IEEE ICCE 2014

2014 IEEE Fifth International Conference on Communications and Electronics (ICCE)

Pullman Beach Resort Da Nang
July 30 - August 1, 2014

IEEE

WELCOME

Welcome to the 2014 IEEE Fifth International Conference on Communications and Electronics (IEEE ICCE 2014) integrated in a USB version.

ICCE is becoming a reputable biennial international conference series in the scientific community on the areas of Electronics and Communications recently. Following the past successful conferences, the fifth IEEE ICCE 2014 looks for significant contributions to various topics in communication engineering, networking, microwave engineering, signal processing and electronics engineering. The conference will also include tutorials, workshops, and technology panels given by world-class speakers.

At the conference, two hundred and twenty eight (228) papers from more than 30 countries have been submitted. Among these submissions, ninety six (96) regular full papers, which will be submitted for inclusion into IEEE Xplore and twenty nine (29) poster papers have been accepted for presentation and will be organized in 18 regular and 2 special sessions. The Opening Session will host 4 keynotes. Four tutorials offered by the conference widely cover the most interesting topics on electronics and communications engineering, whose issues related to Quality of Mobile Multimedia Experience, Immersive Visual Communication with Depth, Crowdsourcing, and Free Space Materials Characterization.

The technical program focuses on hot topics on the fields of Communications Networks and Systems, Signal Processing and Applications, Microwave Engineering, and Electronic Systems. Besides, the special sessions on Crowdsourcing and Crowdsourcing Applications, and on Information Hiding and Security in Communications: Recent Developments have been added to the technical program of this event.

Beside the printed proceeding, this edition in USB is designed for readers to locate the papers by session or authors. Papers are originated as electronic files and were converted to Adobe Acrobat PDF file format for a crossplatform access. Even though the viewing quality on your monitor may vary, all papers have been printed clearly.

Enjoy your reading through!

IEEE Catalog Number: CFP1416B-USB
ISBN: 978-1-4799-5050-8

TABLE OF CONTENTS

Abnormal event detection using multimedia information for monitoring system

Do-Dat Tran, Thi-Lan Le, Thi-Thanh-Hai Tran
International Research Institute MICA
HUST – CNRS/UMI-2954 – Grenoble INP, Hanoi University of Science and Technology
Hanoi, Vietnam
{do-dat.tran, thi-lan.le, thanh-hai.tran}@mica.edu.vn

Abstract**—** The health smart home has been become an attractive domain for both research and industrial. One of the main issues in smart home is abnormal event detection that allows alarming and making assistance as soon as possible. In this paper, we analyze and exploit multimedia information (image and audio) to resolve the problem of abnormal event detection. The interested abnormal events of our study include falling, lying motionless on the floor, staying in rest room, being out door in long time, abnormal speech (e.g. screaming, shouting) and abnormal non-speech (e.g. breaking and falling sounds). The experimental results evaluated by false alarm rate and sensitivity show that audio and image information complement each other and the use of both information allows detecting a large number of interested abnormal events which can be used in monitoring system for healthcare application.

Keywords—abnormal event; detection; classification; tracking; image; video; sound; multimedia;

I. INTRODUCTION

Given the development of micro-electronic technology, information technology and automation technology, smart homes, smart buildings more and more widely have been developed and built. The development has been extended into other domains, and health care is a domain that has much interest in of researchers. The health smart home concept is a promising and cost-effective way of improving access to home care for the elderly and disabled. Many research and development projects are ongoing, funded by international and governmental organizations. We can cite here some research projects in which multimedia information are processed for monitoring and event detection tasks [1]–[5]. In [1], [5], cameras were used to automatically detect and recognize activities, abnormal events such as falling of people living in house. In studies [4], [6], [7] sound and speech recognition were applied to detect audio events which can be useful to elderly or dependant people in case of problem.

In our study, we would like to deal with more abnormal events number that could happen to elderly, patients, disabled by exploiting multimedia information (video and audio signals). The developed system is intended to detect accidents such as falling, being motionless and accidents which would probably produce abnormal sounds.

The proposed system then is composed of two main modules: video based abnormal event detection and audio based abnormal event detection. Our main contributions are three-fold: i) we have created a publicity dataset which contains video and audio stream of six abnormal events. This dataset could be used to encourage research on abnormal event detection based on multimodal information; ii) we proposed a method for video based abnormal detection that combines motion template and localization information; iii) we made a comparative study on audio classification.

The remaining of this paper is organized as follows. In the section II, we make an overview of our proposed system for abnormal event detection. We describe in detail two main modules for abnormal event detection based on video and audio in section III. Section IV shows experimental results. Section V concludes and gives ideas for future works.

II. ABNORMAL EVENT DETECTION SYSTEM OVEVIEW

In daily life elderly, dependant people (with physical disabilities for instance) and patient sometime who live alone can meet an accident in their home that could harm them. That is why many research and development projects have been conducted in order to automatically detect abnormal and emergency events.

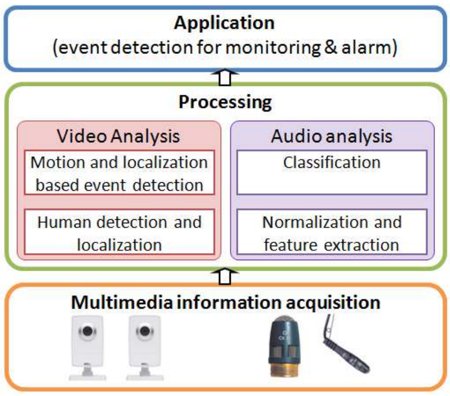

Fig. 1. System architecture.

In the framework of our study, we consider six abnormal events: (1) *falling* (2) *lying motionless on the floor,* (3) *staying in rest room,* (4) *being out door in long time,* (5) *abnormal speech (e.g. screaming, shouting)* and (6) *abnormal non-speech (e.g. breaking and falling sounds).*

The Fig. 1 shows the architecture of our proposed system. It contains three main layers: (1) **Acquisition**: it captures visual and audio signal from cameras and microphones equipped in the environment; (2) **Processing**: it contains video

978-1-4799-5051-5/14/$31.00 ©2014 IEEE

analysis and one for audio analysis modules. All core processing / analyzing are performed at this layer; (3) **Application:** this layer is generic, it allows all kinds of video / audio based application. In the context of this work is the event detection.

III. Abnormal Event Detection Based On Video/Audio Information

In the context of our work, we try to explore both visual and audio information to recognize the events of interest. In this section, the capacity of each modality (video and audio) for abnormal event detection is analyzed. We first present our propositions for each modality separately. At the end of this section, we will give some discussions on the possibility to combine audio and visual information

A. Video Based Event Detection

Among the 6 abnormal events of interest, we analyze in this section the capacity of using video for recognizing 4 events that are (1) *falling (from the bed or during walking)*, (2) *lying motionless on the floor*, (3) *staying too long in the rest room*, (4) *being out of the room too long*. To do this, we propose a method that combines motion templates and localization information. In this section, we first present human detection, tracking and localization module. Then, we describe abnormal event detection method.

1) Object detection, tracking and localization

To be able to detect the presence of people and their location in 2D image plane as well as 3D room space (real world) we carry out three main tasks: background modeling, human detection and human tracking.

a) Background Modeling

The object detection assures an automatic initialization of the tracker as well as provides observations for data association. In our context, the camera is fix but scene can contain moving background (like waving curtain) and illumination variations. To handle to this, we proposed to use segmentation technique based on codebook technique [8]. This method is experimentally shown to be more efficient in time/memory and precision, meaning some moving elements in the background is considered as background. For technical detail, see the original paper [8].

b) Moving Object detection

Once the background model is built, given each video frame, the moving objects detection is carried out by differencing the current image with the background model. To remove noises, we threshold the different image. Morphological operators are then applied, followed by connected component analysis to group pixels into blobs.

We can observe in Fig. 2 that using background subtraction, people are detected but their localization is not really perfect: the bounding boxes are mostly bigger than human. Sometime, a part of background is considered as a false alarm. To remove this kind of false alarms, first, we extend all bounding boxes then apply HOG-SVM [9] based human detector on each extended bounding box for verification. We notice also that by this way, we can remove some false alarms occurred when we apply HOG-SVM detector on the whole image.

Fig. 2. Detection results obtained from differencing current frame with background frame.

Beside, to avoid missed detection caused by HOG-SVM, we will keep detection that satisfy conditions to be still a human (ratio between width and height, percentage of foreground pixels and the bounding box) to keep tracking longer.

(a) (b)

Fig. 3. (a) Detection results (black rectangles) by applying HOG-SVM on whole image. (b) Detection results (red rectangle) by applying HOG-SVM on the extended region (green rectangle). The false alarm in (a) (smaller rectangle) is now removed in (b), the localization of human is more precise.

c) Tracking

For tracking human, we propose to use the traditional Kalman filter that has been shown to be good enough in lot of surveillance applications. Observation and process noise are supposed as white noise with Gaussian distributions.

(a) (b)

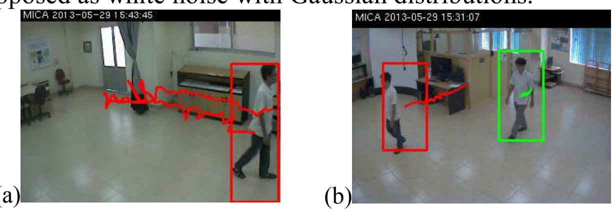

Fig. 4. (a) Tracking results: (a) Bounding box represents the current location of the human, the red line is his trajectory; (b) Each person and his trajectory are represented by a color in multiple human tracking.

In our case, we would like to build a multiple human tracking, so we need to do a more complex track – observation association. The association between a track and an observation will be selected based on a match measure that is the Euclidian distance between two HOG descriptors. If a track does not find an observation (missed detection), we keep this track in several frames until it find an observation in the next frame. After important missed observations, we delete this track. For all remaining observations, we create new tracks. Fig. 4 shows some examples of human detection and tracking.

2) Abnormal event detection

Among these four events, the first event (fall) has attracted many works in the computer vision community. According to [10], this event can be decomposed in four phases. The prefall phase corresponds to daily life motions, with occasionally sudden movements directed towards the ground like sitting down or crouching down. The critical phase, corresponding to the fall, is extremely short. This phase can be detected by the movement of the body toward the ground or by the impact shock with the floor. The post-fall phase is generally characterized by a person motionless on the ground just after the fall. It can be detected by a lying position or by an absence of large motion. A recovery phase can eventually occur if the person is able to stand up alone or with the help of another person. Fig. 5 illustrates different phases of a fall out of bed.

Fig. 5. Illustration of different phases of the event fall out of bed.

There are a number of works have been proposed for fall event detection. These works can be divided into two categories. The works belonging to the first category try to model and to recognize the fall events by using finite state machine, HMM (Hidden Markov Model) [11] while the second compute the motion templates such as MHI (Motion History Image) [12]. In this paper, for the fall event, inspired the work of [13], we propose a fall event algorithm combining both object localization output and MHI.

Among 4 events of interest, the third and the fourth events ("*staying too long in the rest room*" and "*being out of the room too long*") are inferred directly from the output of human localization. The first and the second events ("*falling from the bed or during walking*" and "*lying motionless on the floor*") are recognized by the algorithm illustrated in Fig. 6.

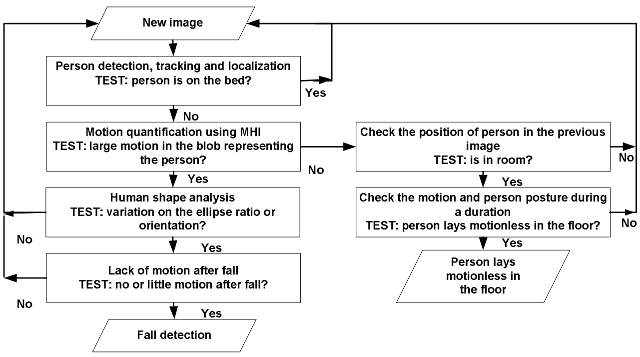

Fig. 6. Recognition algorithm for two events ("Falling from the bed or during walking " and "Lying motionless on the floor")

Since the first event is described by a large motion while the second is represented by a small motion. We use MHI to estimate person motion as method of [13]. However, the main difference of our method and that of [13] is the "*Person detection, tracking and localization*" step. Based on this step, we verify the hypothesis: "*is the person on the bed?*". The result of this verification allows to remove false detection because if person lays motionless in the bed, this is normal situation. Moreover, if the system knows that the person is on the bed, it does not need to do fall event detection.

The human shape analysis is done by approximating a person by an ellipse defined by its center, its orientation and the length of its major and minor semi-axes. The approximated ellipse gives us information about the shape and orientation of the person in the image. We compute the orientation standard deviation and standard deviation of ratio between major and minor semi-axes in duration of time (e.g. 1s) and base on the following observations: (1) If a person falls perpendicularly to the camera optical axis, then the orientation will change significantly and orientation standard deviation will be high. If the person just walks, orientation standard deviation will be low; (2) If a person falls parallelly to the camera optical axis, then the ratio will change and the standard deviation of ratio between major and minor semi-axes will be high. If the person just walks, this measure will be low.

B. *Audio based Event Detection*

Sound classification problem has been studied for a long time and applied widely in many applications [3], [4], [6], [7], [14]–[17]. Early study on speech/non-speech classification were presented in [14], in which authors concentrated in discriminating speech from music on broadcast. In the studies [3], [4], [6], [16] sound classification were applied for building abnormal sound detection and classification systems which would be used for surveillance applications. And in these studies abnormal sounds are defined as those coming from emergency situations, such as scream, groan, cry of patient, or sounds of falls or break of objects

The problem of detecting abnormal sounds could be solved by classifying sounds. Captured sound is firstly classified into speech and non-speech. At this step, there should be a speech/non-speech discriminator. Secondly, in turn, speech sound is classified into normal speech and abnormal speech; and non-speech sound is classified into normal non-speech and abnormal non-speech. At this step, it requires two other discriminators: normal/abnormal speech, and normal/abnormal non-speech. For the purpose of study and developing the three discriminators, it is needed to build four sound corpora: normal speech, abnormal speech, normal non-speech, and abnormal non-speech.

1) *Corpora construction*

a) *The Normal Speech*

Normal speech is the sound coming from patient's throat when he/she is in good health. From VNSpeechCorpus [18], which is a Vietnamese speech corpus used for automatic speech recognition, a normal speech corpus is extracted and used for our study. This corpus is composed of speech signals of 10 speakers (5 male and 5 female) with the age from 20 to 40. Besides, we built a new supplemental speech corpus in simulated room. This supplemented corpus includes speech signals from 10 new speakers (7 male and 3 female) aging from 21 to 37.

b) *The Abnormal Speech*

Abnormal speech is the sound coming from patient's throat when he/she is in bad health or in a serious situation, e.g. fall, faintness, sick, etc. Due to the difficulties of building this type of corpus in real condition, we carried out building the abnormal speech corpus by recording the sound in the simulated room. The ten speakers (7 males and 3 females) were asked to make sound like scream, shout and cough. Besides, we collected sounds corpus (scream, shout, cough, cry, moan) from internet, films, and sound effect CDs.

c) The Normal Nonspeech

Normal non-speech is the sound coming from things or sounds of normal life. Those audio signals could originate from, for instance, doors (closed), door bell, chairs (dragged), drawers (opened and closed), liquid (pour in and out), glasses, cups, dishes, bowls, thermos flask, etc. In our study the patient's room of the 108 Hospital in Hanoi, Vietnam, has been considered as the pattern room. We collected things from the patient's room, such as bed, chair, table, cupboard, glasses, cups, etc., then we recorded their sounds or sounds from their clashes. The corpus were recorded in the recording studio and the simulated room. We also recorded sounds of objects in bathroom, such as shower, faucet, flushing of toilet, etc. To enrich this corpus, we also collected those sounds from internet and sound effect CDs.

d) The Abnormal Nonspeech

Abnormal non-speech is the sound coming from things in monitored room when person is in bad health. For instance, if the patient falls, he/she will probably break a glass, then the sound of a broken glass is considered as an abnormal non-speech. And when he falls, he will probably make a chair tumbled down, in this case, the sound of a tumbled chair is another signal of this corpus. In our studio, we tumble chairs, break glasses and bowls and cups, etc., to record sounds of abnormal non-speech. Then once again, we enlarge the corpus by collecting more from internet and sound effect CDs.

The obtained sound corpora is shown at TABLE I. All signals are mono, recorded at 16 kHz and quantized at 16 bit.

TABLE I. SOUND COPORA

Corpus	Number of signals	Duration (s)	Size (MByte)
Normal Speech	323	3039	92,7
Abnormal Speech	403	907	26,1
Normal Non-speech	2,119	2285	74,7
Abnormal Non-speech	812	761	23,4

2) Sound based abnormal event detection

Process of developing three discriminators are similar, two following steps are performed.

Selecting a feature set for a discriminator: we start with 11 features which were applied in [3], [7]: zero-crossing rate (ZCR), f0, centroid, energy, roll-off, band energy ratio (BER), bandwidth, linear spectral pairs (LSP), Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), and pitch. Then PCA (Principal Component Analysis) is applied to select the most suitable features for each discriminator. A set of features will be selected if their total variability exceeds a certain threshold (95% for instance).

Choosing a classification model for the discriminator: The selection criterion is discrimination ratio, which is the ratio between correctly discriminated samples and tested samples. In this study, artificial neuron network (ANN), Gaussian mixture model (GMM), decision tree (DT), and support vector machine (SVM) are chosen as considered classification models.

a) The Speech/Nonspeech Discriminator

The all four corpora are used to build and to evaluate this discriminator. From these databases we extract 11 features. The application of PCA on these features shows that the 9 features MFCC, ZCR, f0, centroid, energy, roll-off, BER, bandwidth, and LSP have a total variability of 95%. Therefore, they are chosen to form the feature set of this discriminator. This feature set is used to evaluate the performance of each classification model in discriminating speech/non-speech signal.

b) The Normal/Abnormal Speech Discriminator

The normal speech and abnormal speech databases are used to construct this discriminator. To find appropriate discriminant features, PCA is applied. Experiments prove that a set of 7 features (ZCR, f0, centroid, energy, roll-off, bandwidth, and pitch) can satisfy 96% of the total variability. Basing on this set, the 4 models are tested to distinguish between normal and abnormal speech.

c) The Normal/Abnormal Nonspeech Discriminator

Experiments for this discriminator are based on the normal non-speech and the abnormal non-speech corpora. The PCA was also applied; the 7 features ZCR, f0, centroid, energy, roll-off, bandwidth, and LSP occupy 95% of the total variability. In the next stage, this set is then fed to all 4 models to find the best.

TABLE II. DISCRIMINATION RATIO [%] OF THREE DISCRIMINATORS WITH TWO CLASSIFICATION MODELS

Model	Parameter	Speech / Non-Speech	Abnormal /Normal Speech	Abnormal /Normal Non-Speech
ANN	40 hidden neurons	96.00	93.46	90.02
DT [17]	Deviance criterion	94.80	91.96	94.88

The preliminary results of our study presented in [17] showed that DT model gives better discrimination ratio for all three discriminators in comparison to three remaining models. Thus, we had chosen DT model to construct three discriminators. However, on the one hand, when implementing three discriminators using obtained DT model in simulated room we found that the execute time is important because the size of DT model is quite large. On the other hand, by more deeply study with ANN, we can get the discrimination ratio closed to that of DT model TABLE II. , and the run time of three ANN based discriminators is much better DT based discriminators. Therefore, the ANN model is finally chosen for developing sound classification system in our study.

C. Discussion on combination of audio and video abnormal event detection

In [19] three types of inter-modal relations are presented, they are trigger, integration and collaboration relations. From the 6 considered abnormal events in our study, we found that there is only the falling event which can produce abnormal video and audio information simultaneously. In case of falling event, when a person falls, he/she can shout or scream or breaking sound is generated. Thus, the trigger relation which models the triggering of one modality processing by the detection of an event in another modality was considered. Based on the algorithm that is applied for detecting falling event and the experimental implementation, we found that the trigger relation did not give a good result because the delay of decision from video information is quite large; it is not logical if the system uses this decision to trigger the abnormal sound detection module. Consequently, the two abnormal detection modules are implemented and worked independently, the processed information is sent to central processing module in parallel.

IV. RESULTS EXPERIMENTS

A. *Experiment Scenario and Evaluation Measure*

To evaluate our system, we need to setup environment and define scenarios. We carry out experiments in two environment conditions: (1) at the show room of MICA Institute – simulated room (2) at patient room at 108 Hospital in Hanoi, Vietnam. TABLE III. gives some information of our testing environment. Both two environments have similar structure, as illustrated in Fig. 7. In these environment, we equip two IP cameras (AXIS M1054) and two high quality microphones (AKG CK31) to capture visual and audio information.

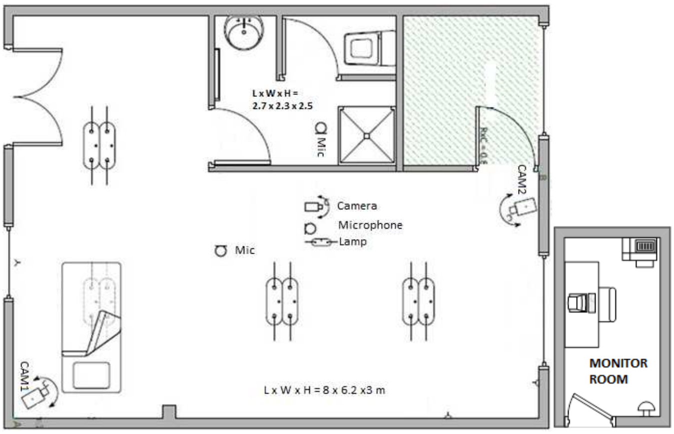

Fig. 7. The health smart room with the audio video captures, detected abnormal events will be alarmed at the monitor room.

The number of participants to the experiments are 20 (10 male and 10 female), and are 10 (5 male and 5 female) aging from 25 to 40 years old for the first condition and the second condition respectively. Subjects are asked to play the following scenario 5 times. In the scenario, the person is asked to (1) Enter to the room; (2) walk and say something; (3) Sit on the bed; (4) Lay on the bed and scream; (5) Lay motionless on the bed; (6) Fall from the bed and shout; (7) Lay motionless on the floor; (8) Get up and walk to the table; (9) Go toward the rest room and say something; (10) Get into the rest room; (11) pour water; (12) drop an stainless steel tray or a plastic chamber-pot; (13) Stay long in the rest room; (14) Get out with a plastic vase; (15) Fall on the floor and drop the plastic vase; (16) Get out the room. We perform our system in the computer with the following configuration: Intel(R) Core(TM) i5-2520M CPU @ 3.2 GHz x 2, RAM 4GB

TABLE III. DESCRIPTION OF THE EXPERIMENTAL ENVIRONMENT

	Simulated room	Patient room
Size (L x W x H)	(9.2 x 8.8 x 3.0) m	(8.0 x 6.2 x 3.0)m
Main door	01	01
Windows	02	02
Toilet inside	01	01
Objects	Bed, medical cabinet	Bed, medical cabinet
Lighting condition	Neon and daylight through windows and door	Neon and daylight through windows and door
Video/audio captures	2 IP cameras AXIS M1054, 2 microphones AKG CK31	2 IP cameras AXIS M1054, 2 microphones AKG CK31

Fig. 8. A corner of the room with the mounted camera and micophone

We measure two criteria below:

$$F.A.R = \frac{FP}{TP+FP} \; ; \; Sensitivity = \frac{TP}{TP+FN} \qquad (1)$$

where TP (True Positive) is the number of correct events detected; FP (False Positive) is the number of wrong events detected, and FN (False Negative) is the number of lost events. The smaller F.A.R and the greater Sensitivity are, the better system is.

B. *Experimental results*

The obtained results of two experiments (simulated and hospital room) are shown in Tab. IV and Tab. V respectively. The total number of event for the first experiment at each times is 200 (40 Event 1, 40 Event 2, 20 Event 3, 20 Event 4, 40 Event 5, 40 Event 6) while the total number of event for the second experiment is 100. The experimental results show that our system is capable to detect six events of interest with a high value of sensitivity and low value of false alarm rate. The overall result obtained in the simulated room is better than that of hospital room. The reason is that our training databases are collected in the simulated room therefore the chosen parameters of the model are more suitable for the simulated room condition.

Among 6 events, "*falling from the bed or during walking*", "*lying motionless on the floor*", "*abnormal speech*" and "*abnormal non-speech*" obtains the best result in term of Sensitivity. The value of F.A.R obtained for these events is also high. This means that our system detect these events whenever they happen. But, it also has several false positives. This result is acceptable in the context of surveillance system for peoples with special need because these events are important and the miss of these events can cause the major health problem. The recognition results of "*staying too long in the rest room*" and "*being out of the room too long*" are relatively good. The main reason is that these events are recognized by using the results of object localization and tracking module. The doors of our testing room are transparent. Therefore, the illumination effect cause some false detections.

TABLE IV. THE OBTAINED SENSITIVITY AND FALSE ALARM RATE OF 6 EVENTS OF INTEREST WITH 20 SUBJECTS (EVENT 1: FALLING FROM THE BED OR DURING WALKING; EVENT 2: LYING MOTIONLESS ON THE FLOOR; EVENT 3: STAYING TOO LONG IN THE REST ROOM; EVENT 4: BEING OUT OF THE ROOM TOO LONG; EVENT 5: ABNORMAL SPEECH; EVENT 6: ABNORMAL NON-SPEECH)

Events	1	2	3	4	5	6
Sensitivity at times #1	0.88	0.98	0.75	0.80	1	0.84
F.A.R at times #1	0.00	0.11	0.00	0.08	0.09	0.08
Sensitivity at times #2	0.88	0.9	0.75	0.85	0.91	0.88
F.A.R at times #2	0.07	0.15	0.00	0.13	0.02	0.18
Sensitivity at times #3	0.93	0.95	0.80	0.85	0.98	0.92
F.A.R at times #3	0.02	0.02	0.00	0.06	0.01	0.11
Sensitivity at times #4	0.93	0.98	0.95	0.80	1	0.85
F.A.R at times #4	0.00	0.03	0.00	0.00	0.07	0.2
Sensitivity at times #5	0.93	0.95	0.80	1.00	0.96	0.83
F.A.R at times #5	0.00	0.02	0.00	0.05	0.08	0.13
Average Sensitivity	**0.91**	**0.95**	**0.81**	**0.86**	**0.97**	**0.86**
Average F.A.R	**0.018**	**0.067**	**0.00**	**0.064**	**0.054**	**0.14**

TABLE V. THE OBTAINED SENSITIVITY AND FALSE ALARM RATE OF 6 EVENTS OF INTEREST WITH 10 SUBJECTS

Events	1	2	3	4	5	6
Sensitivity at times #1	0.95	0.8	1	0.5	1	0.93
F.A.R at times #1	0.1	0.083	0	0	0.033	0.083
Sensitivity at times #2	0.95	0.8	0.9	0.8	0.95	1
F.A.R at times #2	0.1	0	0	0	0.083	0.083
Sensitivity at times #3	0.95	0.95	0.6	0.6	0.9	0.8
F.A.R at times #3	0.033	0.116	0	0	0.033	0.083
Sensitivity at times #4	0.75	0.75	0.8	0.6	0.97	0.92
F.A.R at times #4	0.067	0.083	0.05	0	0.05	0
Sensitivity at times #5	0.85	0.95	0.8	0.85	0.88	0.95
F.A.R at times #5	0.067	0.033	0.05	0.05	0.058	0.05
Average Sensitivity	**0.89**	**0.85**	**0.82**	**0.67**	**0.94**	**0.92**
Average F.A.R	**0.073**	**0.063**	**0.02**	**0.01**	**0.051**	**0.059**

V. CONCLUSIONS

In this paper, we have presented an abnormal event detection using multimedia (audio and video) for monitoring system. We have analyzed the environment setup, introduced audio based abnormal event detection and video based abnormal event detection. We have performed also two experiments in two different conditions. The obtained results show that our system is capable to detect 6 events of interest with a high value of sensitivity and low value of false alarm rate. In the future, we would like to improve this system by (1) improve object detection tracking and localization method by taking into account illumination change and object occlusion; (2) extend the system so that it can detect a larger number of events of interest

ACKNOWLEDGEMENTS

This study was done in the framework of the International cooperation project 10/2011/HĐ-NĐT and was supported by the Research Grant from Vietnam's National Foundation for Science and Technology Development (NAFOSTED), No. FWO.102.2013.08.

REFERENCES

[1] Abowd G., Mynatt E., and Rodden T., "The human experience of ubiquitous computing," *IEEE Pervasive Comput.*, vol. 1, pp. 48–57, 2002.

[2] Bonhomme S., Campo E., Estève D., and Guennec J., "Prosafe-extended, a telemedicine platformto contribute tomedical diagnosis," *Telemed. Telecare*, vol. 14, no. 3, pp. 116–119, 2008.

[3] D. Istrate, M. Vacher, E. Castelli, and C.-P. Nguyen, "Sound processing for Health Smart Home," in *International Conference on Smart homes and health Telematic*, 2004.

[4] D. Istrate, J. Boudy, H. Medjahed, and J.-L. Baldinger, "Medical remote monitoring using sound environment analysis and wearable sensors," in *BioMED 2009*, Heraklion, 2009, pp. 517–532.

[6] Guillaume Sacco, Véronique Joumier, et al, "Detection of activities of daily living impairment in Alzheimer's disease and mild cognitive impairment using information and communication technology," *Clin. Interv. Aging*, vol. 2012:7, pp. 539 – 549, Dec. 2012.

[6] Michel Vacher, Anthony Fleury, Francois Portet, Jean-Francois Serignat, and Norbert Noury, "Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living," in *New Developments in Biomedical Engineering*, InTech, 2010, p. 714.

[7] J. Chen, "Bathroom Activity Monitoring Based on Sound," in *PERVASIVE 2005*, Munich, 2005, pp. 47–61.

[8] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time Foreground Background Segmentation Using Codebook Model," *Real-Time Imaging Spec. Issue Video Object Process.*, vol. 11, no. 3, pp. 172–185, Jun. 2005.

[9] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, vol. 1, pp. 886 – 893.

[10] Noury, N., Rumeau, P., Bourke, A.K., ÓLaighin, G., and Lundy, J.E., "A proposal for the classification and evaluation of fall detectors," *IRBM*, vol. 29, no. 6, pp. 340–349, Dec. 2008.

[11] Vinay Vishwakarma, Chittaranjan Mandal, and Shamik Sural, "Automatic Detection of Human Fall in Video," presented at the The 2nd international conference on Pattern recognition and machine intelligence 2007, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 616–623.

[12] Bobick, A.F. and Davis J.W., "The recognition of human movement using temporal templates.," *IEEE Trans. Pattern Anal. Mach. Intell. 2001*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[13] C. Rougier, Alain St-Arnaud, Jacqueline Rousseau, and Jean Meunier, "Video Surveillance for Fall Detection," in *Video Surveillance*, 2011, pp. 357–382.

[14] J. Saunders, "Real time discrimination of broadcast speech/music," in *ICASSP' 96*, Atlanta, Georgia, USA, 1996.

[15] S. Shafiee, "A two-stage speech activity detection system considering fractal aspects of prosody," *Pattern Recognit. Lett.*, vol. 31, no. 9, pp. 936–948, 2010.

[16] Cheung-Fat Chan and Eric W.M. Yu, "An abnormal sound detection and classification system for surveillance applications," in *EUSIPCO 2010*, Aalborg, 2010, pp. 1851–1855.

[17] C. P. Nguyễn and Đ. Đ. Trần, "Sound classification for event detection Application into medical telemonitoring," in *Comantel'13*, Ho Chi Minh city, Vietnam, 2013, pp. 330–333.

[18] V. B. Le, D. D. Tran, E. Castelli, L. Besacier, and J.-F. Serignat, "Spoken and written language resources for Vietnamese," in *LREC'04*, Lisbon, Portugal, 2004, pp. 599–602.

[19] T. Martin, "Vers une reconnaissance multimodale du texte et de la parole pour l'analyse de documents vidéos pédagogiques," PhD thesis, La Rochelle University, 2009.