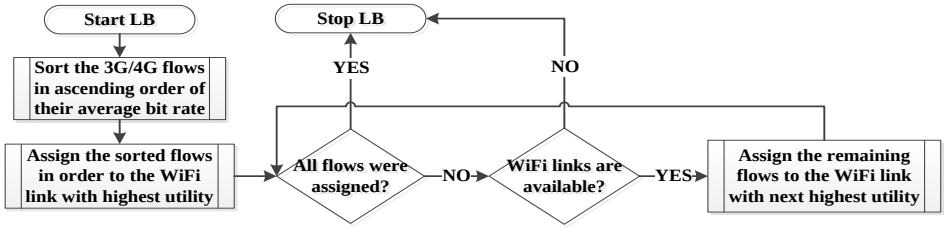

Figure 3: The proposed LB algorithm for cellular-WLAN heterogeneous networks.

4. The evaluation methodology and the test scenarios

In order to evaluate the performance of the proposed VHO and LB mechanisms the following methodology was used: using a real test site the RSS and the ATR parameters of the WLANs composing the heterogeneous network were acquired and the data obtained were fed into a system level simulator, developed in the UCONNECT FP7 project. The simulation performed replicates a real scenario in which a MT or a MR is moving in the coverage area of several WLANs and VHO and LB operations are taking place according to the proposed algorithms. The WLAN test network, located in a university campus, includes 5 WiFi (802.11g) Access Points (APs), and it is presented in *Fig. 4*. The RSS variations experienced by the mobile terminal during its journey are depicted also in *Fig. 4*. By assigning to each WiFi AP non or partially overlapping channels the estimated BLER parameter was kept smaller than 0.1. The background traffic (generated using the *Iperf* tool), for each of the five APs, is presented in *Table 1* for two scenarios. The ChBPF parameter was measured for each AP in different ranges of the RSS parameter (see *Fig. 4*).

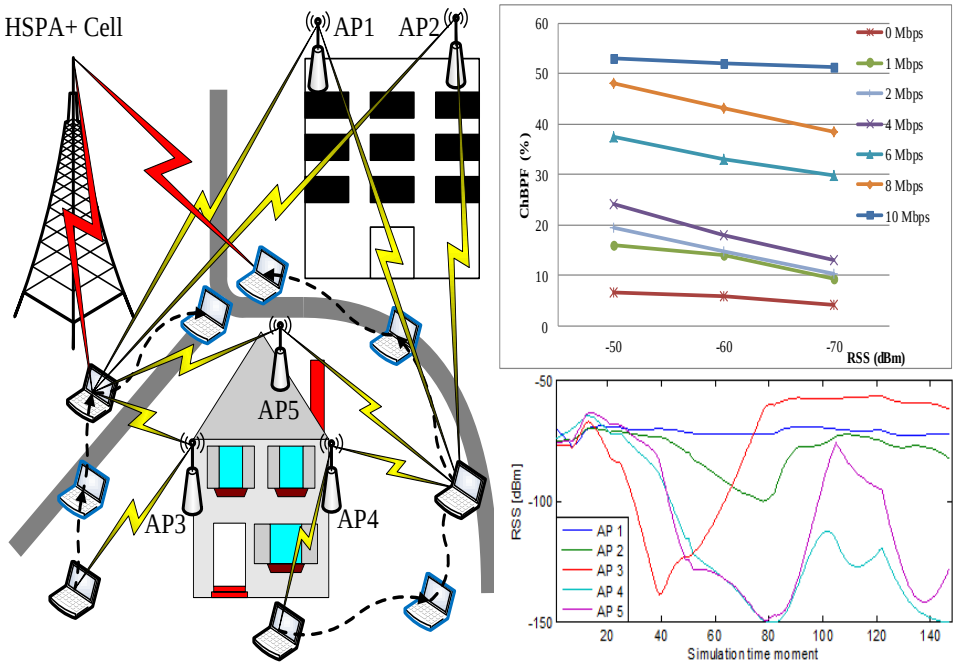

Figure 4: 3G-WLAN and WLAN-WLAN VHO test scenario. ChBPF versus RSS for different traffic values. RSS variation along the route of the monitoring terminal.

5. Simulation results

One of the targets of this study was to evaluate the performance of various VHO target network selection mechanisms. We considered, besides the RSS and ATR based decisions, the decision process based on the utility function, which was computed for each AP using relation (3). The weights of the parameters considered were assigned empirically according to *Table 2*.

Table 1: Background traffic passing through APs and the cost of the transmitted MByte.

AP	Scenario 1	Scenario 2	Cost of MByte
AP1	6Mbps	1Mbps	0.1
AP2	8Mbps	2Mbps	0.2
AP3	10Mbps	10Mbps	0.15
AP4	1Mbps	8Mbps	0.25
AP5	0Mbps	0Mbps	0.3

Table 2: Weights considered in the different test scenarios.

Weighting method index	Algorithm	Weights		
		ATR	RSS	Cost
1	ATR based	1	0	0
	RSS based	0	1	0
	ATR+RSS based	0.4	0.6	0
	ATR+RSS+Cost based	0.3	0.6	0.1
2	ATR based	1	0	0
	RSS based	0	1	0
	ATR+RSS based	0.5	0.5	0
	ATR+RSS+Cost based	0.33	0.33	0.33
3	ATR based	1	0	0
	RSS based	0	1	0
	ATR+RSS based	0.6	0.4	0
	ATR+RSS+Cost based	0.6	0.3	0.1
4	ATR based	1	0	0
	RSS based	0	1	0
	ATR+RSS based	0.55	0.45	0
	ATR+RSS+Cost based	0.45	0.35	0.2

In *Fig. 5* and *Fig. 6* the achievable average transfer rate is presented for Scenario 1 and 2 (see *Table 1*) for the decision and weighting methods considered. These results show that the ATR based decision offers the best performance and the RSS based decision the worst one, but the measurement precision of ATR is lower, thus the decision process cannot consider only the ATR parameter. One solution is to select the VHO target network based on the utility function. The combined usage of the RSS and ATR parameters reduces

the influence of the ATR measurement imprecision, while keeping the average achievable tranfer rate approximately the same. One can also notice that the decision which takes into consideration the cost of the networks has lower performance than the algorithms which neglect the cost parameter, because in this case the network with the highest achievable rate is not always selected.

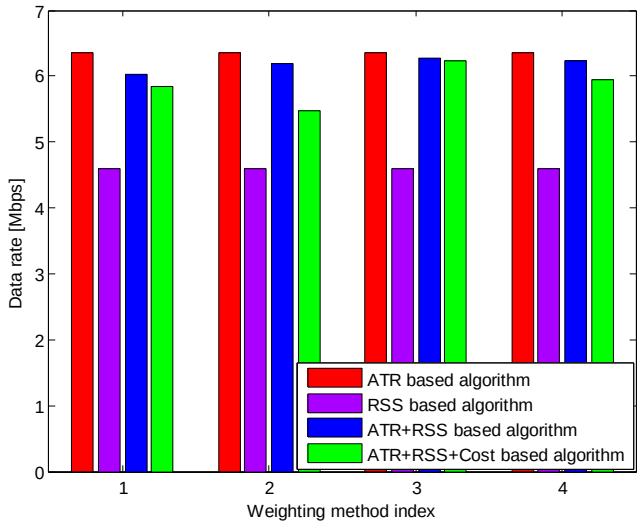

Figure 5: Average achievable transfer rate obtained in the case of Scenario 1.

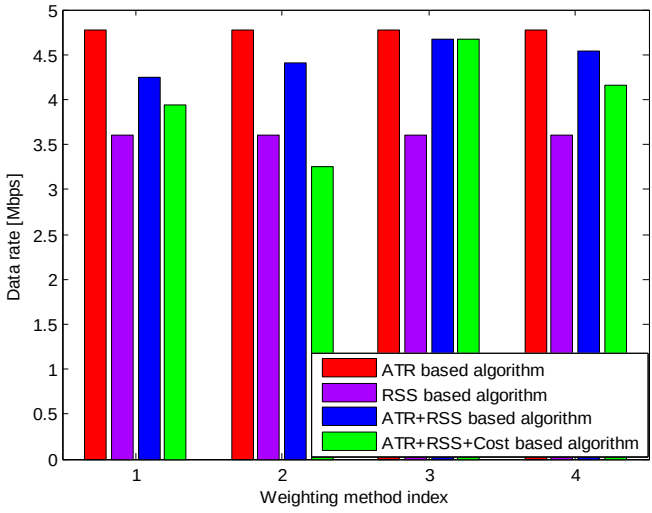

Figure 6: Average achievable transfer rate obtained in the case of Scenario 2.

Another target of this study was to evaluate how much a 3G+ link can be offloaded when using several WiFi links and the proposed LB algorithm. This test scenario involves a MR equipped with a 3G+ interface and with several WiFi 802.11g interfaces. The measured average downlink ATR of the 3G+ link was ≈ 11Mbps (see *Fig. 7*). The MR is moving on a path covered by the 5 WiFi APs presented in *Fig. 4* and having the background traffic given in *Table 1* (Scenario 1). The test flows are represented by 10 video streaming flows with 800kbps average rate and 10 web radio flows with 128kbps average rate. All these test flows can be carried by the HSPA+ interface, with an average usage ratio of ≈ 90% (usage ratio of the available capacity, i.e. the ATR). The selection of the target WLANs is performed using the RSS and ATR parameters and equal weights, i.e. weighting method 2 in *Table 2*.

In the considered scenarios a single WiFi can completely offload the 3G+ link if all the available transmission resources of the WLANs can be used by the MR. However, in a real situation the WLAN resources allocated to a user/MR are limited in order to ensure fairness between the users. *Fig. 8* gives the average usage of the 3G+ link if 2, 3 respectively 4 WiFi interfaces are used and the transmission rate is limited to 6% respectively to 3% of the 802.11g WiFi link rate (i.e. 54Mbps).

Figure 7: 3G+ link ATR parameter measurement.

The results presented in *Fig. 8* show that a LB process over 2 – 3 WiFi 802.11g networks, having "normal" background traffic can offload significantly the 3G+ link of a MR, while reducing the network usage cost.

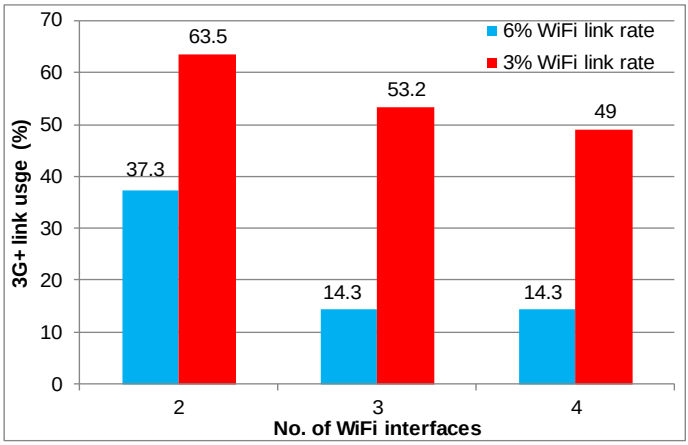

Figure 8: 3G+ link usage in the considered LB scenarios.

6. Conclusion

The paper proposes a VHO and a LB decision algorithm for cellular-WLAN heterogeneous networks. The proposed decision algorithms act based on the NSI and traffic information stored in the central database of the network architecture. Both algorithms aim at offloading the cellular networks by using, when possible, public WLANs. The LB algorithm represents an extension of the VHO algorithm and it is intended for mobile routers. Computer simulations performed in a scenario involving a 3G+ network, 5 WiFi APs and real NSI and traffic information show that the proposed VHO decision algorithm performs better compared to "classical" ones. The simulations performed highlight also the potential of the LB algorithm in offloading the 3G/4G links if several WLANs are available and the MR is equipped with 2 – 3 WiFi interfaces.

Acknowledgements

This research has received funding from the European Union's Seventh Framework Programme under grant agreement nº FP7-SME-2012-315161.

References

[1] Macriga, A., Kumar, A., "Mobility Management for Seamless Information flow in Heterogeneous Networks Using Hybrid Handover," *IJCSNS International Journal of Computer Science and Network Security*, vol. 10, no. 2, pp. 61–70, February 2010.

[2] Sarakis, L., Kormentzas, G., Guirao, F., M., "Seamless Service Provision for Multi Heterogeneous Access," *IEEE Wireless Comm.*, vol. 16, no. 5, pp. 32–39, October 2009.

[3] Zarai, F., Smaoui, I. ,Bonnin, J-M., Kamoun, L., "Seamless Mobility in Heterogeneous Networks," *Int. Journal of Next-Generation Networks*, vol. 2, no. 4, pp. 12–31, Dec. 2010.

[4] Kassar, M., Kervella, B., Pujolle, G., "An Intelligent Handover Management System for Future Generation Wireless Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 791691.

[5] Liu, B., Martins, Ph., Bertin, Ph., "Cross-Layer Design of the Inter-RAT Handover between UMTS and WiMAX," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, Article ID 763614.

[6] Frei, S., Fuhrmann, W., Rinkel, A., Ghita, B., V., "Improvements to Inter System Handover in the EPC Environment," in *Proc. of 4th IFIP International Conference on New Technologies, Mobility and Security NTMS 2011*, Paris, France, February 2011, pp. 1–5.

[7] Solouk, V., Ali, B., M., Wong, K., D., "Vertical Fast Handoff in Integrated WLAN and UMTS Networks," in *Proc. of 7th International Conference on Wireless and Mobile Communications ICWMC 2011*, Luxembourg, June 2011, pp. 59–64.

[8] Lot, M., Sdralia, V., Pishella, M., et al., "Cooperation of 4G Radio Networks with Legacy Systems," in *Proc. of 14th IST Mobile & Wireless Communications Summit*, Dresden, Germany, June 2005.

[9] Sartori, L., Elayoubi, S-E., Fourestie, B., Nouir, Z., "On the WiMAX and HSDPA coexistence," in *Proc. of IEEE ICC 2007*, Glasgow, UK, June 2007.

[10] Bazzi, A., Pasolini, G., Andrisano, O., "Multiradio Resource Management: Parallel Transmission for Higher Throughput?," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 763264.

[11] Macriga, G., A., Surya, V., S., "Location Management and Resource Allocation Using Load Balancing in Wireless Heterogeneous Networks," *Advances in Computer Science and Information Technology Networks and Communications*, vol. 84, pp. 383–393, 2012.

[12] Son, H., Lee, S., Kim, S. C., Y., Shin, S., "Soft Load Balancing over Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no.4, pp. 2632–2638, July 2008.

[13] Wang, N., Shi, W., S., Fai, Liu, Y., "Flow Diversion-based Vertical Handoff Algorithm for Heterogeneous Wireless Networks", *Journal of Computational Information Systems*, vol. 7, no. 13, pp. 4863–4870, 2011.

[14] Alonso-Zárate, J., Kartsakli, E., Alonso, L., Katz M., Verikoukis, C., "Multi-Radio Cooperative ARQ in Wireless Cellular Networks: A MAC Layer Perspective", *Telecommunications Systems*, vol. 52, no. 2, pp. 375–385, February 2013.

[15] Chen, C.-Y., Chang, K.-D., Chao H.-C., Kuo, S.-Y., "Ubiquitous IMS Emergency Services over Cooperative Heterogeneous Networks," in *Proc. of 2009 International Conference on Wireless Comm. and Mobile Computing*, Leipzig, Germany, June 2009.

[16] Shi, W., Li, B., Li, N., Xia, C., "A Network Architecture for Load Balancing of Heterogeneous Wireless Networks," *Journal of Networks*, vol. 6, no. 4, pp. 623–630, April 2011.

[17] Polgar, Z., A., Rus, A., B., Kiss, Z., I., Consoli, A., Ayadi, J., Egido, M., "Ubiquitous Connectivity Platform for Public Transport Communication Services," in *Proc. of Future Network & Mobile Summit 2013*, Lisboa, Portugal, July 2013, pp. 1–8.

[18] Hosu, A., C., Kiss, Z., I., Ivanciu, I., A., Polgar, Z., A., Consoli, A., Egido, M., "Ubiquitous Connectivity Platform for Intelligent Public Transportation Systems," in *Proc. of 10th ITS European Congress*, Helsinki, Finland, June 2014.

[19] Saaty, T., L., "Decision making with the analytic hierarchy process," *International Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, 2008.

[20] Wang, L., Kuo, G., S., "Mathematical Modelling for Network Selection in Heterogeneous Wireless Networks - A Tutorial," *IEEE Surveys and Tutorials*, vol. 15, no. 1, pp. 271–292, February 2013.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 47–58

Estimating the Level of Conflict Based on Audio Information Using Inverse Distance Weighting

Gábor GOSZTOLYA

MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary,
e-mail: ggabor@inf.u-szeged.hu

Manuscript received July 30, 2015; revised October 19, 2015.

Abstract: In recent years it has become possible to extract non-trivial information from audio sources. One such task is to determine the intensity of conflicts arising in speech recordings, based solely on audio information sources. This intensity is expressed as a real number, therefore this task is essentially a regression one, the objective being to estimate a given numeric score. As the number of examples in these tasks are limited, a kNN-like solution may work well in these problems. Such an approach is the Inverse Distance Weighting (IDW) algorithm, which is also a suitable choice as it is computationally cheap. By applying this method on the conflict intensity estimation task using the SSPNet Conflict Corpus, we were able to reach the level of performance of baseline SVM.

Keywords: speech technology, conflict detection, regression, KNN, inverse distance weighting.

1. Introduction

In the past, within the field of speech technology, most of the researchers' efforts were devoted to speech recognition. But in recent years they have turned their attention to other areas as well like emotion detection [25, 10], speaker verification [17], speaker age estimation [5], detecting social signals like laughter and filler events [1, 10, 12], and estimating the amount of physical or cognitive load during speaking [20, 11, 14]. What these tasks have in common is that what is considered noise in speech recognition (i.e. non-verbal audio information) becomes important, while what was relevant in speech recognition (i.e. what the speaker actually said) becomes irrelevant.

Such a task is to determine the level of conflict from the audio. Conflicts influence the everyday lives of people to a significant extent, either in their public or personal lives, and they are one of the main causes of stress [23]. With

the rise of socially intelligent technologies, the automatic detection of conflicts can be the first step of handling them properly.

In this study we focus on the automatic estimation of the level of conflict in televised political debates. This is mainly a regression task [2], i.e. we have to match a score as closely as possible, as the level of conflict is expressed as one numerical value. Of course, from an application point of view, a categorical approach looks more practical, where the question is whether there a conflict present or not, and if so, we want to know what its level is. This in fact means that the task is turned into a classification one [6]. However, this categorization may be readily performed by setting up intervals for the conflict score; therefore we approached this task mainly from a regression point of view.

Although such recordings can be obtained quite easily, their annotation can be rather expensive; hence it is preferable to use a machine learning method that works well for small-sized training sets. One such algorithm for classification is the K Nearest Neighbours method (*kNN*), where the label of the given utterance to be classified is determined by simple majority voting of its *K* nearest neighbours. Of course, the distance function used and the value of *K* have to be determined, but these are not major requirements (especially when compared to the parameters of other machine learning methods like Artificial Neural Networks (ANNs) [3] and Support Vector Machines (SVM) [19, 24]).

Another advantage of this method is its low computational cost if both the train and test sets consist of just a small number (e.g. hundreds) of examples − especially when compared to high-complexity approaches like SVM and AdaBoost [18, 4]. A similar approach for regression is Inverse Distance Weighting (IDW) [22]. In it, the function value of a given point is calculated by computing the weighted sum of the function value of the training points, where the weight of a training point is inversely proportional to its distance from the point to be evaluated.

The structure of this paper is as follows. First, we describe the audio corpus used for conflict intensity estimation, and the evaluation methodologies. Then we describe the original and an improved version of the IDW algorithm. After, we explain the slight modifications made that we felt necessary to use IDW for this task. Then we present and analyse our results got from applying them on the development and test sets. Lastly, we draw some conclusions and make some suggestions for future study.

2. The SSPNet Conflict Corpus

We performed our experiments on the (freely accessible) SSPNet Conflict Corpus [15]. It contains recordings of Swiss French political debates taken from the TV channel “Canal9”. It consists of 1430 recordings, 30 seconds each,

making a total of 11 hours and 55 minutes. The ground truth level of conflicts was determined by manual annotation performed by volunteers not understanding French (French-speaking people were excluded from the list of annotators). Each 30-second long clip was tagged by 10 annotators, and in the end we got a score in the range [-10, 10], 10 meaning a high level of conflict and -10 meaning no conflict at all. The data was later used in the Conflict sub-challenge of the Interspeech 2013 ComParE Challenge [21].

The database contains both audio and video recordings, and the annotators were able to rely on both sources. In the latter experiments, however, attention was focused only on the audio information for a number of reasons. Firstly, the annotators judged the level of conflict in a similar way based on the two sources: the correlation of the scores was 0.95 [15]. Furthermore, in a television political debate, audio can be a more reliable indicator: the subjects can hear all the participants, but they can only see the one that the cameraman of the debate has chosen, which is not the one speaking in many cases (especially in the heat of a debate when several persons may be speaking at the same time).

3. Inverse Distance Weighting

Inverse Distance Weighting (IDW) was introduced by Shepard in 1968, originally for interpolating surfaces from irregularly-spaced data [22]. Later it was used for other interpolation tasks as well [9, 26]. This method (sometimes called "Shepard's algorithm") estimates the target score of a given point by the weighted sum of the input scores, and the weight of a training point is inversely proportional to its distance. Given a set of sample points $x_1, \ldots, x_N$, score values $f_1, \ldots, f_N$ and a distance function $d(x,y)$, for a point $y \neq x_i$, its score $F(y)$ will be

$$F(y) = \sum_{i=1}^{N} w_i f_i, \tag{1}$$

where w_i is the weight of the ith input point. It is defined by

$$w_i = \frac{d(x_i, y)^{-c}}{\sum_{j=1}^{N} d(x_j, y)^{-c}}, \tag{2}$$

where $c > 0$. Inserting this into Eq. (1) we get

$$F(y) = \frac{\sum_{i=1}^{N} d(x_i, y)^{-c} f_i}{\sum_{i=1}^{N} d(x_i, y)^{-c}}. \quad (3)$$

The value of c regulates the relative importance of closer and more distant points: for larger values of c, the closer points are more important, while using smaller values of c tends to equalize the weights. It is a global method in the sense that to determine the score of a test example, all training points are used, no matter how far away they are. A simple extension to make this method local was suggested by Franke and Nielson [8], who introduced the limiting parameter R. Their formula for determining the weights is

$$w_i = \frac{\left(\frac{(R - d(x_i, y))_+}{Rd(x_i, y)} \right)^c}{\sum_{j=1}^{N} \left(\frac{(R - d(x_j, y))_+}{Rd(x_j, y)} \right)^c}, \quad (4)$$

where $(v)_+$ denotes $\max(v, 0)$.

4. Experimental setup

Speech recognition usually decomposes the speech signal of an utterance into small-equal sized parts (*frames*), from which it is easy to extract the same number of features for machine learning. In the current task, however, we have to estimate the level of conflict for the whole 30 second-long utterance, therefore features which describe the whole recording are preferred. A straightforward choice is to compute the standard features (e.g. MFCC and filter banks) for each frame, then calculate the minimum, maximum, mean and standard deviation of these values.

In our experiments we used the feature set introduced in [21]. It contained 6373 features overall, extracted by using the tool openSMILE [7]. The set includes energy, spectral, cepstral (MFCC) and voicing-related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. Of course, as this is a quite general feature set, not all attributes are useful for our current task; now, however, we focused on the application of IDW, and did not experiment with any kind of feature selection.

Following standard machine learning practice, the available data was split into training, development and test sets. The first one was used for training purposes, i.e. IDW estimation was done using the points belonging to this set. The development set was used to find the meta-parameters of the learning algorithm, i.e. *c* and *R* by choosing the values which led to the best results by training on the training set and evaluating on the development one. Next, using the "optimal" *c* and *R* values, we evaluated our model on the test set; in this case we used the points of both the training and development sets as training points. We used the division described in [21], so 793 recordings were used for model training, whereas 240 and 397 were used for the development and test sets, respectively.

A straightforward choice for measuring the similarity of the reference and the estimated values is cross-correlation. For the signals $X = x_1, \ldots, x_n$, and $Y = y_1, \ldots, y_n$, it is defined as

$$CC(x,y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_X\sigma_Y}, \tag{5}$$

where μ_X and μ_Y are the mean and σ_X and σ_Y are the standard deviation values of *X* and *Y*, respectively. Another choice for measuring the difference between the two series is the Root-Mean-Square Error (RMSE), defined as

$$RMSE(x,y) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}. \tag{6}$$

While cross-correlation measures the tendency of the two signals, RMSE measures the actual difference between the values; this means that in a regression task it may be sensitive to the scaling of results.

Another possibility is to turn this task into a classification one. We also carried out experiments for this, following the setup described in [21], where non-negative conflict scores were considered as *high* ones, while negative ones were converted into the class label *low*. Methods applied on such two-class classification problems can be measured by a number of metrics, all of which are based on the values of the confusion matrix. There, T_P will be the number of true positives (i.e. the occurrences of class *high* that were classified correctly) and F_P the number of false positives (the *low* occurrences classified as *high*), while the values T_N (true negatives) and F_N (false negatives) are defined in a similar way. (The sum of the four values will be *n*.) Then accuracy will simply be the ratio of correctly classified examples, i.e.

$$Accuracy = \frac{T_P + T_N}{n}. \tag{7}$$

If we treat our task as an information retrieval one, meaning that we are interested in the detection of occurrences of the positive class only (in our case, class *high*), we can measure our performance by means of precision and recall. Precision measures how many of the identified examples actually belonged to this class, i.e.

$$\mathrm{Pr}\,ecision = \frac{T_P}{T_P + F_P}, \tag{8}$$

whereas recall expresses how many of the examples actually belonging to the positive class were found; i.e.

$$\mathrm{Re}\,call = \frac{T_P}{T_P + F_N}. \tag{9}$$

As there is clearly a tradeoff between these two scores, they are usually aggregated via F-measure (or F_1-score), defined as the harmonic mean of the two values, i.e.

$$F_1 = \frac{2 \cdot \mathrm{Pr}\,ecision \cdot \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} = \frac{2 \cdot T_P}{2 \cdot T_P + F_N + F_P}. \tag{10}$$

Using the concept of recall, we can define another variant of accuracy, namely the Unweighted Average Recall (UAR) or True Positive Rate (TPR), expressed as the mean of the recall values for all the classes. In a two-class set-up it is equal to

$$UAR = \frac{\frac{T_P}{T_P + F_N} + \frac{T_N}{T_N + F_P}}{2}. \tag{11}$$

Accuracy is sensitive to class distribution, whereas UAR can be viewed as an accuracy which is balanced class-wise. For this task and this dataset in the past, regression metrics (especially cross-correlation) were used [15], and we also find this approach more logical, so we will follow this in our study. However, we will also view the task as a classification one, where we will primarily rely on the UAR score, just as it was common in some earlier studies on this dataset [21, 16].

5. Applying IDW for estimating the conflict scores

Shepard's algorithm and Franke's modified version were developed for generating surfaces based on sparsely distributed input points in a two-dimensional space and a function value. In a large-dimension regression task they might require some minor changes in order to perform well (and in our case there were 6373 features). To achieve this, we included some minor pre-processing and post-processing steps, which we will now describe in detail.

First, we used the Euclidean distance metric; that is, for two points $y = y_1, y_2, \ldots, y_n$ and $z = z_1, z_2, \ldots, z_n$, their distance $d(y,z)$ was simply

$$d(y,z) = \sqrt{\sum_{i=1}^{n} (y_i - z_i)^2} \tag{12}$$

and in our preliminary tests we found that applying other distance functions yielded somewhat worse results. To prevent confusion caused by differently-scaled features (where a few of them might dominate the distance, whereas other, perhaps more important attributes might simply be ignored because of initial scaling), feature normalization was clearly required. For this reason, first all the vectors were normalized so that they had a standard deviation of 1. A couple of features had a standard deviation of 0, which were discarded, but this step clearly did not lead to any information loss (as it meant that the value of these features was the same for all examples).

After performing the IDW procedure, the resulting values were quite small compared to the real ones, perhaps because of the high dimensionality of the input data. To handle this issue, the resulting scores were also normalized: the mean was set to zero, and they were multiplied by a factor such that the standard deviation of the results became equal to the one of the scores of the training set. Next, scores falling below or above the limits of the scores of the training set were set to the minimum or maximum score, respectively, and each value was rounded to one decimal place.

Franke's method has two parameters, namely c and the limit value R. As for the latter, we decided to express it via the function of $maxd = \max(d(x,y))$ for all possible values of x and y (of the training set); that is, $R = r \cdot maxd$. Eventually when $r = 1$, all the training points were considered, whereas for lower values the more distant points were ignored. When no training points were found in the R-sized neighbourhood, the conflict score of the closest training point was used. We optimized the parameters cross-correlation, for UAR and for F-measure; we used linear SVM in regression (the SMOReg method in Weka [13]) mode as the baseline. (Note that this method was used as the baseline approach for ComParE

2013 [21]; the only difference is that the *c* parameter was tuned to maximize UAR, while we optimized CC as well).

5. Results

The results when optimizing for cross-correlation can be seen in Table 1.

Table 1: Scores obtained by optimizing for cross-correlation.

Method		CC	RMSE	Acc.	UAR	F_1
dev	IDW, c = 13.56	0.805	2.390	80.83%	80.67%	79.28%
	IDW, r = 0.15	0.816	2.314	81.67%	81.46%	80.00%
	SVM	0.828	2.427	74.58%	73.40%	66.30%
test	IDW, c = 13.56	0.782	2.654	80.60%	80.47%	77.94%
	IDW, r = 0.15	0.768	2.727	79.35%	79.23%	76.57%
	SVM	0.804	2.414	83.63%	82.35%	79.37%

Here, IDW achieved practically the same level of performance as SVM for all the metrics on the development set; Franke's method was somewhat better than the basic IDW algorithm. On the test set, however, the standard IDW method proved to be more stable, and Franke's variation (case $r = 0.15$) showed signs of overfitting. Shepard's method performed slightly worse than the baseline SVM, but the difference is not that big.

Table 2: Scores obtained by optimizing for UAR.

Method		CC	RMSE	Acc.	UAR	F_1
dev	IDW, c = 7.22	0.801	2.430	82.08%	81.95%	80.72%
	IDW, r = 0.69	0.808	2.383	82.50%	82.39%	81.25%
	SVM	0.806	2.330	80.42%	79.55%	75.65%
test	IDW, c = 7.22	0.775	2.702	79.09%	79.29%	76.88%
	IDW, r = 0.69	0.765	2.725	80.86%	80.55%	77.91%
	SVM	0.826	2.271	84.64%	83.87%	81.46%

Upon examining the classification results (see Table 2), it can be seen that the IDW classification performance significantly exceeded that of SVM for the

development set in its basic form, and using the variation developed by Franke and Nielsen (case r = 0.69) even surpassed this. (This variant also performed better judging from the regression scores.) However, on the test set the best variation with r = 0.69 performed slightly worse than the baseline SVM, although the difference is again not that big. Still, in our opinion even matching the score of the SVM is a good result for an algorithm that has such low computational requirements as IDW.

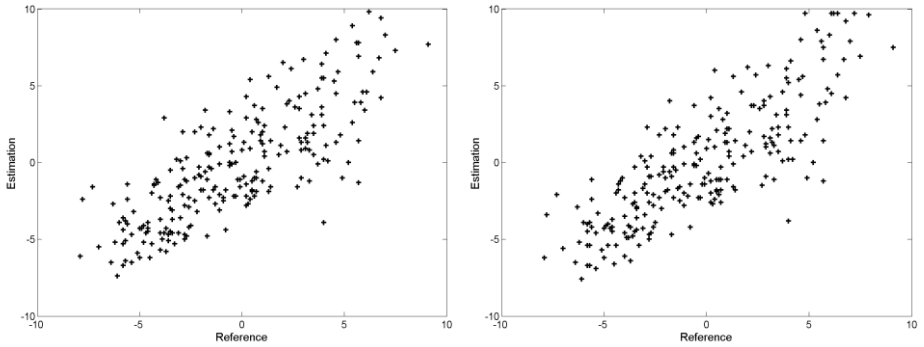

Figure 1: The estimated scores got as a function of the reference values, using IDW optimized for cross-correlation; Shephard's (left) and Franke's (right) methods.

Fig. 1 shows the regression scores in the function of the reference scores for the development set, obtained using the IDW algorithm with c = 13.56 (Shephard's method, left) and with $r = 0.15$ and $c = 7.68$ (Franke's algorithm, right). The strong correlation between the two values can clearly be seen; overall, the points produced by Franke's method seem a bit more packed, which is confirmed by both the higher CC and lower RMSE scores. It is understandable, though, as in this case we had one more parameter to set.

Fig. 2 shows the corresponding scores we got with the value c = 7.22 (Shephard's method, left) and with $r = 0.69$ and $c = 5.44$ (Franke's algorithm, right). This time we optimized for the UAR score, which is reflected in the lower cross-correlation value, resulting in somewhat more scattered points. The reason for this is that UAR only measures which point falls into which quarter of the chart (i.e. both the reference and the estimated scores are non-negative, both are negative, etc.), while the actual difference between the expected and the estimated scores is completely ignored.

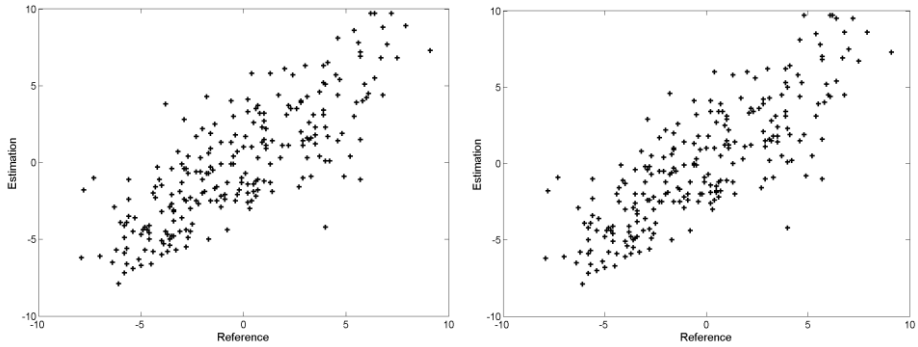

Figure 2: The estimated scores got as a function of the reference values, using IDW optimized for UAR; Shephard's (left) and Franke's (right) methods.

An interesting observation is that the optimal *c* values for Shepard's method were somewhat higher (13.56 and 7.2) than those of Franke's algorithm (7.68 and 5.44). This might be because for such a regression task training points which fall closer should be more important than those further away; this can be realized in the basic IDW method by using high values of c. When using the version developed by Franke and Nielsen, however, we can simply do this by choosing the right R value; then *c* can be set to a lower value as well.

Finally we should note that there were higher accuracy scores among the participants of ComParE 2013. (Although the cross-correlation scores were not always reported, since the official metric of the Challenge was UAR even for this regression task.) The more successful attempts, however, performed some kind of feature selection [16] or extracted new features from the utterances [12], while in this study we applied a different machine learning method for the regression task of conflict score estimation. Of course, it could be beneficial to use some kind of feature selection method for IDW as well, but this is clearly the subject of future work.

6. Conclusions

Regression tasks are quite rare in speech technology, but one exception is the detection of the intensity of conflicts based on speech recordings. We applied the Inverse Distance Weighting method to this task, which was originally developed for estimating surfaces on the basis of just a few sparsely and unevenly distributed reference points. After making a few minor alterations, this method outperformed the baseline SVM in terms of classification accuracy, and gave only slightly worse results in terms of regression scores. Taking into account the fact that IDW has low computational requirements and we can add

further training points without having to retrain a complicated model, we think that this method is a valid tool for conflict intensity estimation in particular, and speech technology regression tasks in general.

Acknowledgements

This publication is supported by the European Union and co-funded by the European Social Fund. Project title: Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073.

References

[1] Beke, A., Neuberger, T. "Automatic laughter detection in Hungarian spontaneous speech using GMM/ANN hybrid method," in *Proc. SJUSK, Copenhagen, Denmark*, 2013

[2] Berk, R. A. "Statistical Learning from a Regression Perspective," Springer Verlag, 2008.

[3] Bishop, M. C., "Neural Networks for Pattern Recognition," Clarendon Press, Oxford, 1995.

[4] Busa-Fekete, R., Kégl, B., "Accelerating AdaBoost using UCB," in *Proc. KDDCup 2009 (JMLR W&CP), Paris, France*, 2009, pp. 111–122.

[5] Dobry, G., Hecht, R. M., Avigal, M., Zigel, Y., "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 7, pp. 1975–1985, 2011.

[6] Duda, R. O., Hart, P. E., Stork, D. G. "Pattern Classification", John Wiley & Sons, 2001.

[7] Eyben, F., Wöllmer, M., Schuller, B., "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia, Firenze, Italy*, 2010, pp. 1459–1462

[8] Franke, R., Nielson, G., "Smooth interpolation of large sets of scattered data," *Int. Jour. for Num. Meth. in Eng.*, vol. 15, pp. 1691–1704, 1980.

[9] Gemmer, M., Becker, S., Jiang, T., "Observed monthly precipitation trends in China 1951–2002," *Theor. and Appl. Climat.*, vol. 77, no. 1, pp. 39–45, 2004.

[10] Gosztolya, G., Busa-Fekete., R., Tóth, L., "Detecting autism, emotions and social signals using AdaBoost," in *Proc. Interspeech, Lyon, France*, 2013, pp. 220–224

[11] Gosztolya, G., Grósz, T., Busa-Fekete., R., Tóth, L., "Detecting the intensity of cognitive and physical load using AdaBoost and Deep Rectifier Neural Networks," in *Proc. Interspeech, Singapore, Singapore*, 2014, pp. 452–456

[12] Grézes, F., Richards, J., Rosenberg, A., "Let me finish: Automatic conflict detection using speaker overlap," in *Proc. Interspeech, Lyon, France*, 2014, pp. 200–204

[13] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[14] Kaya, H., Özkaptan, T., Salah, A. A., Gürgen S. F., "Canonical Correlation Analysis and Local Fisher Discriminant Analysis based multi-view acoustic feature reduction for physical load prediction," in *Proc. Interspeech, Singapore, Singapore*, 2014, pp. 442–446

[15] Kim, S., Valente, F., Filippone, M., Vinciarelli, A., "Predicting continuous conflict perception with Bayesian Gaussian Processes," *IEEE Trans. Aff. Comp.*, vol. 5, no. 2, pp. 187–200, May. 2014.

[16] Räsänen, O., Pohjalainen, J., "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *Proc. Interspeech, Lyon, France*, 2013, pp. 210–214

[17] Reynolds, D. A., Quatieri, T. F., Dunn, R. B., "Speaker verification using adapted Gaussian Mixture Models," *Dig. Sign. Proc.*, vol. 10, no. 1, pp. 19–41, 2000.

[18] Schapire, R. E., Singer, Y., "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, 1999.

[19] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C., "Estimating the support of a high-dimensional distribution," *Neur. Comp.*, vol. 13, no. 7., pp. 1443–1471, 2001.

[20] Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y.: "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & Physical load," in *Proc. Interspeech, Singapore, Singapore*, 2014, pp. 427–431.

[21] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech, Lyon, France*, 2013, pp. 148–152.

[22] Shepard, D., "A two-dimensional interpolation function for irregularly-spaced data," in *Proc. 23rd ACM Nat. Conf., New York, NY, USA*, 1968, pp. 517–524.

[23] Spector, P., Jex, S., "Development of four self-report measures of job stressors and strain: interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory," *Jour. of Occup. Health Psych.*, vol. 3, no. 4, pp. 356–367, 1998.

[24] Tax, D. M., Duin, R. P., "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.

[25] Tóth, S. L., Sztahó, D., Vicsi, K., "Speech emotion perception by human and machine", in *Proc. COST Action, Patras, Greece*, 2012, pp. 213–224.

[26] Verbunt, M., Gurtz, J., Jasper, K., Lang, H., Warmerdam, P., Zappa, M., "The hydrological role of snow and glaciers in alpine river basins and their distributed modeling," *Jour. of Hydr.*, vol. 282, no. 1, pp. 36–55, 2003.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 59–67

Analyzing F0 Discontinuity for Speech Prosody Enhancement

György SZASZÁK, Miklós Gábriel TULICS, Ákos Máté TÜNDIK

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
e-mail: {szaszak; tulics}@tmit.bme.hu
e-mail: akos.tundik@nokia.com

Manuscript received June 28, 2015; revised September 11, 2015.

Abstract: This research is interested in assessing the pros and contras of using an overall continuous versus a disrupted, not overall defined F0 estimate and compare formal and informal speech styles in this regard. During the evaluation we keep in mind the elaboration of an algorithm capable of handling both speaking styles. The approaches are evaluated in an automatic phonological phrasing task, using a formal and informal speech corpus. A phonological phrasing component is a prosodic unit that has its own stress and intonation contour, which may continue in the next unit. For the different speaking styles we use two speech databases: BABEL for formal speaking style and The Hungarian Spoken Language Database for informal speaking style, both in Hungarian language. Three alternatives of F0 post-processing are compared, ranging from a natural F0 contour disrupted at unvoiced places, over a partial interpolation to an overall continuous contour estimation defined for all unvoiced speech segments. Whereas in formal speech, the more continuous the F0 contour is the better detection rates are observed for phonological phrases, for informal speech a partial interpolation of F0, preserving some fragmentation, yields better results. These results also show that discontinuity of F0 can be an important cue in human perception of informal speech, and also means that the idea of trying to de-spontanize spontaneous speech, in order to be able to treat them in the conventional way, seems to be doubtful.

Keywords: speech prosody, event detection, machine learning, speech signal processing.

1. Introduction

Speech prosody is an important building block of spoken language, carrying information related to several modalities, i.e. prosody provides cues for broad segmentation of the speech stream, carries stress/emphasis, has a discourse

function, reflects speaker attitude and emotions etc. An important speech technology application exploiting prosody is automatic prosodic event detection, i.e. stress detection, or automatic phrasing (segmentation for prosodic units) of the speech flow. Prosodic event detection can be a basic pre-processing step in content or discourse analysis, speech-to-speech translation, that is, in automatic speech understanding applications in general. The three basic acoustically measurable features building up prosody are F0 (fundamental frequency), energy and duration, and they can sometimes be complemented with other, less frequently used features such as jitter, shimmer, harmonics-to-noise ratio (HNR), sub-band energies, etc. In the realisation of prosodic constituents, like stress or intonation patterns, the three basic features interact. The contribution of different features in accomplishing given linguistic functions is language dependent [10], for example, duration is an important cue of stress in American English or German [15], whereas in Hungarian, F0 is believed to be the dominant cue of stress with duration playing almost no role in it [3].

Extracting energy is a basic task, which can usually be carried out without complications. Extraction of duration patterns, on the other hand, may pose problems, especially if the underlying segmental structure (phone segmentation) of the speech signal is unknown. The biggest challenge in this field remains the accurate extraction of F0 [4]. Although several reliable algorithms are known, the F0 estimate is often corrupted by doubling/halving errors [12]. Moreover, the F0 contour is not continuous (it is undefined for unvoiced speech segments), but the human perception of F0 is capable of keeping track of the pitch as if it were continuous. Recent research has found that using a continuous F0 estimate can be advantageous in several applications [5]. Indeed, in speech technology applications, F0 is often interpolated to overcome problems caused by discontinuity [11]. An alternative to this approach has recently been proposed based on probabilistic features [6].

The current paper is interested in exploring these advantages and eventual disadvantages in an automatic prosodic event detection task. A special preference of the evaluation is to keep in mind the elaboration of an algorithm capable of handling both read (formal) and spontaneous (informal) speaking styles. Several studies have proposed phrasing approaches for read and slightly spontaneous speech [1, 2]. In [3] an automatic prosodic phrasing system was implemented for read speech which was able to gain a reliable phrasing down to the phonological phrase level, and also to separate intonational phrase level from the underlying phonological phrase level. The accuracies were ranging between 70-80%. In this approach, prototypes of phonological phrases were clustered, modelled by Hidden Markov Models/Gaussian Mixture Models (HMM/GMM) based on acoustic-prosodic features. In [7] an unsupervised

learning approach for clustering prosodic entities in Hungarian spontaneous speech was described, clustering of such characteristic prototypes led only to partial success.

In this paper the authors compare read and spontaneous speech processing in a prosodic event detection related task, and aim at exploring the advantages and disadvantages of using a continuous vs. disrupted F0 data stream. Three ways of F0 estimations are studied: a continuous, overall interpolated F0 estimation is compared to cases where F0 is fragmented or interpolated only partially. These approaches are evaluated in automatic phonological phrasing system, both for a formal and informal speech corpora. We also believe that these experiments may lead us to a better understanding of spontaneous speech.

This paper is organized as follows: first, the used speech databases are presented, followed by a brief overview of the embedding prosodic event detection system. Experiments are described thereafter, followed by the presentation of the results. Finally, conclusions are drawn.

2. Material and method

This section describes speech databases and basic processing tools used for the experiments later.

2.1 Speech Databases

We use two databases with different speaking styles for the experiments; a formal and an informal one. Both databases got prosodic annotation, representing the top layers of the prosodic hierarchy [9], as follows:

- Intonational Phrases (IP): The IP is the prosodic unit positioned at the top of the prosodic hierarchy, just below the utterance level. The IP is interpreted as a segment of speech that has its own complete prosodic contour, where the first word is accentual (in Hungarian). IP is often found between two pauses.
- Phonological Phrases (PP): The IPs can be divided into PPs: A PP is a prosodic unit that has its own stress and intonation contour, but this contour can be continued in next PP. Syntactically, PP is often related to clitic groups, at least in formal speaking style.

2.1.1 BABEL

The Hungarian BABEL is a read speech corpus (formal speaking style), recorded in a low-noise environment [13]. 60 native Hungarian speakers (30-30

male/female) of varying age and professional background read the utterances. A subset of the database is composed of paragraphs of 5-6 sentences. From this subset, we randomly selected 300 sentences from 22 speakers, and labelled them manually for phonological phrases (2067 in total), according to the 7 types described in [3]. The PP annotation is such that it reveals the IP boundaries unambiguously as well.

2.1.2. Hungarian Spoken Language Database

The Hungarian Spoken Language Database (BEA) [8] is the first Hungarian spontaneous speech database that involves several hundred speakers and a rich speech material in semi-informal and informal speaking styles. The speech material contains different kinds of spontaneous narratives and discourses about personal life, everyday topics, but also includes some sentence repetitions. Informal utterances from 8 speakers (4 female and 4 male) were selected from the database. This subcorpus was manually annotated by two experts for IPs (398 in total) and PPs (751 in total). Again, all phrase boundaries are aligned with word boundaries (this requirement is relatively easy to fulfil in fixed stress Hungarian; however, if stress were unbound, it would not be impossible either).

2.2. Automatic Segmentation for Phonological Phrases

This section describes the automatic PP segmentation algorithm used in the experiments. In this paper, we chose PP segmentation to analyse the effects of varying the post-processing for F0. Before moving on to the experiments, we provide here a brief overview of the system.

As already mentioned, PPs constitute a prosodic unit, characterized by an own stress and some preceding/following intonation contour. The intonation contour might be incomplete (following in the next PP or truncated in informal speech), however, it is specific, hence PPs can be classified/modelled separately, in a data-driven machine learning approach. The distinction between PPs consists of two components: the strength of stress the PP carries and the PPs intonation contour. In this way 7 different types of PPs are distinguished. Acoustic modelling is done with HMM/GMM models, and PP segmentation is carried out as a Viterbi alignment of PPs for the utterances requiring segmentation. The overall approach is documented in detail in [3], featuring as well the 7 types chosen for modelling.

As acoustic-prosodic features, fundamental frequency (F0) and wide-band energy (E) are used. Syllable duration is not used for Hungarian as it was not found to be a distinctive cue in this task [3]. F0 extraction and post-processing alternatives for F0 are described in the respective section later. For energy

computation a standard integrating approach is applied with a window span of 150 ms. Frame rate is 10 ms. First and second order deltas are appended to both F0 and E streams. The evaluation of the PP segmentation is carried out in 10-fold cross-validation. First a PP alignment is generated with models trained on utterances different from the one under segmentation. The generated PP alignment is then compared to the reference obtained by manual labelling. Detection is regarded to be correct if the boundary is detected within the TOL=100-250 ms vicinity of the reference. Here, TOL defines a tolerance interval, ranging between 100 and 250 ms. These values are chosen so to be in the order of a length of a syllable, on average, and much less than average PP length. Once all utterances have the automatic PP segmentation ready, the following performance indicators are evaluated for the PP boundaries:

- recall

$$\left(RCL = \frac{TP}{TP + FN} \right) \tag{1}$$

- precision

$$\left(PRC = \frac{TP}{TP + FP} \right) \tag{2}$$

- the average time deviation (ATD) between the detected and the reference PP boundary. TP refers to the number of true positive, FN to the number of false negative and FP refers to the number of false positive retrieved PP boundaries.

3. Overall and Partial Interpolation for F0

F0 was extracted using the Snack toolkit [14]. F0 data have been subjected to error correction resulting from octal halving/doubling. Keeping the energy unchanged, further post-processing of F0 varies as follows:

- apply only and octal correction, then use the F0 contour as produced by a conventional pitch tracker (Snack V2.2.10 in our case);
- use a continuous contour, interpolated at all unvoiced parts;
- use a partially F0 interpolated contour. This means that the interpolation is omitted if the length of the unvoiced interval is higher than 250 ms, or if the starting F0 value is significantly higher than the value before the unvoiced segment. The criteria used was:

$$F0_{former} \times 1.1 < F0_{current} \tag{3}$$

Speech containing longer periods of silence than 250 ms can hardly be considered as fluent speech. In such situations speakers often may not reset their pitch, so IP (and hence the co-occurring PP) boundaries can be marked by silence. In such cases silence is an acoustic marker by itself and we prevent the interpolation, which may mask the PP boundary. If a medium strength pitch reset occurs it might be smoothed by the F0 interpolation. In the vicinity of long plosives for example, microprosodic disturbances can give false phrase boundary detection. The factor of 1.1 is set to avoid these false detections. These were the motivations to constrain the disruption of the F0 contour in the partial interpolation scenario.

Table 1: Precision (PRC) and recall (RCL) in operating points defined by PRC = RCL and ATD for the 3 evaluated scenarios for formal and informal speech styles TOL = 100 ms).

Style	F0 interpolation	PRC = RCL [%]	ATD [ms]
Formal style	None	62.9	93.0
	Partial	68.2	80.1
	Total	81.2	55.9
Formal style	None	57.7	52.9
	Partial	69.7	44.1
	Total	66.3	44.8

4. Results

Results are shown in Table 1, involving all three tested scenarios for formal and informal speaking styles separately. PP detection rates have different characteristic depending on the speaking style.

Regarding precision (PRC) and recall (RCL), they highly depend on a parameter influencing PP insertion likelihoods during the PP segmentation done with Viterbi alignment (see the PRC-RCL curve in *Fig. 1*). Therefore, segmentation results are shown for operating points where precision and recall are equal (PRC=RCL).

The settings of the tolerance interval TOL also influence results (see *Fig. 2*). We may consider that TOL=100ms corresponds roughly to the length of a half syllable on average and hence TOL=200ms is in the order of the length of a syllable.

As expected, formal and informal speech styles show different characteristics. For read speech, PP detection results are better if the F0 contour is overall interpolated, whereas for the informal speaking style, the partially interpolated F0 contour gives the best results. In case of partial interpolation,

the interpolation was applied only if the length of the unvoiced segment was below the limit of 250 ms, and pitch reset was not suspected.

The results suggest that in the perception of spontaneous speech characterized by informal speaking style, the discontinuity of F0 plays an important role. This also means that the idea of trying to de-spontanize spontaneous speech to that of read speech, in order to treat it with conventional algorithms or tools developed for read speech, seems to be doubtful.

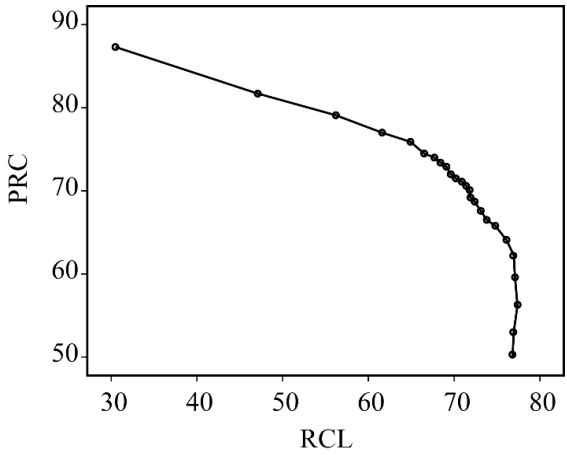

Figure 1: Precision [%] and recall [%] as influenced by the PP insertion likelihood in the automatic PP segmentation. Read speech, total F0 interpolation, TOL = 100ms.

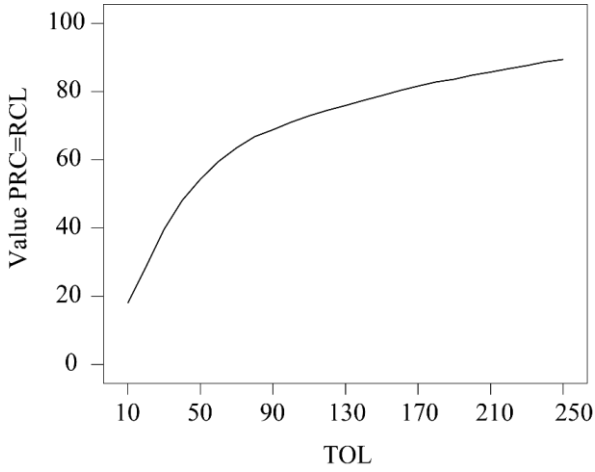

Figure 2: Precision and recall in operating points defined by PRC = RCL [%] depending on TOL [ms]. Read speech, total F0 interpolation.

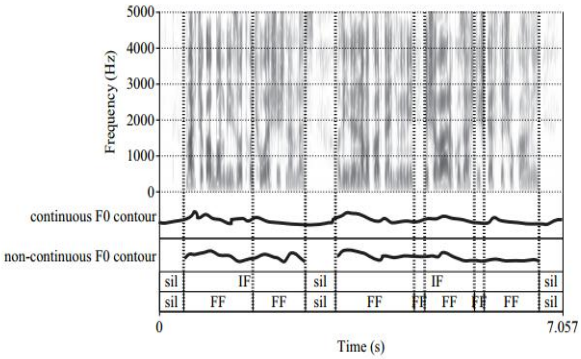

Figure 3: An example with continuous and partially interpolated F0 contours with IP and PP labelling.

5. Conclusions

In this research the authors examined the effect of F0 discontinuity in a phonological phrase segmentation task and drawn a comparison between overall interpolated (continuous) and partially interpolated (fragmented) F0 estimates. Results indicate that the overall interpolated F0 estimate is useful only in read or formal speaking styles. Overall interpolated F0 contour showed better results compared to the partially interpolated one in the phonological phrase segmentation task. The difference between the two methods' precisions was relative 19.1%. The best performing system yielded 81.2% recall and precision in the operating point where type I and type II errors are equal. In terms of precision, this result is fairly comparable to the reported accuracy of other phrase detection tasks by excellent recall rates. Furthermore, a partially interpolated F0 estimate which leaves longer unvoiced periods or pitch reset untouched outperformed the total interpolation one by relative 5.1% for informal speaking style in the same phonological phrase segmentation task. The best precision and recall in the operating point characterized by equal errors was 69.7, which is acceptable for spontaneous speech in international comparison. It follows that piecewise interpolation of F0 seems to be better for spontaneous speech in speech technology applications. Results also suggest considerations regarding human speech perception, but for those we need further investigations with targeted experiments.

Acknowledgements

The authors would like to thank the support of the Hungarian Scientific Research Fund (OTKA) under contract ID PD 112598, titled "Automatic Phonological Phrase and Prosodic Event Detection for the Extraction of Syntactic and Semantic/Pragmatic Information from Speech."

References

[1] Veilleux, N. M., Ostendorf, M., "Prosody/parse scoring and its application in atis," in *Proceedings of the workshop on Human Language Technology*, 1993, pp. 335–340.

[2] Gallwitz, F., Niemann, H., Nöth, E., Warnke, W., "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication* 36, 1-2 (2002) 81–95.

[3] Szaszák, G., Beke, A., "Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech," *Journal of Language Modeling* 1, 0 (2012) 143–172.

[4] Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S., "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, 17, 6 (2009) 365–375.

[5] Yu, K., Young, S., "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, 5, 19 (2011), pp. 1071–1079.

[6] Garner, P. N., Cernak, M., Motlicek, P., "A Simple Continuous Pitch Estimation Algorithm," *IEEE Signal Processing Letters* 20, 1 (2013) 102–105.

[7] Beke, A., Szaszák, Gy., "Unsupervised clustering of prosodic patterns in spontaneous speech, Text, Speech and Dialogue," *Lecture Notes in Computer Science*, 7499 (2012) 648–655.

[8] Neuberger, T., Gyarmathy, D. Gráczi, T. E., Horváth, V., Gósy, M., Beke, A., "Development of a large spontaneous speech database of agglutinative Hungarian language, Text, Speech and Dialogue," *Lecture Notes in Computer Science* 8655 (2014) 424–431.

[9] Selkirk, E., "The Syntax-Phonology Interface," in the International Encyclopedia of the Social and Behavioral Sciences, Pergamon Press, Oxford, 2001, pp. 15407–15412.

[10] Hirst, D., and Di Cristo, A., "Intonation Systems: A Survey of Twenty Languages," Cambridge University Press, New York 1989, p. 256.

[11] Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 2494–2498.

[12] Murray, K., "A study of automatic pitch tracker doubling/halving Errors," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001, 16 pp. 1–4.

[13] Roach, P. S., Amfield, S., Bany, W., Baltova, J., Boldea, M., Fourcin, A., Goner, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meiste, E., Vicsi, K., "BABEL: An Eastern European Multi-language database," in *Proeedings of the International Conference on Speech and Language* 1996, pp. 1033–1036.

[14] Sjölander, K., Beskow, A., "Wavesurfer – an open source speech tool," in *Proceedings of the 6th International Conference of Spoken Language Processing*, 2000, 4, pp. 464–467.

[15] Tamburini, F., Wagner, P., "On Automatic Prominence Detection for German," *Proceedings of Interspeech*, 2007, pp. 1809–1812.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 68–77

Identifying Chains of Software Vulnerabilities: A Passive Non-Intrusive Methodology

Béla GENGE, Călin ENĂCHESCU

Department of Informatics, Faculty of Sciences and Letters,
Petru Maior University of Tg. Mureş, Romania
e-mail: bela.genge@ing.upm.ro, ecalin@upm.ro

Manuscript received May 4, 2015; revised August 20, 2015.

Abstract: We present a novel methodology to identify chains of software vulnerabilities in computer networks. Vulnerabilities constitute software bugs, which may enable attackers to perform malicious operations. Attacks against systems, however, embrace various software flaws on different machines interconnected by networks. As a result, vulnerability chains may give the attacker the ability to compromise a series of hosts, and to reach his goals. This paper shows that an attacker may infer software vulnerabilities by leveraging passive network monitoring tools, and may construct vulnerability chains in the attempt to penetrate security perimeters. We present an approach to automatically build vulnerability chains based an automated vulnerability reconstruction reports from National Vulnerability Database, and passive network analysis tools such as PRADS. The approach is experimentally validated in a laboratory test infrastructure.

Keywords: software vulnerability, vulnerability chain, network sniffer, National Vulnerability Database.

1. Introduction

Vulnerabilities constitute software flaws that may enable attackers to run random code sequences, to escalate privileges, and ultimately to take complete control of underlying Operating Systems (OS). Therefore, extensive effort has been invested in the description and rapid dissemination of software vulnerabilities.

However, starting with their description, the wide dissemination of software vulnerabilities faces several challenges. To begin with, vulnerability reports need to be identically formulated across various institutions and databases. Then, a structured format needs to be provided to enable automated processing

and reasoning with various tools. To facilitate the sharing of vulnerability-related information, the Common Vulnerabilities and Exposures (CVE) was introduced in 1999 by the MITRE Corporation. One of the most well-established vulnerability databases is the National Vulnerability Database (NVD), and it builds on the information provided by CVE.

Besides individual vulnerabilities, security experts need to account for the complexity of various cyber attacks embracing different vulnerabilities of software running across different host. In this context, attackers may adopt chains of software vulnerabilities, in which case exploits of vulnerabilities on one particular host may give access to running a different set of exploits on another set of hosts. Therefore, security experts need to be aware of possible vulnerability chains in order to limit the attacker's ability to reach critical assets.

In an attempt to address this challenge, this paper proposes a methodology to build vulnerability chains based on passive assessment of network traffic. The methodology embraces recent advancement in the field of passive network asset discovery [1], and automated vulnerability reconstruction [2]. Subsequently, it builds on the well-established field of attack trees [3], and their extension with probabilistic computations as proposed in the work of Nai Fovino, et al. [4]. The methodology is experimentally evaluated in the context of a simplified laboratory-scale scenario.

The remainder of this paper is organized as follows. Section 2 provides an overview of related methodologies and it emphasizes the main novelty of the technique proposed in this work. Then, Section 3 provides an overview of vulnerability reports and of the tools employed in this work, which is followed by a detailed description of the proposal. Next, Section 4 provides the results of experiments conducted with the proposed method. Finally, the paper concludes in Section 5.

2. Related work

In the field of vulnerability assessment we find various methodologies, which may be categorized in two classes. In the first class we find approaches from the field of active vulnerability assessment. Here we mention *Nessus*, an "all-in-one" vulnerability assessment tool [11]. *Nessus* actively probes services in order to test for known vulnerabilities and possible service configuration weaknesses. It relies on plug-ins which are specifically written to test for the presence of particular vulnerabilities. It generates a comprehensive report which contains descriptions on discovered assets and vulnerabilities, but also recommendations for improving system security. Next, we mention *ZMap* [12], a network scanner that provides valuable information on discovered services to vulnerability assessment tools.

In the field of passive network asset discovery we can find several tools such as *p0f* [13] and *PRADS* [1], which rely on user-specified signatures to distinguish between specific products and version numbers. These tools generate a list of discovered assets from network traffic capture files.

Finally, we mention the work of Cheminod, et al. [10], which mostly relates to ours. In their work, Cheminod, et al. extended the structure of CVE reports with entries on vulnerability causes and effects in the attempt to build an automated reasoning framework on vulnerability chains. However, the approach requires the modification of each new CVE from other open databases, i.e., NVD. Therefore, we believe that the methodology proposed in this work represents a significant improvement since it leverages open and available CVE reports to build probabilistic vulnerability chains.

3. Proposed methodology

A. Overview of vulnerability reports

As mentioned above, vulnerability reports provide a description of software flaws. Each item in CVE is identified by the year and the order they were included in the database. For example, the recently reported software vulnerability identified by CVE-2015-0699 refers to Cisco Unified Communications Manager's Interactive Voice Response (IVR) component, and the ability of an attacker to launch an SQL injection attack, which may in turn permit the execution of arbitrary SQL statements. The CVE was added to the database in 2015, and was given sequence number 0699.

One of the most well-established vulnerability databases, the National Vulnerability Database (NVD), builds on the information provided by CVE. NVD is often viewed as the "ground truth" for software vulnerability assessment [5]. At the heart of every CVE entry lies the Common Platform Enumeration (CPE), "a standardized method of describing and identifying classes of applications, operating systems, and hardware devices present among an enterprise's computing assets" [6]. CPE names provide information on software vendor, name, version, language, edition, etc.

Each CVE contains various entries among which we mention a summary of the vulnerability, a list of CPEs to which the vulnerability applies, and a scoring on its severity and exploitability. Vulnerability scores are given in the Common Vulnerability Scoring System (CVSS) format. CVSS scores range from 0 to 10, 0 being the lowest (low severity), and 10 being the highest (highest severity) vulnerability score that can be assigned to a specific CVE entry.

B. Overview of automated vulnerability reconstruction tool: ShoVAT

The vulnerability chain reconstruction methodology proposed in this work embraces the advanced features exposed by ShoVAT, as described in our previous work [2]. ShoVAT stands for *Shodan-based vulnerability assessment tool*, and it aims at automatically identifying vulnerabilities in Internet-facing services. ShoVAT uses *Shodan* search engine [7, 8] to discover services and service-specific data, e.g., service banners. Conversely, this work builds on data extracted from network traffic and on ShoVAT's ability to automatically identify software.

Besides service-specific data, ShoVAT employs NVD as a source of CVE reports. CVEs are downloaded once per-day, as they are regularly updated in NVD, and are used to reconstruct an in-memory representation of the relationship between CPEs and CVEs. The memory-resident model is based on bipartite-hypergraphs, where each CPE may be associated to several CVEs, and each CVE may be associated to several CPEs. Besides these, efficient hash functions are implemented in order to provide efficient access to various data structures.

Finally, ShoVAT's output is a comprehensive report detailing the discovered hosts, services, CPEs and vulnerabilities.

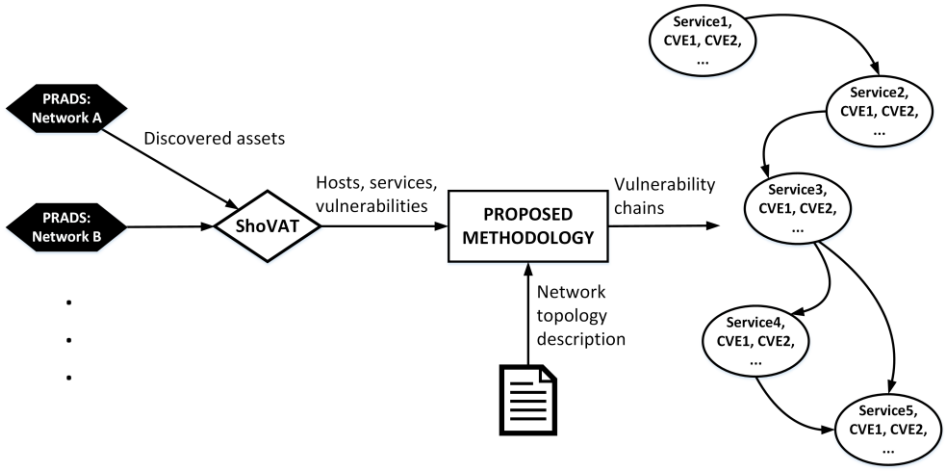

Figure 1: Architecture and components in the proposed methodology.

C. Proposed approach

The proposed methodology for vulnerability chain construction leverages various tools and well-established methodologies (see *Fig. 1*). The network asset discovery is provided by the Passive real-time asset detection system

(PRADS) [1]. Next, the data discovered by PRADS is processed by ShoVAT [2] in the attempt to identify software vulnerabilities. Its output constitutes a list of hosts, services, and CVE reports.

The proposed methodology unfolds several significant steps. At first, it takes a description of the network topology, given as a list of hosts (IP addresses), services (port numbers), and physical connections. This particular entry file also contains logical connections between services, that is, traffic flows between various software and hosts. An obvious enhancement to this data collection methodology could be the adoption of automated network topology discovery mechanisms. However, while such approaches might reduce the time needed to discover the network, their provisioning at various network locations is not trivial. Subsequently, several issues might rise along the way. In this respect, the time dimension of the network discovery phase is a significant aspect, since specific software and inadvertently, communication protocols, might be executed at various points in time.

It is noteworthy that the methodology described in this paper also embraces automated asset discovery methodologies. These are implemented with the help of PRADS, which provides a list of software and OS descriptions for each asset. An example output of PRADS is the following:

```
10.0.0.100,0,80,6,CLIENT,[http:Mozilla/5.0 (Windows
NT 6.1; WOW64; rv:30.0) Gecko/20100101 Firefox/30.0],
1,1404207546
```

In the example above PRADS identified a Firefox Web browser client, version 30.0. The detection is based on patterns described in PRAD's input files. Therefore, the accuracy of the detection depends on various factors, and it is a significant element, which influences the accuracy of results of the proposed methodology as well. An example in this sense is PRAD's ability to detect specific OS. The detection leverages the structure of network packets, specific bits, and so on. However, in many cases the procedure may produce false results due to the lack of sufficient information to accurately distinguish specific OS versions [9]. An example in this sense is the following output, as produced by PRADS:

```
10.0.2.121,0,49251,6,SYN,[8192:127:1:48:M1460,N,N,S:.
:Windows:2000 SP2+, XP SP1+ (seldom 98):link:
ethernet/modem],1,1404207849
```

According to this example, PRADS states that the OS version might be Microsoft Windows 200 SP2+, yet it may also be Microsoft Windows XP SP1+, and in rare occasions Microsoft Windows 98. Therefore, such output

poses significant challenges even in the hand of experienced security experts. Nevertheless, it should be noted that since PRADS is a passive network asset discovery tool, it is limited to such outputs, and a more accurate result may be only achieved by leveraging active detection methodologies.

In the next phase, PRAD's output is processed by ShoVAT's automated vulnerability identification modules. Since ShoVAT was intended to process data from Shodan, we extended ShoVAT with a new module written in the Python language in order to process PRAD's output. ShoVAT takes PRAD's output and reconstructs CPE names based on asset names and version numbers. Then, it extracts all known CVEs for all matching CPEs. As an example in this sense, for the previous example on Firefox ShoVAT identifies the following CPE:

```
cpe:/a:mozilla:firefox:30.0
```

Then, ShoVAT searches for known CVE and gives 52 matches. This yields a number of 52 known vulnerabilities that are associated to Mozilla Firefox version 30.0.

Based on these aspects, in the next phase the methodology at hand builds a graph-based representation of nodes, vulnerabilities, and vulnerability scores. The most significant issue with the construction of such graphs, however, is the actual exploitability of vulnerabilities. That is, for each CVE we need to establish if an attacker may exploit the vulnerability, and what are the consequences of the exploited vulnerability.

Unfortunately CVEs do not formalize these concepts, and they are summarized in an intuitive, natural language-based form. In fact, related research on building vulnerability chains [10] extended the structure of CVE with entries that specifically identify the possible exploits. However, the developed methodology is not accessible to the public, and it entails the extension of each CVE with such specific entries. Therefore, the adoption of the methodology as proposed in [10] may not be feasible in real systems due to the high maintenance requirements for keeping a separate and up-to-date database of extended CVEs. This is also easily understandable since at the moment the entire NVD contains more than 100 thousand CVEs, and their expansion with exploitability-specific information would be highly unreasonable.

Based on these facts, in this work we adopt the numerical scoring available in CVE as a measure of vulnerability exploitability. As already stated, CVEs include a set of CVSS values, among which we also find various sub-scores such as *exploitability* and *impact*. Since these have already been assigned by security experts, we adopt these numbers in the proposed methodology. More specifically, since their values are in the range from 1 to 10, we use them in the

form of probability assignments. In other words, the exploitability sub-score will represent the probability of successful exploitation (after division by 10), denoted by P_{EXP}, while the impact sub-score will represent the probability of escalating privileges (after division by 10), denoted by P_{IMP}. As a result, the probability of successful escalation of privileges by means of software vulnerability exploitation is computed as $P_{SE} = P_{EXP} \bullet P_{IMP}$.

Then, from the host's point of view, which encapsulates various software and vulnerabilities, the probability of privilege escalation includes all software and vulnerabilities found on that particular host. More specifically, the privilege escalation for a host i is defined according to [4] as $P_{Hi} = 1 - Prod_k(1 - P_{SEik})$, where $Prod_k$ computes the product of k successful privilege escalation probabilities for host i. The chain probability between two hosts is computed as $P_{Hi} \bullet P_{Hj}$.

Finally, given the above equations, we can calculate the probability for an attacker to successfully reach his target. For this purpose we use well-known graph-based algorithms, which maximize/minimize the cost of a specific path in the graph.

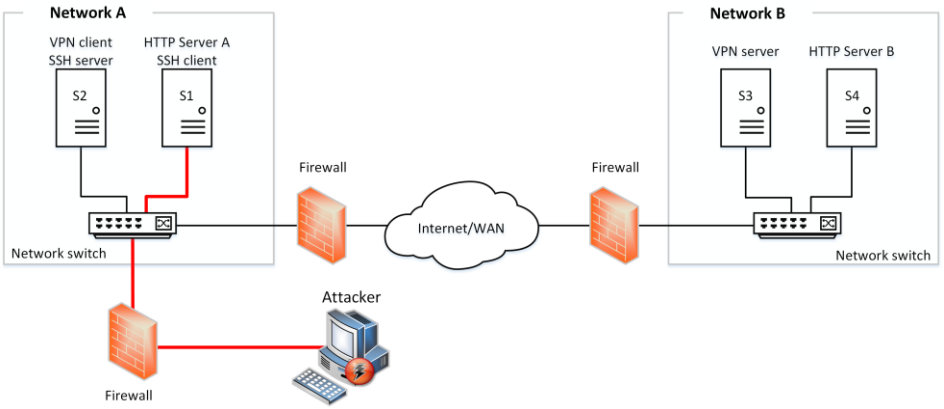

Figure 2: Experimentation setting.

4. Experimental results

The experimentation assessment assumes a simplified topology (see *Fig. 2*) consisting of two networks, i.e., *Network A* and *Network B*. Network A hosts two nodes: S1, running HTTP server A and SSH client, and S2, running a VPN client to Network B and SSH server. Network B hosts another two nodes: S3, running VPN server, and S4, running HTTP server B.

The attacker is located outside of the two networks. Both networks are protected by firewall, however, certain services are allowed through. As such, the attack is permitted to access HTTP server A on S1, and VPN client in Network A can connect to VPN server in Network B. In the case of the two HTTP servers we use Apache Web server version 2.4.4.

Based on this scenario we use PRADS to identify network assets on both networks and we feed this data to ShoVAT. Finally, we use the proposed methodology to build the vulnerability chain.

With the data discovered by PRADS, ShoVAT reconstructs the following CPEs associated to HTTP server A:

```
cpe:/a:apache:http_server:2.4.4
cpe:/a:apache:apache_http_server:2.4.4
```

Based on these CPEs ShoVAT then identifies several vulnerabilities. For the sake of the example at hand we select one particular example in order to illustrate the application of the proposed approach. More specifically, we select CVE-2013-2249, a vulnerability from 2013 affecting all Apache HTTP servers before version 2.4.5.

According to NVD, the aforementioned vulnerability is associated to `mod_session_dbd.c` in the `mod_session_dbd module`, which may allow an attacker in certain cases to run various attack vectors. The vulnerability has a CVSS score of 7.5, therefore, a *high* score. The impact sub-score is of 6.4, while the exploitability is of 10. Therefore, an attacker might exploit the vulnerability with a probability of 1, yet, the probability of successful escalation of privileges, is of 0.64.

As a result, a remotely located attacker, outside the perimeters of Network A, may successfully exploit the vulnerability with the probability of 0.64. Then, the attacker can open an SSH connection to S2, from where it gains access to S3. This entails, however, that the VPN configuration between S2 and S3 does not require further verification of credentials.

Once the attacker reaches this point, i.e., S3, it can adopt the same vulnerability in Apache HTTP server to escalate privileges on S4 with a probability of 0.64. As a result, given the two cascading, i.e., series, probabilities, we compute that the probability for an attacker to reach S4 remotely is of $0.64 * 0.64 = 0.4096$. Therefore, the attacker might have almost 50% of chances to reach his goals from outside Network A.

Based on these results, we consider that the proposed approach represent a powerful instrument in the identification of highly critical vulnerability chains. As such, the methodology may be adopted by security specialists in order to reduce the attacker's probability of success. In the scenario at hand, for

instance, by assuming that HTTP Server A has only non-critical vulnerabilities, with a probability of successful escalation of privileges of 0.15, the overall success probability for the entire vulnerability chain decreases to 0.096. Therefore, once the vulnerabilities have been identified, the methodology provides the ability to conduct what-if calculations, which can also aid network designers to identify possible solutions and security improvements.

5. Conclusion

We presented a methodology to construct software vulnerability chains. The approach builds on passive network asset discovery tools such as PRADS [1] and non-intrusive automated vulnerability identification features exposed by ShoVAT [2]. The output of these tools is processed in the attempt to construct a graph representation of network, hosts, software, and vulnerabilities. The work also proposes metrics to quantify in a probabilistic framework the successful exploitation of vulnerabilities. In this respect we adopt the impact and exploitability sub-scores as provided by CVSS. The applicability of the proposed approach is demonstrated in a laboratory-scale test scenario. Future research will focus on additional tests and measurements in various scenarios.

Acknowledgements

The research presented in this paper was supported by the European Social Fund under the responsibility of the Managing Authority for the Sectoral Operational Programme for Human Resources Development, as part of the grant POSDRU/159/1.5/S/133652.

References

[1] Fjellskal, E., "Passive real-time asset detection system (PRADS)," 2009, [Online; available at: http://gamelinux.github.io/prads/].

[2] Genge, B., Enachescu, C., "ShoVAT: Shodan-based vulnerability assessment tool for Internet-facing services," *Security and communication networks*, Wiley, 2015 (In Press), [Online; available at: http://www.ibs.ro/~bela/Papers/SCN2015.pdf].

[3] Schneier, B., "Attack trees," *Dr. Dobbs journal*, vol. 24, no. 12, pp. 21–29, 1999.

[4] Nai Fovino, I., Masera, M., and De Cian, A., "Integrating cyber attacks within fault trees", *Rel. Eng. & Sys. Safety*, vol. 94, no. 9, pp. 1394-1402, 2009.

[5] Nannen, V., "The Edit History of the National Vulnerability Database," *Master's Thesis, ETH Zurich*, Switzerland 2012.

[6] Cheikes, B., Waltermire, D., and Scarfone, K., "Common platform enumeration: Naming specification version," *Technical Report NIST Inter-agency Report 7695*, NIST August 2011.

[7] Matterly, J., "Shodan," 2015, [Online; available at: http://www.shodanhq.com].

[8] Bodenheim, R., Butts, J., Dunlap, S., and Mullins, B., "Evaluation of the ability of the Shodan search engine to identify Internet-facing industrial control devices", *International Journal of Critical Infrastructure Protection*, vol. 7, no. 2, pp. 114-123, 2014.

[9] Richardson, D.W., Gribble, S.D., Kohno, T., "The limits of automatic OS fingerprint generation", *in Proc. of the 3rd ACM Workshop on Artificial Intelligence and Security*, AISec '10, ACM: New York, NY, USA, 2010.

[10] Cheminod, M., Bertolotti, I.C., Durante, L., Maggi, P., Pozza, D., Sisto, R., and Valenzano, A., "Detecting Chains of Vulnerabilities in Industrial Networks", *IEEE Transactions on Industrial Informatics*, vol. 5, no. 2, pp. 181-193, 2009.

[11] Rogers, R., "Nessus Network Auditing", Syngress publishing, 2008.

[12] Durumeric, Z.,Wustrow, E., and Halderman, J.A., "Zmap: Fast Internet-wide scanning and its security applications", *in Proc. of the 22Nd USENIX Conference on Security*, SEC'13, USENIX Association: Berkeley, CA, USA, pp. 605–620, 2013.

[13] Zalewski, M., "p0f v3: Passive fingerprinter", 2012, [Online; available at: http://lcamtuf.coredump.cx/p0f3/.

Acta Universitatis Sapientiae

The scientific journal of Sapientia University publishes original papers and surveys
in several areas of sciences written in English.
Information about each series can be found at
http://www.acta.sapientia.ro.

Acta Universitatis Sapientiae
Electrical and Mechanical Engineering

Sapientia University

Scientia Publishing House

ISSN 2065-5916
http://www.acta.sapientia.ro

Information for authors

Acta Universitatis Sapientiae, Electrical and Mechanical Engineering publishes only original papers and surveys in various fields of Electrical and Mechanical Engineering. All papers are peer-reviewed.

Papers published in current and previous volumes can be found in Portable Document Format (PDF) form at the address: `http://www.acta.sapientia.ro`.

The submitted papers must not be considered to be published by other journals. The corresponding author is responsible to obtain the permission for publication of co-authors and of the authorities of institutes, if needed. The Editorial Board is disclaiming any responsibility.

The paper must be submitted both in MSWord document and PDF format. The submitted PDF document is used as reference. The camera-ready journal is prepared in PDF format by the editors. In order to reduce subsequent changes of aspect to minimum, an accurate formatting is required. The paper should be prepared on A4 paper (210×297 mm) and it must contain an abstract of 200-250 words.

The language of the journal is English. The paper must be prepared in single-column format, not exceeding 12 pages including figures, tables and references.

The template file from http://www.acta.sapientia.ro/acta-emeng/emeng-main.htm may be used for details.

Submission must be made only by e-mail (`acta-emeng@acta.sapientia.ro`).

One issue is offered to each author free of charge. No reprints are available.

Contact address and subscription:
Acta Universitatis Sapientiae, Electrical and Mechanical Engineering
RO 400112 Cluj-Napoca
Str. Matei Corvin nr. 4.
E-mail: acta-emeng@acta.sapientia.ro

Printed by Digital Color Company
www.digitalcolorcompany.ro