Audio-Visual Signal Processing in a Multimodal Assisted Living Environment

Alexey Karpov [1,5], *Lale Akarun* [2], *Hülya Yalçın* [3], *Alexander Ronzhin* [1], *Barış Evrim Demiröz* [2], *Aysun Çoban* [2] *and Miloš Železný* [4]

[1] St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[2] Boğaziçi University, İstanbul, Turkey
[3] İstanbul Technical University, İstanbul, Turkey
[4] University of West Bohemia, Pilsen, Czech Republic
[5] University ITMO, St. Petersburg, Russia

karpov@iias.spb.su, {akarun,baris.demiroz,aysun.coban}@boun.edu.tr,
hulyayalcin@itu.edu.tr, ronzhinal@iias.spb.su, zelezny@kky.zcu.cz

10.21437/Interspeech.2014-267

Abstract

In this paper, we present some novel methods and applications for audio and video signal processing for a multimodal environment of an assisted living smart space. This intelligent environment was developed during the 7th Summer Workshop on Multimodal Interfaces eNTERFACE. It integrates automatic systems for audio and video-based monitoring and user tracking in the smart space. In the assisted living environment, users are tracked by some omnidirectional video cameras, as well as speech and non-speech audio events are recognized by an array of microphones. The multiple objects tracking precision (MOTP) of the developed video monitoring system was 0.78 and 0.73 and the multiple objects tracking accuracy (MOTA) was 62.81% and 72.31% for single person and three people scenarios, respectively. The recognition accuracy of the proposed multilingual speech and audio events recognition system was 96.5% and 93.8% for user's speech commands and non-speech acoustic events, correspondingly. The design of the assisted living environment, the certain test scenarios and the process of audio-visual database collection are described in the paper.

Index Terms: elderly healthcare, acoustic event detection, speech recognition, video surveillance, assistive technology

1. Introduction

Nowadays in the framework of the EU FP7 program, there is a special research and development direction focused on creation of assistive smart spaces, homes and intelligent environments called Ambient Assisted Living (AAL). It specifies the future of assistive information technologies aimed for information support and care of disabled and elderly people. AAL European program includes several International projects (e.g. AGNES, DOMEO, HAPPY AGEING, HOPE, SOFTCARE, etc. [1]), and some other national projects like Sweet Home, CIRDO, ALADIN, homeService, WeCare [2-4].

With the advances in real time monitoring systems, intelligent environments equipped with audio and video sensors offer promising solutions for home care and independent living applications. They also improve the quality of the care the users are provided with, especially it concerns the growing population of elderly people and people with disabilities. There are numerous techniques involving different modalities such as special bracelets equipped with sensors to get information from the human's body, special beds sensing the person laying on it, infrared or colour cameras mounted to the walls, and audio sensors detecting any emergency calls [5, 6]. Among these, video cameras and microphones installed in the smart environment are pervasively used, and these modalities are taken into account in this study for daily monitoring of single elderly persons.

Visual data is mainly used for person tracking and automatic activity recognition in the assisted environment. In action recognition, videos are rapidly segmented in time and an activity is defined as an ordered set of actions (e.g. cooking is an activity whereas stirring is an action). Human action understanding has many application areas concerning security, surveillance, assisted living, edutainment, etc. With an activity recognition unit, it is possible to detect some set of actions involving emergency states such as falling, and to warn the system users of the extraordinary situation as it is discussed in [7]. Human activity recognition modules consist of people tracking, describing their motions by means of visual clues such as keypoints and classifying the performed actions [8, 9]. The detection and tracking of people are based on lower level background/foreground segmentation techniques such as Gaussian mixture models (GMM) [10, 11], and codebook model [12]. The blobs found after foreground segmentation are used for extracting features, which describe the related person. Good survey on the widely-used interest point detection and feature extraction methods is given in [13]. Recent works show that the interest point descriptors are mainly used in cameras with low distortion, but the use of them alongside with omnidirectional (fisheye) cameras is a new direction for video surveillance.

Reasoning based only on the visual information is not robust since the illumination may change severely, occlusions may occur and so that the action may be misclassified. The use of audio signal processing makes the system more robust and maintainable. Audio signals consist of human speech and environmental sounds such as knocking at the door and water coming out of a pipe. Automatic speech recognition (ASR) is a process of converting speech into a sequence of words by means of algorithms implemented as a software or hardware module. Recent ASR systems exploit mathematical techniques such as Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Bayesian Networks methods, etc. [14].

Although natural speech is the most informative audio signal, other auditory events may also convey useful information. Therefore detection and classification of acoustic events may be used for people activity detection and acoustic monitoring (e.g. applause or laugh during a dialog, chair movements, cough, cry, etc.) [15]. There are some recent publications on automatic detection of individual acoustic events such as cough [16-18], sounds of human fall [19-22], cry and scream [23], distress calls [24, 25], etc. Most of these

systems use standard methods for ASR like classifiers based on HMMs, GMMs, support vector machines (SVMs) or ANN with some extra features: zero-crossing rate, audio signal energy, event duration, impulse form, etc. Classification and/or detection of acoustic events is a new field of analysis of acoustic scenes [26]. Systems for detection of acoustic events may be used in different environments like a meeting room [27], hospital [28], kitchen [29] or even bathroom [30].

In this paper, we present a multimodal assisted living environment that was designed and researched during the 7th International Summer Workshop eNTERFACE held in Pilsen [31]. In this research, we used some omnidirectional cameras installed in the environment for single and multiple people tracking and room monitoring, as well as some microphone arrays for automatic detection and processing of speech and non-speech acoustic events.

2. Methodology and methods

Our study deals with processing audio and video modalities. Multimodal user interfaces based on audio-visual signal processing allow organizing natural interaction and communication between a user and a smart environment.

2.1. Design of the assisted living environment

The design of the assisted living room (studio of 64 square meters) is shown in Figure 1. It is an ordinary room with a hard floor (linoleum) that contains 2 chairs, 2 tables and a sink, as well as 2 omnidirectional cameras and 4 stationary microphones. First camera was mounted on the ceiling and the other one is on the side wall. The frame resolution is 640×480 pixels and the frame rate is 8 fps for both cameras. The cameras (Mobotix) offer higher resolution, but with a lower fps (frames per second) that was not sufficient for tracking.

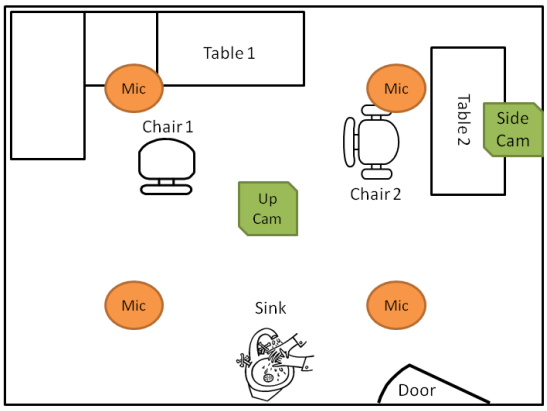

Figure 1: *Overall design of the assisted living environment.*

4 condenser microphones Oktava MK-012 connected to a multi-channel external sound board M-Audio ProFire 2626 have been installed in the smart room too. The microphones are located on the ceiling on a grid so that to take the input clearly from the person with the nearest microphone.

2.2. Video-based environment monitoring

Two video cameras have been installed in the assisted living environment for monitoring the room space. Information we need to gather for understanding the underlying positions and actions of the people is extracted from the non-stable regions in the scene (foreground regions). GMMs are used to segment the foreground objects; shadow removal is additionally applied to enhance the result by removing artefacts due to small changes in chromaticity and intensity [11]. Each pixel is considered individually and labelled as foreground or background using a statistical approach based on multiple Gaussian distributions. Background pixels form a Gaussian with a higher weight since they stay for longer time in the scene in this model whereas the foreground objects form new Gaussians with smaller weights and leave the scene. Thus, they fire smaller values given the background model, which is used to identify them. Another challenge is the existence of shadows in the environment. The shadow regions have a small amount of chromaticity and intensity difference to the background. For a given pixel value, a specific amount of chromatic distortion and intensity distortion from the background model is allowed [10].

We use state-of-the-art methods for processing blobs (a group of pixels) to infer higher level information about the objects such as the appearance or the performed activity [13]. We apply morphological closing to fill the holes and connect the nearby pixels to form a unified connected component since the foreground segmentation module works on individual pixels. Under the assumption of single user scenario, an area threshold is used and all the components having smaller area are removed. Then, the largest connected component is selected to be the person and it is tracked. In multiple people tracking, we seek for good matching of blobs based on descriptors extracted from the interest points on the blobs.

Due to the requirement of having real-time feedback in the assisted living environment, FAST feature point detection is preferred to slower alternatives such as Harris or Difference of Gaussians [32]. The keypoints are extracted only from the bounding rectangles of the found foreground parts for each frame. The use of bounding rectangles instead of the actual segmentation has provided better results, since foreground detection can result in smaller regions than the actual foreground. After locating the keypoints, corresponding BRIEF (Binary Robust Independent Elementary Features) descriptors are extracted from the image [33]. BRIEF keypoint descriptors are preferred to SURF or SIFT for speed reasons. BRIEF descriptors are found using relatively simple intensity difference tests and reported to be highly discriminative even when using relatively few bits. An additional advantage is that it uses binary representation that makes the comparisons efficient by using Hamming distance. BRIEF is originally designed to work on greyscale images, but we extended it to work on colour images to improve the descriptor power. Also we followed the colour boosting transformation in the opponent colour space [34].

For a given frame, for every blob's keypoint, the closest keypoint is searched in several previous frames. Matched keypoint's blob's label count is incremented. Each label is assigned to the blob in such a way that total number of counts is maximized. This is known as an assignment problem, and is solved with Hungarian algorithm. If any blob is left unlabeled, a new label is created and assigned to that blob.

2.3. Audio-based environment monitoring

The recognizer of speech and non-speech audio events is based on HMMs modeling and calculates Mel-frequency cepstral coefficients (MFCCs) from multi-channel audio signals [35, 36]. The system uses HMM-based Viterbi

algorithm for finding an optimal unit from incoming audio signal. Architecture of the audio and speech recognition system is presented in Figure 2. All speech commands, audio events and a garbage model (undefined audio sound) are described by a context-free grammar that allows the system to output only one event in each recognition hypothesis. The recognition process is quite fast (the speed factor SF < 0.1 real-time) and the recognition result is available just after voice/audio activity detection (energy-based VAD).

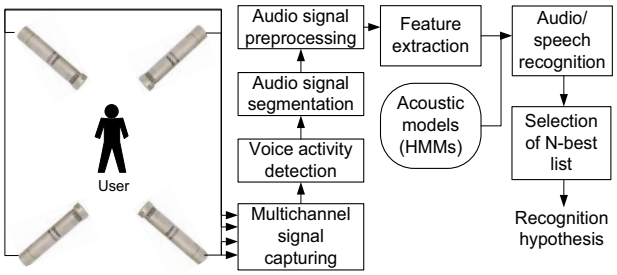

Figure 2: *Architecture of audio event/speech recognition system*

The recognition system is speaker-dependent due to system's aim and usability issues. The developed ASR system is multilingual and it is able to recognize and interpret some speech commands both in Russian and English. Figure 3 presents a tree classification of user's speech commands and acoustic events, which are modelled in the system.

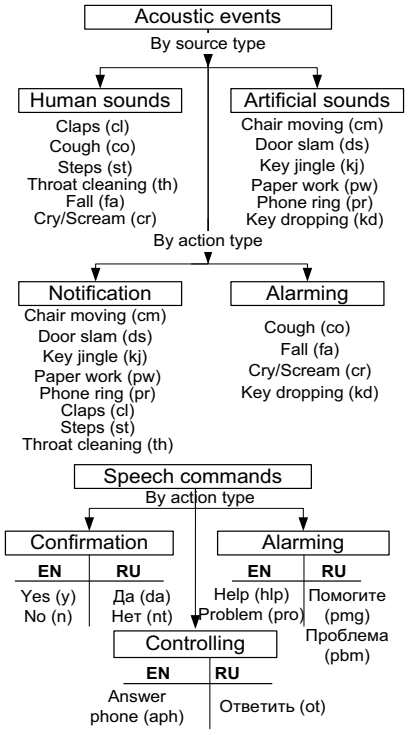

Figure 3: C*lassification of audio events/speech commands.*

The recognition lexicon contains 5 English/Russian spoken commands (words or phrases), plus 12 non-speech acoustic events for different types of human's activities.We also defined a set of alarming speech and audio events, which can be a signal on a critical situation with the user in the assisted living room: *X*={"Help", "Problem", "Cry", "Cough", "Fall", "Key/object drop"}. To train statistical acoustic models of the recognizer, an audio corpus has been collected in room conditions with an acceptable level of background noise (25-30 dB). In total, we have recorded over 1.3 hours of audio data of several potential users performing certain scenarios.

3. Multimodal database collection

Two scenarios were developed as a prototype for basic actions performed by people in a smart home. First scenario involves single person and simulates an emergency situation. The latter one is a more challenging and involves up to 3 subjects, where the subjects occlude each other at some frames.

Scenario 1 (audio-visual data) has the following steps: Enter room from the door (open & close); Walk to table 1; Pick up a glass of water from the table 1; Walk to the chair 1; Sit on the chair 1; Drink water; Cough after drinking; Stand up; Walk to the table 1; Release the glass; Walk to the sink; Wash hands; Exit the room (open & close); Enter the room (open & close); Walk to the chair 2; Sit on the chair 2; Phone rings on table 2; Say "Answer phone"; Talk on the phone; Say "Hello"; Say "I'm fine"; Say "Don't worry"; Say "Bye"; Stand up; Walk to the table 1; Pick up a metallic cup from the table 1; Free walk; Drop the cup on the floor and leave it; Free walk; Fall; Cry for "help".

Scenario 2 (only video data) has the following steps: P1 (person 1) enters the the room; P1 walks to the chair 1; P1 sits on the chair 1; P2 enters the room; P2 walks to the table 1 passing from the right-side of P1; P1 stands up; P1 walks to the table 1 (on the left-side of P2); P1 and P2 shake hands; P3 enters the room; P1 walks to the middle and P3 walks to the middle; P1 and P3 meet each other and shake hands; P2 leaves the room by passing from the right-side of the room; P1 and P3 walk to the table 2; P1 and P3 stay in front of the table 2; P1 walks to the door and exits; P3 walks to the door and exits.

During the multimodal database collection at the eNTERFACE Workshop we recorded audio-visual samples of Scenario 1 from 6 different subjects (potential users). They were free to perform the fall (on the hard floor that produces a sound) as it was comfortable for them. In the case of Scenario 2, we have collected 18 samples from 5 different subjects interchanging the roles of users (P1, P2, and P3) in the scenario. All videos were recorded in MJPEG format, which is the original format provided by the cameras.

We recorded audio data in regular acoustic conditions, where a factor of entering of a new person into the room during the recording session was avoided in order to remove significant external noises. Five check points were selected in the room for collecting the audio corpus. 4 of them were located on the floor under each microphone (located on the ceiling 2.5 meters above the floor) and the last one was in the centre of the room. Each of 5 subjects performed the following sequence of actions: (1) come to a check point; (2) give a speech command or simulate a non-speech audio event for 5 times with a pause between utterances; (3) move to the next check point. All speech commands and acoustic events were simulated 100 and 200 times, correspondingly. In total, above 2800 wave files were recorded, 44% of these audio files are non-speech events and 56% of them are speech commands. 70% recordings for each subject were used for system's training and the rest of the data were used in the experiments. This corpus (SARGAS DB) has been registered in Rospatent (#2013613086 on 25.03.2013) and is available on request.

4. Experimental results

4.1. Experiments with the visual modality

For implementation and evaluation of the video-based monitoring, an open-source computer vision library Open CV was used [37]. Figure 4 shows video frames from the top camera with the examples of position detection of a single person and three persons.

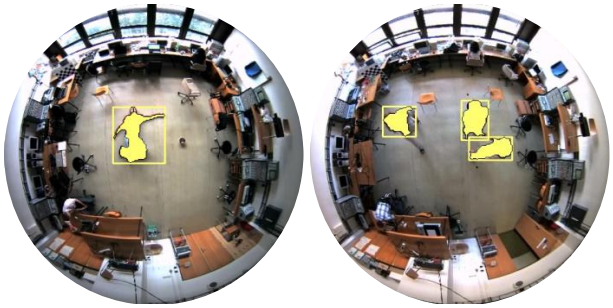

Figure 4: *Examples of position detection of a single lying person (left) and three persons (right) from the top camera.*

The evaluation metrics used for person tracking are Multiple Object Tracking Precision MOTP and Accuracy (MOTA) [38]. MOTP (overlap) metric can be interpreted as the average ratio of intersection and the union of the true object and hypothesized boundaries. For 10 test videos recorded in Scenario 1, MOTP and MOTA were calculated as 0.78 and 64.85% respectively, whereas the miss, false alarm and mismatch rates were 2.92%, 32.11% and 0.10%.

For 36 test videos recorded in Scenario 2, MOTP (overlap) and MOTA were calculated as 0.73 and 72.31% respectively where the miss, false alarm and mismatch rates were 23.47%, 3.74% and 0.47%. Since Scenario 2 contains multiple people occluding each other, its miss ratio is higher compared to the Scenario 1. When the performance of the video-based tracker is investigated in detail, it was observed that miss ratio is significantly higher for the videos acquired from top camera in both scenarios. This can be explained by the extended duration of the occlusion that is present in videos of the top camera.

We also proposed a simple method for person's fall detection. We use the fisheye camera on the ceiling, and if the area (blob) of the tracked person is suddenly increased (in 1.5-2 times) it can be a signal that he/she has fallen and lies. At that we use also some restrictions: the user may lie on a bed, on a sofa or on a chair, otherwise it is interpreted as an emergency situation (his/her fall) signal.

4.2. Experiments with the audio modality

The automatic system for recognition of speech commands and audio events was evaluated using audio recordings made in the assisted living environment (evaluation part of the audio corpus). Also all audio-visual data recorded in the Scenario 1 were used in our experiments. Tables 1-2 show recognition accuracy [39] (after VAD) in the form of confusion matrices for speech commands and non-speech audio events, correspondingly. Presented results show that the most of speech commands were recognized with a high accuracy (recognition rate) over 90%, but there were some mistakes as well. For example, the system recognizes the English command "Problem" as the Russian word "Проблема" (same meaning), because pronunciations are very close.

Table 1. *Accuracy of speech command recognition.*

Speech commands	Recognition accuracy, %									
	aph	hlp	n	pro	y	d	nt	ot	pmg	pbm
aph	100									
hlp		85								15
n			100							
pro				94						6
y					100					
d						100				
nt							100			
ot								100		
pmg				2					95	3
pbm				9						91

Table 2. *Accuracy of recognition of acoustic events.*

Audio events	Recognition accuracy, %											
	cl	cm	co	cr	ds	fa	kd	kj	pw	pr	st	th
cl	100											
cm		100										
co			100									
cr				100								
ds				2	98							
fa			1			63					36	
kd	15						75		10			
kj								100				
pw	4								96			
pr										100		
st			6								94	
th												100

The presented results show that the lowest accuracy was observed for the non-speech audio event "Fall". In 1/3 of such cases this event was recognized as the event "Steps".

5. Conclusion and future work

We presented the multimodal assisted living environment that was developed during the 7th Summer Workshop on Multimodal Interfaces eNTERFACE (project "Multimodal assisted living environment"). It uses sub-systems for audio- and video-based monitoring and user tracking. The developed system for speech and audio events recognition is multilingual and it is able to recognize commands in 2 languages. The audio corpus (SARGAS DB), containing over 2800 wave files with user's speech commands and simulated acoustic events, was collected. For the evaluation part of the corpus the recognition accuracy of speech and acoustic events was 96.5% and 93.8%, correspondingly. In our experiments on video-based user tracking, the multiple object tracking precision (MOTP) of the proposed monitoring system was 0.78 and 0.73 for single person and three people scenarios, respectively. The multiple object tracking accuracy MOTA was 62.81% and 72.31% for these scenarios, correspondingly.

In future work, we are going to merge two systems for audio- and video-based monitoring in one assistive system to improve accuracy of multimodal event detection. Also the multimodal system will be applied to real operation conditions and evaluated with attraction of end-users. This research is partially supported by the Government of Russian Federation (Grant № 074-U01).

6. References

[1] Ambient Assisted Living (AAL) Program. Online: http://www.aal-europe.eu, accessed on 20 June 2014.

[2] Portet, F., Vacher, M., Golanski, C., Roux, C. and Meillon, B. "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects", Personal and Ubiquitous Computing, 32(1): 1-18, 2011.

[3] Christensen, H., Casanueva, I., Cunningham, S., Green, P. and Hain, T. "homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition", In Proc. 4th Workshop on Speech and Language Processing for Assistive Technologies SPLAT-2013, France, 29-34, 2013.

[4] Alemdar, H., Yavuz, G., Özen, M., Kara, Y., Incel, Ö., Akarun, L. and Ersoy, C., "Multi-modal fall detection within the WeCare framework", In Proc. 9th ACM/IEEE International Conference on Information Processing in Sensor Networks IPSN'10, New York, USA. 436-437, 2010.

[5] Nakashima, H., Aghajan, H., Augusto, J. C. and Nakashima, H., "Handbook of Ambient Intelligence and Smart Environments", in H. Nakashima, H. Aghajan, & J.C. Augusto [Ed]. Boston, MA: Springer Verlag. 2009.

[6] Papadopoulos, A., Crump, C. and Wilson, B., "Comprehensive home monitoring system for the elderly", In Proc. WH'10 Wireless Health, New York, USA. 214-215, 2010.

[7] Kara, Y. and Akarun, L., "Human action recognition in videos using keypoint tracking", In Proc. 19th Signal Processing and Communications Applications Conference SIU-2011, Antalya, Turkey, 1129-1132, 2011.

[8] Poppe, R., "A survey on vision-based human action recognition", Image and Vision Computing, 28(6): 976-990, 2010.

[9] Weinland, D., Ronfard, R. and Boyer, E., "A survey of vision-based methods for action representation, segmentation and recognition", Computer Vision and Image Understanding, 115(2): 224-241, 2011.

[10] KaewTraKulPong, P. and Bowden, R., "An improved adaptive background mixture model for real-time tracking with shadow detection", Proc. European Workshop Advanced, 1(3):1-5, 2001.

[11] Zivkovic, Z. and van der Heijden, F., "Efficient adaptive density estimation per image pixel for the task of background subtraction", Pattern Recognition Letters, 27(7): 773-780. 2006.

[12] Kim, K., Chalidabhongse, T., Harwood, D. and Davis, L., "Real-time foreground–background segmentation using codebook model", Real-Time Imaging, 11(3): 172-185. 2005.

[13] Tuytelaars, T. and Mikolajczyk, K., "Local Invariant Feature Detectors: A Survey", Foundations and Trends in Computer Graphics and Vision, 3(3): 177-280. 2007.

[14] Besacier, L., Barnard, E., Karpov, A. and Schultz, T., "Automatic speech recognition for under-resourced languages: A Survey", Speech Communication, 56(1): 85-100, 2014.

[15] Temko, A., Malkin, R., Zieger, C., Macho, D. and Nadeu, C., "Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems", IV Jornadas en Tecnología del Habla, Zaragoza, 5-11, 2006.

[16] Drugman, T., Urbain, J. and Dutoit, T., "Assessment of audıo features for automatıc cough detectıon", In Proc. EUSIPCO-2011, Barcelona, Spain, 1289-1293, 2011.

[17] Takahashi, S., Morimoto, T., Maeda and S., Tsuruta, N., "Cough detection in spoken dialogue system for home health care", In Proc. INTERSPEECH-2004, Jeju, Korea, 1865-1868, 2004.

[18] Huynh, T.H., Tran, V.A. and Tran, H.D., "Semi-supervised tree support vector machine for online cough recognition", In Proc. INTERSPEECH-2011, Florence, Italy, 1637-1640, 2011.

[19] Zigel, Y., Litvak, D. and Gannot, I., "A Method for Automatic Fall Detection of Elderly People using Floor Vibrations and Sound - Proof of concept on human mimicking doll falls", IEEE Trans. on Biomedical Eng., 56(12):2858-2867, 2009.

[20] Zhuang, X., Huang, J., Potamianos, G. and Hasegawa-Johnson, M., "Acoustic Fall Detection Using Gaussian Mixture Models and GMM Supervectors", In Proc. ICASSP-2009, Taipei, Taiwan, 69-72, 2009.

[21] Li, Y., Zeng, Z., Popescu, M. and Ho, K.C., "Acoustic Fall Detection Using a Circular Microphone Array", In Proc. IEEE Int. Conf. Engineering in Medicine and Biology Society EMBS-2010, Buenos Aires, Argentina, 2010.

[22] Miao, Yu., Naqvi, S.M., Rhuma, A. and Chambers, J., "Fall detection in a smart room by using a fuzzy one class support vector machine and imperfect training data", In Proc. ICASSP-2011, Prague, Czech Republic, 1833-1836. 2011.

[23] Zabidi, A., Khuan, L.Y., Mansor, W., Yassin, I.M. and Sahak, R., "Classification of Infant Cries with Asphyxia Using Multilayer Perceptron Neural Network", In Proc. 2nd International Conference on Computer Engineering and Applications ICCEA-2010, Bali, Indonesia, 204–208, 2010.

[24] Aman, F., Vacher, M., Rossato, S. and Portet, F., "In-Home Detection of Distress Calls: The Case of Aged Users", In Proc. INTERSPEECH-2013, Lyon, France, pp. 2065-2067, 2013.

[25] Vacher, M., Lecouteux, B., Istrate, D., Joubert, T., Portet, F., Sehili, M. and Chahuara, P., "Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home", In Proc. INTERSPEECH-2013, Lyon, France, 2062-2064, 2013.

[26] Wang, D. and Brown, G., "Computational Auditory Scene Analysis: Principles, Algorithms and Applications", Wiley-IEEE Press, 2006

[27] Temko, A. and Nadeu, C., "Acoustic event detection in meeting-room environments", Pattern Recognition Letters, 30:1281-1288, 2009.

[28] Vacher, M., Istrate, D., Besacier, L., Castelli, E. and Serignat, J., "Smart audio sensor for telemedicine", In: Proc. Smart Object Conference, 15-17, 2003.

[29] Stäger, M., Lukowicz, P., Perera, N., Büren, T., Tröster, G. and Starner, T., "Sound button: Design of a low power wearable audio classification system", In Proc. IEEE International Symposium on Wearable Computers, 12-17, 2003.

[30] Jianfeng, C., Jianmin, Z., Kam, A. and Shue, L., "An automatic acoustic bathroom monitoring system", In Proc. IEEE Int. Symp. on Circuits and Systems, 1750-1753. 2005.

[31] 7th International Workshop on Multimodal Interfaces eNTERFACE'11: Online: http://enterface11.zcu.cz, accessed on 20 June 2014.

[32] Rosten, E. and Drummond T., "Machine learning for high-speed corner detection", In Proc. 9th European Conference on Computer Vision ECCV-2006, LNCS 3951, Graz, Austria, 430-443, 2006.

[33] Calonder, M., Lepetit, V. and Fua, P., "BRIEF: Binary Robust Independent Elementary Features", In Proc. ECCV-2010, Springer LNCS 6314, Crete, Greece, 778-792, 2010.

[34] van de Sande, K., Gevers, T. and Snoek, C., "Color descriptors for object category recognition", In Proc. European Conference on Color in Graphics, Imaging and Vision, 2:378–381, 2008.

[35] Rabiner, L. and Juang, B., "Speech Recognition", In J. Benesty, M. M. Sondhi, & Y. Huang [Ed], Springer Handbook of Speech Processing. Springer, New York. 2008.

[36] Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D. and Ronzhin A., "Large vocabulary Russian speech recognition using syntactico-statistical language modeling", Speech Communication, 56(1): 213-228, 2014.

[37] Bradski, G. and Kaehler, A., "Learning OpenCV: Computer vision with the OpenCV library", O'Reilly Media, 2008.

[38] Bernardin, K. and Stiefelhagen, R., "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics", EURASIP Journal on Image and Video Processing, ID 246309, 2008.

[39] Karpov, A., Kipyatkova, I., Ronzhin, A. "Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis", In Proc. INTERSPEECH-2011, Florence, Italy, 3161-3164, 2011.