Near real-time human silhouette and movement detection in indoor environments using fixed cameras

A. Christodoulidis
University of Central Greece,
Department of Computer Science and Biomedical Informatics,
Lamia, Greece

+30 22310 66901
axristodoulidhs@ucg.gr

K.K. Delibasis
University of Central Greece,
Department of Computer Science and Biomedical Informatics,
Lamia, Greece

+30 22310 66901
kdelibasis@yahoo.com

I. Maglogiannis
University of Central Greece,
Department of Computer Science and Biomedical Informatics,
Lamia, Greece

+30 22310 66931
imaglo@ucg.gr

ABSTRACT
The automated human behavior modeling is highly desired in the context of an assistive environment. In this paper, we describe a software video processing and analysis system to assist the near real time detection of human activity. The video data are acquired indoors from fixed cameras in the living environment. The proposed system uses image-processing techniques to segment the human figure from the background, suppress the appearance of its shadow and detect the path and velocity of its motion. Detail performance measurements of the proposed algorithm are given in terms of execution time. Initial results are presented for a small number of video sequences.

Categories and Subject Descriptors
Performance, Design, J.3 LIFE AND MEDICAL SCIENCES, *Medical information systems*

General Terms
Algorithms, Measurement, Performance, Reliability, Experimentation, Security, Human Factors.

Keywords
Human activity detection, video processing, background subtraction, shadow suppression.

1. INTRODUCTION
The advances in video acquisition, communication and processing have enabled the development of many applications and systems related to human activities. Such pervasive human-centered systems are able to understand the human state (identity, emotions and behavior) in assistive environments using audiovisual and biological signals. Their goal is to offer services such as support for the aged/disabled/chronic patients, detection of critical situations from audiovisual content, biosignals and neurophysiology analysis for the detection of pathology (e.g. Alzheimer's disease, epilepsy, etc), as well as for treatment follow-up.

The monitoring of human physiological data, in both normal and abnormal situations of activity, is vital also for the purpose of emergency event detection, especially in the case of elderly people living on their own. Special interest is paid in the detection of the severity of the case that can indicate injury level and assistance request type. Several techniques have been proposed for identifying such distress situations using either motion, audio or video data from the monitored subject and the surrounding environment.

The paper presents the design and the initial implementation of a video-processing module that may be used for patient activity interpretation and emergency recognition in cases like elder falls. The module utilizes video data captured from fixed cameras in the user's living environment. Appropriate image- and video-processing techniques are applied in order to segment the human figure from the background, suppress the appearance of its shadow and detect the path and velocity of its motion. The performance of the implemented algorithms in sample videos is also presented.

The rest of the paper is organized as follows; Section 2 discusses related work in the context of activity recognition using video sequences. Section 3 describes the proposed methodology and details the implemented algorithms for human figure segmentation and correction, while Section 4 presents the conducted experiments along with the corresponding results. Finally, Section 5 concludes the paper.

2. RELATED WORK
Although the concept of activity recognition using video data is not so new, it is still an open issue that preoccupies several computer vision research groups worldwide [1] – [6]. Information regarding the movement and activity is frequently acquired through visual tracking of the patient's position, while low level features are extracted and combined with existing libraries of image patterns. For instance in [1] authors present a logical language called Probabilistic Activity Description Language (PADL) in which users can specify activities of interest. In [2]

PETRA12, June 6 - 8, 2012, Heraklion, Crete, Greece.

low-level features are extracted from a video stream during a short time period for describing human actions, and the extracted sequence of the low-level features is used for describing human activity. A simple surveillance camera is utilized in [3] for measuring motion quantity and detecting active state. The ultimate goal is to find the detecting point of suspicious activity, and estimation of the degree of the suspicious activity. In [4] a human detection and tracking algorithm for indoor environments, while at the feature level an adaptive learning method to estimate the physical location and moving speed of a person is presented. This information is exploited at the action level as a hierarchical decision tree and dimension reduction methods for human action recognition.

In [1] overhead tracking through cameras provides the movement trajectory of the patient and gives information about user activity on predetermined monitored areas. Unusual inactivity (e.g., continuous tracking of the patient on the floor) is interpreted as a fall. Similarly, in [6] omni-camera images are used in order to determine the horizontal placement of the patient's silhouettes on the floor (case of fall). Success rate for fall detection is declared at 81% for the latter work.

A different approach for detecting user activity is the use of non-visual channels for capturing information [7]. Such systems integrate devices like accelerometers, gyroscopes and contact or proximity sensors. Sound signals collected by the environment may also be utilized. In this work we focus only on the video channel for activity processing and analysis, however the proposed algorithms can be easily integrated into more complex systems, which exploit additional information channels.

3. MATERIALS AND METHODS

3.1 Overall approach

The aim of our research is to implement efficient algorithms for assisting video-based human activity detection systems. A generalized pipeline for human activity detection is shown in Figure 1(a), where video sequences are segmented and further analyzed, whereas activity detection is also assisted by non-visual channels (sensor inputs) as well as libraries of known activity patterns.

The contribution of this work is on the processing and analysis of the video channel in assistive systems. More specifically, we propose a method for background modelling and human shadow suppression in order to achieve robust and real-time human figure segmentation. Furthermore, we propose a method for extracting important information of the human activity, such as the path of the human centre of mass and its velocity. Qualitative results are presented for a number of monocular video sequences, using an indoor video-camera based on the roof, with a fish-eye lens.

The proposed algorithm may be briefly described as following:

A model of the background is being updated with each current frame. Initial segmentation is performed by subtracting and thresholding the background model from the current frame. If the number of segmented pixels is below a preselected threshold (no human activity detected), then the process is repeated for the next frame. If human activity is detected, then the following steps take place:

- A region of the current frame is selected using the smallest bounding box that encloses all the segmented pixels.
- For the part of the frame within the bounding box, the algorithm for shadow detection is executed and the final segmentation of the human silhouette is performed.
- A two-step process is also applied to calculate the centre of mass of the human.

In case more frames are available in the video sequence, the algorithm is repeated, otherwise, the path of the movement of the human is constructed and the algorithm terminates. The block diagram of the proposed algorithm is illustrated in Figure 1(b).

Figure 1(a). A generic architectural scheme for video-based human activity characterization, while the dashed bounding box indicates the focus of this work

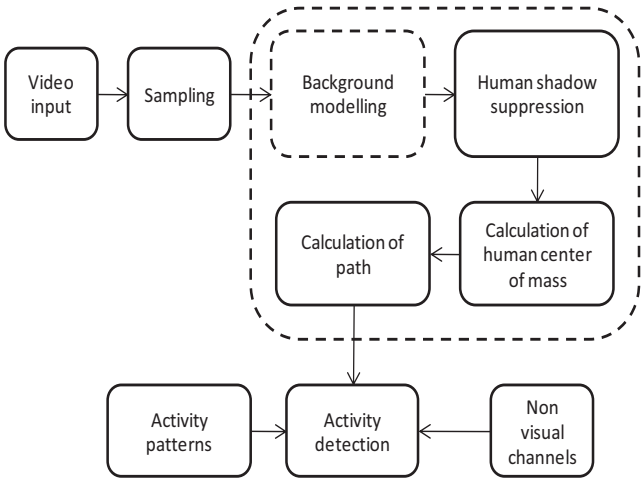
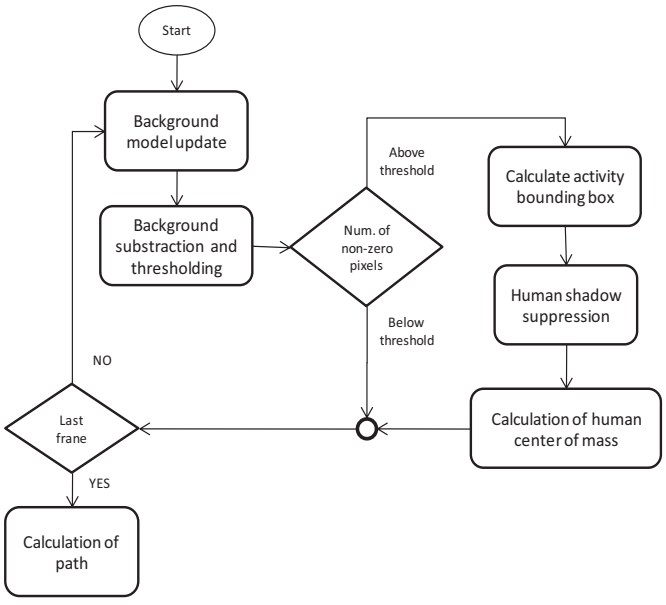


Figure 1(b). Block diagram of the overall proposed algorithm



3.2 Human segmentation

Human segmentation is performed by background subtraction and suppression of the corresponding shadow. These two discrete steps are described in the following subsections.

3.2.1 Background modelling

The human activity, as a fixed monocular camera records it, can be analyzed by segmenting the changing pixels of the human against the background that is considered to be slowly changing with respect to the human motion. Most of the video processing algorithms employ a segmentation technique based on background subtraction. The background however may not be considered constant, due to a number of reasons, including random changes (e.g. background object removing), change of light conditions, video suppression artifacts). A number of techniques have been proposed for background modelling, i.e. constructing a model of the background that is slowly being updated using the values of the current video frame. These methods can be separated in two major categories, the non-recursive and the recursive algorithms. In the literature a number of non-recursive algorithms have been proposed, such as the following.

Frame difference [8]: In this method the background model is, simply, defined as the previous frame. Then simple thresholding is employed in order to extract the foreground. This method is very simple to implement but it is prone to a number of segmentation errors.

Background can be modelled by simple median filtering [9]. A frame buffer is constructed and is filled with the N last frames. Then, the background value of each pixel in the model is computed as the temporal median of the pixel across the accumulated frames in the buffer. The class of recursive background modeling algorithms contains techniques that do not use a buffer to store previous frames. Examples of this class of algorithms include the following:

Running average [8]: The background model BG at time t is updated using a running average of the background and the current frame FR, weighted by the learning rate α.

$$BG_{t+1} = aFR_t + (1-a)BG_t \tag{1}$$

An extension of the aforementioned method is the running Gaussian average that was proposed by Wren, C., et al. [10].

Mixture of Gaussians: Stauffer and Grimson [11] first described the method of background modelling in video sequences using the mixture of Gaussians (MoG). According to this method the values of each pixel are modeled as a number of weighted Gaussian probability distributions. The probability of observing a pixel at time t with value x_t that belongs to the background in the current frame is defined as:

$$P(x_t) = \sum_{i=1}^{K} \omega_{i,t}\eta(x_i, \mu_{i,t}, \Sigma_{i,t}) \tag{2}$$

where η is a Gaussian probability distribution with weight ω, mean value μ, and the covariance matrix Σ, defined as:

$$\eta(x_t, \mu_t, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_t-\mu_t)^T\Sigma^{-1}(x_t-\mu_t)} \tag{3}$$

The model uses a number of K Gaussian distributions, with K between 2 and 5. The weights and the parameters of the mixture of the Gaussian models are estimated using a number of approaches (i.e. [12]).

An appealing method that we employed in this study was the *approximated median filtering* method as described by McFarlane, and Schofield [13]. According to this method, the background model is initialized as an empty frame. For each subsequent frame, if a pixel has a value greater (less) than the corresponding pixel in the background model, then the value of this pixel in the background model is increased (decreased) by one. This recursive approximation does not suffer the computational cost of calculating the temporal median value of all pixels in the current frame. In our implementation we adapted the algorithm using the HSI (Hue, Saturation, Intensity) chromatic domain, since we also utilize the HIS color space for human shadow suppression.

Initially, the captured video was converted from the RGB to the HSI domain using the following equations (Gonzalez et al. [14]).

$$\begin{aligned} Hue &= \begin{cases} \theta, & \theta \in [0,\pi] \\ 2\pi - \theta, & \text{otherwise} \end{cases} \\ Saturation &= 1 - \frac{3\min(R,G,B)}{R+G+B} \\ Intensity &= \frac{1}{3}(R+G+B) \end{aligned} \tag{4}$$

where R, G and B are the components of the RGB model and the angle θ is defined as

$$\theta = \cos^{-1}\left(\frac{1}{2}\frac{(2R-G-B)}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right)$$

Subsequently, the background model is initialized with the first frame. With each new frame the background model is being updated using the method of approximated median filtering as described previously. Finally, the absolute difference between the current frame and the background is calculated and thresholded. The pixels with absolute intensity difference higher than a preselected threshold are considered as belonging to the human figure. The result of the background subtraction is combined with the shadow suppression to complete the human segmentation task at the next step.

The algorithm for the background modelling in video sequences is given below in pseudocode. The symbols BG and FR denote the current background model and the current video frame respectively, whereas the index I indicates the intensity component of the HSI model.

Algorithm: Background modelling

Input: The intensity component of the current frame FR_I

Output: The background model BG_I

```
initialize BG as the 1st frame
for each frame FR
   for each pixel p in FR
        if FRI(p)>BGI(p)
           BGI(p)++
        else if FRI(p)<BGI(p)
           BGI(p)--
```

3.2.2 Human Segmentation and Shadow suppression

The following calculations are executed for the segmentation of the human from video sequences. First, the current background model is subtracted from the current frame and the difference is thresholded to perform an initial segmentation frame *S*. The value of the threshold was determined experimentally and is given in the Section 4 (Results - par. 4.1). If the number of non zero pixels in *S*, is lower than a threshold T_2 (in that case it is assumed that there is no human activity in the frame), no further processing is performed on this frame and the next frame is loaded. Otherwise the processing steps of shadow detection and suppression and the calculation of the centre of mass of the human figure are executed. In order to reduce complexity and accelerate the execution of the proposed algorithm, the aforementioned processing steps take place within the smallest bounding box that encloses all the segmented pixels region of the current frame *S* is selected (activity bounding box).

The shadow of the moving human is often perceived by the segmentation algorithm as significant intensity pixel change and therefore it is also segmented as foreground moving object. The participation of the shadow to the calculation of the center of mass of the human and to any subsequent human shape estimation will lead to erroneous results. As it was observed in our experiments, the shadow cannot be removed simply by means of thresholding. Therefore, we employ an algorithm based on the color information of the current frame and the background model, as described in [9]. A pixel **p** current frame FR belongs to a shadow if it satisfies the following conditions:

$$a \le \frac{FR_I(\mathbf{p})}{BG_I(\mathbf{p})} \le b \; AND \; D_H(\mathbf{p}) \le \tau_s \; AND \qquad (5)$$
$$\left|FR_S(\mathbf{p}) - BG_S(\mathbf{p})\right| \le \tau_H$$

where BG indicates the background model and the subscripts *H*, *S* and *I* denote the hue, saturation and intensity of the HSI chromatic model and D_H(p) is given by:

$$D_H(\mathbf{p}) = \min\begin{pmatrix} \left|FR_H(\mathbf{p}) - BG_H(\mathbf{p})\right|, \\ 2\pi - \left|FR_H(\mathbf{p}) - BG_H(\mathbf{p})\right| \end{pmatrix}$$

The thresholds α, β, τ_h, τ_s are parameters of the algorithm that control the behavior of the shadow detection algorithm. Their appropriate values can be determined experimentally as in [15], which describes similar approach in the context of intelligent transportation. In this work we chose a set of values that work best with the specific conditions in the area of video acquisition (see Section 4 – par. 4.1).

Algorithm: Segmentation of Human Silhouette
Input: Current frame FR, background model BG
Output: Segmented frame S

```
for each frame FR
    for each pixel p in FR
        if abs(FR_I(p)-BG_I(p))>T1
            S(p)=1
           Calculate D_H(p) in (5)
           if (5) holds then S(p)=0
       else
            S(p)=0;
       end if
    end for
end for
```

The shadow detection is performed on the non-zero pixels of the thresholded background subtracted frame S and the detected pixels are suppressed. The algorithm for the segmentation of the human in video sequences is given below in pseudocode, using the same symbols as in equation (5).

3.2.3 Human Motion estimation

After the human segmentation step is complete, the segmented current frame contains the detected silhouette, which may not constitute a coherent group of pixels. In order to reject or join isolated pixels, the following steps are taken:

1. For each non-zero pixel in the segmented frame S, we retain its value if all its neighbors in a 3x3 vicinity are also set to 1, otherwise the pixel is set to 0.
2. The center of mass (x_m, y_m) of the remaining non-zero pixels is calculated.
3. Each of the remaining non-zero pixels $\mathbf{p}=(x_p,y_p)$ of S retains its value if it satisfies the condition:

   $\left|x_p - x_m\right| + \left|y_p - y_m\right| < T_3$, with T_3=150 pixels, otherwise it is set to 0. The position of the center of mass is then recalculated.

The apparent velocity of the center of mass may be calculated based on its position in the current and previous frame (*t* and *t*-1 respectively) using the following equation.

$$\mathbf{v}_t = \left(x_m^t, y_m^t\right) - \left(x_m^{t-1}, y_m^{t-1}\right) \qquad (6).$$

4. EXPERIMENTAL RESULTS

4.1 Video sources and execution time

The experiments were conducted on a series of video data, acquired using a fixed monocular camera with a wide-angle lens, placed on the roof of a room. The lighting in the room was artificial. The video Frames are coloured, acquired at a 5 frames second, with dimensions of 1024x768 pixels. A typical frame after grayscale conversion is shown in Figure 2. In each frame a human enters the room walks with variable speed, including temporarily standing and also performs a fall and either remains still or continues to move while on the floor.

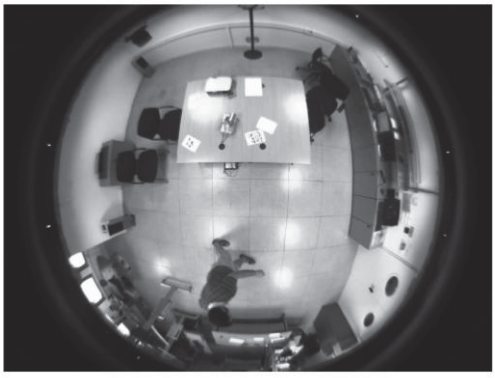


Figure 2. A typical frame from indoors-video sequences, acquired by the single, fish-eye camera positioned on the roof

The proposed video processing algorithms have been implemented using the Matlab programming environment and are executed on an INTEL i5 @ 2.4GHz dual core duo PC with 4 GB of memory. The execution time required for each component of the proposed system is given in Table I in seconds per frame, averaged over all the available frames. These time measurements were made without any optimization of the source code, in the Matlab programming environment. Even with this suboptimal implementation, the proposed algorithm is capable of handling low frame rate video with high-resolution RGB frames. Further source optimisation, exploitation of the Matlab's parallel processing capabilities or rewriting parts of the source code using a faster programming language are expected to substantially accelerate the execution of the proposed algorithmic workflow. The following subsection discusses the accuracy of the proposed methodology in human silhouette and movement detection.

Table I. The average execution time for the major steps of the proposed algorithm

Algorithm	Average time (sec/ frame)
Background subtraction	0.11
Shadow suppression	0.11
Centre of mass	0.22
Total time	0.44

4.2 Results on real video sequences

In this section we provide initial results from the application of the proposed algorithmic workflow on video sources. The presented results were obtained using the following values for the parameters: $T1$=0.15 (assuming maximum pixel value equal to 1), T_2=100 pixels, a=0.7, b=0.9, τ_s=0.5, τ_H=0.8 and T_3=150 pixels, which are found heuristically. Figure 3 shows the resulting segmentation of the human in an two indicative frames. The left column shows the segmentation without shadow suppression, the middle column show the detection of human shadow using the algorithm in par. 3.2.2, whereas the right column shows the final segmentation of the human figure. It can be observed that the shadow detection is quite accurate.

Figure 3. Human silhouette segmentation results, using the proposed algorithm for two indicative frames of the captured video sequences.

The final segmentation of the human from three different frames from two of the available video sequences are superimposed on a single binary frame in Fig. 4(a) and 4(b). In two of the frames of each video the human is in motion (walking) and in the 3rd frame the human is lying on the floor, while not staying still. It can be observed that the human is accurately segmented, with minimal missed pixels or background pixels erroneously segmented as human pixels.

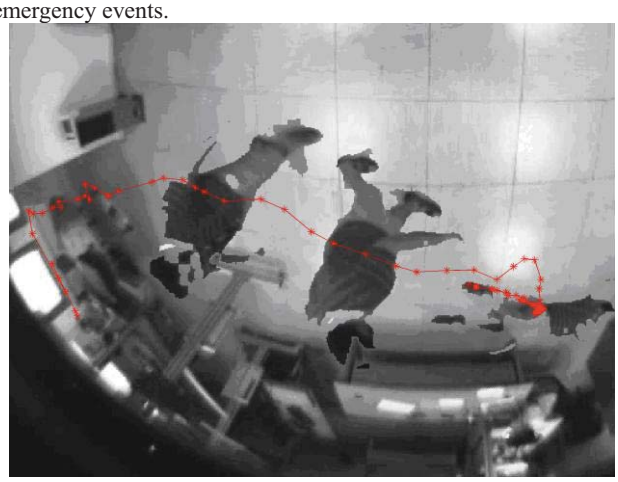
(a)

(b)

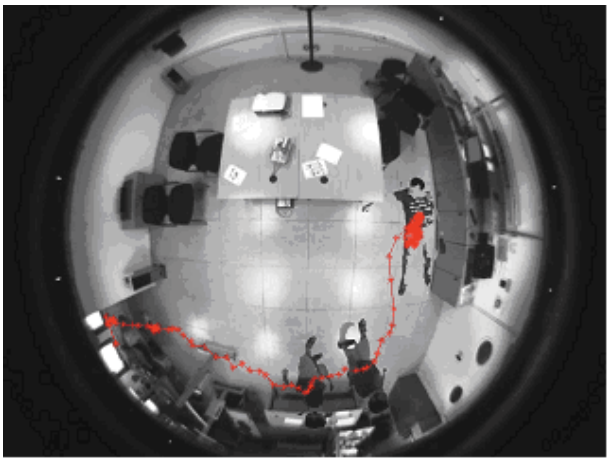
Figure 4. The segmentation of human in three different frames, from 2 indoor video sequences, using the proposed algorithm.

The same video sequences are presented in Figures 5a and 5b. In this set of figures, we show also superimposed the calculation of the position of the centre of mass of the segmented human figure for each frame. The detected human silhouettes from the three-segmented frames are also superimposed for better visualization. It can be observed that the calculation of the centre of mass of the segmented human is also quite accurate. The trajectory of the human path is an important feature that could be exploited for activity recognition. In addition the calculated velocity can assist

the detection of a fall and the evaluation of the severity of such emergency events.



(a)



(b)

Figure 5. The positions of the center of mass of the human, calculated for all available frames, overlaid on the initial frame of two video sequences

5. CONCLUSIONS

A novel algorithmic workflow is presented in this work that achieves segmentation of a moving human in RGB video sequences acquired indoors by a fish-eye, fixed camera at low frame rate. The developed algorithm determines also the trajectory of the centre of human mass. Preliminary results are presented that demonstrate the robustness and the computational efficiency of the proposed algorithms.

The proposed system may be used to model the shape of the human figure and further analyze its path, so that human activity analysis may be performed. Future work includes further enhancement of the proposed system with algorithms that can track multiple persons in indoor environments, as well as handle changes in illumination. Furthermore the current implementation will be optimized and/or rewritten to include parallel execution, so that video sequences with higher frame rates may be processed.

REFERENCES

[1] Albanese, M., Chellappa, R., Cuntoor, N. , Moscato, V., Picariello, A., Subrahmanian, V.S., Udrea, O. 2010, PADS: A Probabilistic Activity Detection Framework for Video Data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 12 (Dec. 2010), 2246-2261.

[2] Murayama, H. and Yamada, K. 2010, Detection of unusual human activity based on sequence of actions with MHI and CDP, In *Proceedings TENCON 2010 - 2010 IEEE Region 10 Conference*, (Nov. 2010), 1663-1667, 21-24.

[3] Takai, M. 2010, Detection of suspicious activity and estimate of risk from human behavior shot by surveillance camera, In *Proceedings of the Second World Congress on Nature and Biologically Inspired Computing (NaBIC)*, 2010, pp.298-304, (15-17 Dec. 2010).

[4] Zhongna Zhou, Xi Chen; Yu-Chia Chung, Zhihai He, Han, T.X.; Keller, J.M. 2008, Activity Analysis, Summarization, and Visualization for Indoor Human Activity Monitoring, Circuits and Systems for Video Technology, *IEEE Circuits and Systems for Video Technology*, 18, 11, 1489 – 1498.

[5] Nait-Charif, H. McKenna, S.J. 2004, Activity summarisation and fall detection in a supportive home environment, In proceedings of the *17th International Conference on Pattern Recognition (ICPR)* 2004, pp. 323-236, Aug. 2004.

[6] Miaou, S. –G., Sung, P.-H., Huang C. –Y. 2006, "A Customized Human Fall Detection System Using Omni-Camera Images and Personal Information", In Proceedings of the *1st Transdisciplinary conference on Distributed Diagnosis and Home Healthcare*, pp. 39-42, 2006.

[7] Doukas, C.N. and Maglogiannis, I. 2011, Emergency Fall Incidents Detection in Assisted Living Environments Utilizing Motion, Sound, and Visual Perceptual Components, *IEEE Transactions on Information Technology in Biomedicine*, 15, 2, (March 2011), 277-289.

[8] Willems, J., Debard, G., Bonroy, B., Vanrumste, B., and Goedemé, T. 2009. How to detect human fall in video? An overview. In *Proceedings of the positioning and contex-awareness international conference* (Antwerp, Belgium, 28 May, 2009), POCA '09.

[9] Cucchiara, R., Grana, C., Piccardi, M., and Prati A. 2003. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 10, (2003), 1337-1442.

[10] Wren, C., Azarhayejani, A., Darrell, T., and Pentland, A. P. 1997. Pfinder: real-time tracking of the human body, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7, (October. 1997), 780-785.

[11] Stauffer, C. and Grimson, W. E. L. 1999. Adaptive background mixture models for real-time tracking. In *Proceedings of the conference on computer vision and pattern recognition* (Ft. Collins, USA, June 23-25, 1999), CVPR '99. IEEE Computer Society, New York, NY, 246-252.

[12] Bouwmans T., Baf F. El and Vachon B., Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey, *Recent Patents on Computer Science* 1, 3 (2008) 219-237.

[13] McFarlane, N. J. B. and Schofield, C. P. 1995. Segmentation and tracking of piglets in images. *MACH VISION APPL.* 8, 3, (May. 1995), 187-193.

[14] Gonzalez, R.C., Woods, R.E., and Eddins, S. L. 2004. *Digital Image Processing using MATLAB 1st edition.* Prentice Hall.

[15] Cucchiara, R., Grana, C., Piccardi, M., Prati A., and Sirotti, S. 2001. Improving shadow suppression in moving object detection with HSV color. In *Proceedings of the conferenece in intelligent transportation system,* (August. 2001), 334-339.