Human Activity Recognition in Smart Environments

Monica-Andreea Dragan* and Irina Mocanu†
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
**Email:monica.dragan@cti.pub.ro*
†*Email: irina.mocanu@cs.pub.ro*

***Abstract*—This paper presents a method for image based human activity recognition, in a smart environment. We use background subtraction and skeletisation as image processing techniques, combined with Artificial Neural Networks for human posture classification and Hidden Markov Models for activity interpretation. By this approach we successfully recognized basic human actions such as walking, rotating, sitting and bending up/down, lying and falling. The method can be applied in smart houses, for elderly people who live alone.**

***Keywords*-Activity recognition, Background subtraction, Hidden Markov Model, Artificial Neural Network, Smart environment;**

I. Introduction

Human activity recognition has received increasing attention in the recent years due to its applications, especially in security industry and medical field.

In order to obtain accurate results, as many information as possible must be retrieved from the environment, enabling the system to locate and track the supervised person in each moment, to detect the position of the limbs and the objects the person interacts or has the intention to interact with. Sometimes, details like gaze direction or hand gestures can provide important information in the process of analyzing the human activity. Thus, the supervised person must be located in a smart environment, equipped with devices such as sensors, multiple view cameras or speakers.

This paper presents an image-based activity recognition approach, on day-time, by retrieving information from one single camera. The scenario of providing daily assisted living for elderly people who live alone in their house was taken into consideration. In this respect, we assume the existence of only one person in the room at one time and the fact that the person is the only moving element in the image.

The rest of the paper is organized as follows: Section II presents some existing methods for human activity recognition. Section III describes the proposed approach, while details about the experimental results and implementation of the model can be found in Section IV. Conclusion and future work are presented in sections V.

II. Related Work

Previous research in human activity recognition revealed multiple different approaches, depending on the information retrieved from the smart environment. The sensor-based and image-based approaches are mostly used. [1].

For a sensor-based approach, either wearable sensors (that can easyly detect actions such as walking and can also provide medical information, such as the frequency of heartbeats) or object-attached sensors (motion or proximity sensors, or sensors used to detect human interactions with certain objects in the room) can be used. A recent approach for wrist motion detection uses depth information from Kinect [2].

Image-based approaches use single [3], [4] or multiple cameras to reconstruct the 3D human pose, to detect the coordinates of the joints and to extract the limbs of the body. The image analysis is possible by isolating the human body from the background. This is achieved using a background subtraction algorithm that adapts to the environmental changes. Several techniques for adaptive background subtraction can be found in [5].

The information retrieved from the smart environment can be analyzed further using machine learning techniques, in order to build activity models and perform pattern recognition. The most used model is the Hidden Markov Model (HMM) - an graphical oriented method to characterize real world observations in terms of state models. Similar to simple Markov chains, a HMM has the property that the probability to be in a certain hidden state at a given time depends only on the previous state (it is independent from the whole transition history). In addition, HMM is based on a second assumption of independence: each observation is independent on the others, thus each observation depends only on the current state. Another good alternative is the Conditional Random Field (CRF) model, which is an undirected graphical method. In addition to HMM, CRF allows dependencies between observations and the use of incomplete information about the probability distribution of a certain observation. A balanced view on the two models can be found in [6].

Several implementations of activity recognition models are already on the market and it is worth being mentioned.

The application developed within the CASAS project at Washington State University [7] is sold as a smart home in a box that can be set up in ordinary houses. The goal of this application is to analyze the resident behavior for a large demographic area, for statistical purpose. For real time activity recognition, a support vector machine (SVM) method was designed. In order to discover activity patterns, a sequence mining technique based on greedy search among the input sets was used.

Other applications such as those presented in [7] and [9] provide assisted living for medical purposes and are already used in hospitals for remote monitoring of patients health.

The approach in [7] involves simple state-change sensors placed on objects in the environment. For activity recognition, the system implements a model-based approach using nave Bayes classifiers. In the training phase of the mathematical model, the user labels its own activities. Eventually, activities such as eating, using the toilet, dressing, shopping or managing medication were detected.

III. Human Activity Recognition Model

In order to design a computational model for human activity understanding, this paper proposes a system that takes into consideration only one person in the room, whose activities are supervised by a single camera. Also, the system does not consider the privacy implications of having a supervising camera in the persons home. The system can be used in medical purpose: to assist elders who live alone in their house.

In order to understand and interpret one persons movements, a model divide in two main levels was used:

1) Firstly, single image frames retrieved from the camera are processed to obtain information about the persons static posture: silhouette, skeleton and posture classification. Information about single frames is stored in a database for further processing.
2) Secondly, a sequence of consecutive images is analyzed, considering the time evolution of certain parameters and matching the sequence with certain precalculated mathematical models for certain activities (Hidden Markov Models are used). These two steps are itemized in this section.

These two steps are described in more details further in this section.

Analyzing single image frames: The image processing techniques used for single frame analysis are: background subtraction and application of several noise reduction and smoothing filters in order to obtain an accurate silhouette. An Artificial Neural Network was used to classify the silhouettes over 6 categories, corresponding to 6 human postures.Additional information concerning angles of the limbs, aspect ratio and image gradient were computed from the silhouette image, to be used in the second step the analysis of a sequence of images.

The computational level for single image frame analysis is described in Fig. 1.

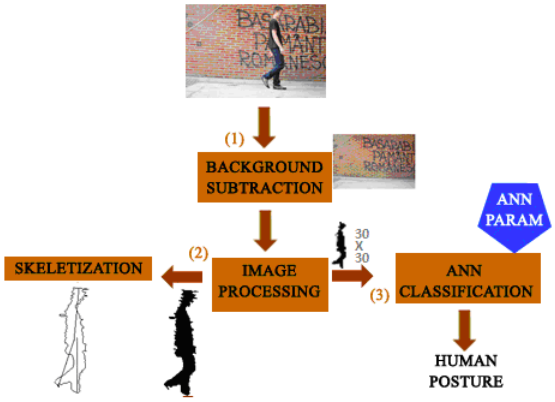

Figure 1. Single image analysis model

This level is composed of the following steps:

1) background subtraction for human detection and silhouette extraction;
2) image processing and skeletisation, which provide several image parameters: height and width of the silhouette, the coordinates of the centroid of the body, horizontal and vertical gradient of the image, the coordinates of the extremities of the body extracted by skeletization;
3) Artificial Neural Network (ANN) classification of the silhouette among 6 postures (standing up right profile/ front/ left profile, sitting down right side/ front/ left side).

Analyzing a sequence of images from the database: The second level for the multi-hierarchical classification model can be seen in Fig. 2.

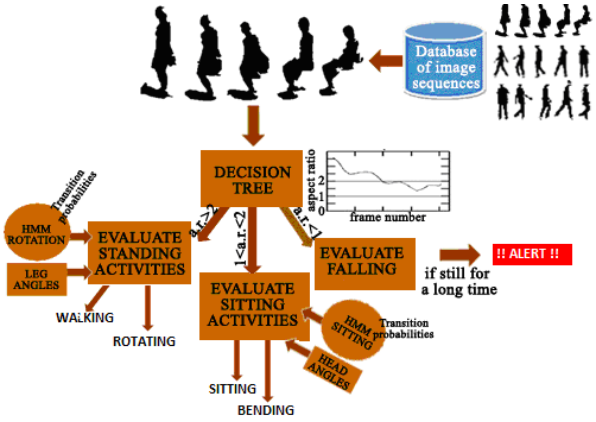

Figure 2. Analyzing model for a sequence of images

Single frame processing is briefly described below:

1) primary classification between activities using a decision tree, which takes into considerations some parameters of the image such as aspect ratio and image gradient;
2) simple activity recognition having as input a sequence of postures classified with ANN, using Hidden Markov Models;
3) high level activity recognition of a more complex movement (a sequence of linked simple activities);

A UML (Unified Modeling Language) sequence diagram showing the workflow for evaluating a sequence of images representing a rotation movement can be seen in Fig. 3. The main steps of the system can be identified in the diagram: after selecting a sequence of processed images from the database, the ANN classification of the posture is performed for each image; by analyzing the parameters of the images and matching the sequences with the corresponding HMM model, a decision is taken. If the sequence contains multiple basic activities, the real complex classifier is recalled and a real time simulation is displayed.

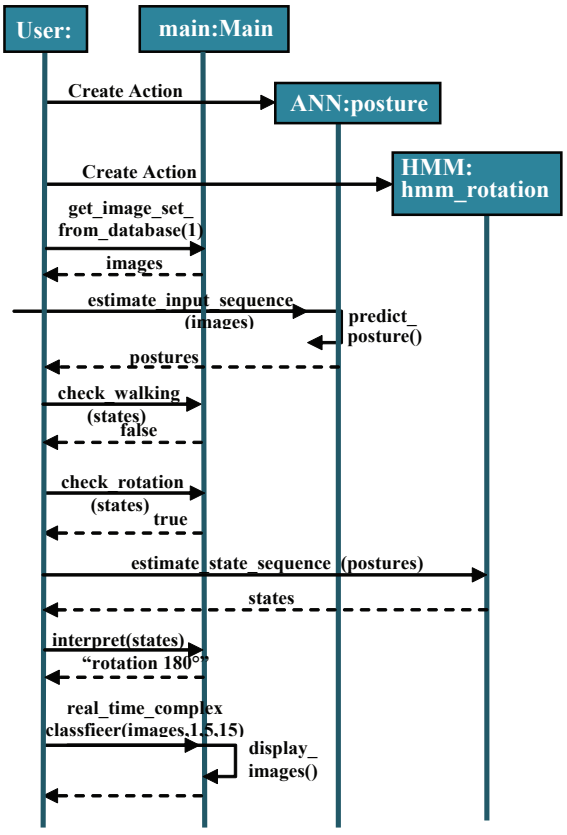

Figure 3. UML sequence diagram for rotation activity (using a seqence of images)

Further, each step will be discussed in more in details.

A. Analyzing a single image frame

1) Background extraction: A non-adaptive background subtraction method for human detection and silhouette extraction was used, in order to process single frames (RGB images). In the present study, this approach was sufficient to achieve the target of the study: activity recognition. An adaptive background subtraction [10] should be considered for future improvement. Before performing the background subtraction, a background image, with no person in the frame, must be provided. By subtracting pixel by pixel, for each color channel (RGB), the image containing the person and the background image, a proper threshold that separates the silhouette from the image background must be found. This threshold is different from a set of images to another, depending on the illumination conditions in the environment. However, this approach alone is not accurate enough because of the shadows that may appear around the moving person and the change of luminosity that may occur in the room during a longer period of time. To remove the shadows, the parameters of the image that remain unchanged regardless the luminosity and the shadows were considered. In this paper, the c1 c2 c3 color invariants model is usused, as suggested in [11]. The c1 c2 c3 invariant color features are defined for each channel in equations 1, 2 and 3.

$$c = arctan\frac{Red}{max(Green, Blue)} \quad (1)$$

$$c2 = arctan\frac{Green}{max(Red, Blue)} \quad (2)$$

$$c3 = arctan\frac{Blue}{max(Red, Green)} \quad (3)$$

By combining the two methods (RGB pixel subtraction Fig. 4a and c1 c2 c3 model Fig. 4b) and by choosing the proper thresholds, in most of the cases an accurate silhouette was extracted (Fig. 4c).

2) Skeletization and silhouette analysis
   a) Skeletization: in order to find the coordinates of the extremities of the body (head, hands, legs) the skeletization algorithm described in [12] was applied. The skeletization process has 4 steps:
      - finding the coordinates of the pixels on the edge of the silhouette (ordered clockwise or counterclockwise) and the centroid of the body;
      - finding the distances between all the points on the edge and the centroid;
      - smoothing the distance vector using a linear smoothing filter;
      - the local maxima of the distance vector are the extremities of the body (Fig. 5); The precision

Figure 4. a) RGB pixel subtraction; b) c1c2c3 shadow subtraction; c) combining the methods a) and b)

of the smoothing filter must be adjusted, in order to obtain exactly the 5 extremes (head, hands, legs).

b) Other parameters: apart from the coordinates of the extremities, the following parameters were extracted from the silhouette and recorded into the database:
   - the height and width of the silhouette;
   - the coordinates of the centroid of the body;
   - the horizontal and vertical gradient distributions of the picture that contain information about the dynamics of the image on the two axis;

c) Artificial Neural Network (ANN) silhouette classification: after performing the background subtraction, a resized 30x30 image of the human silhouette was obtained. The 30x30 image was used further as input for a pre-trained Artificial Neural Network, in order to obtain the classification. The following six human postures were considered:
   - upright profile
   - up front
   - up left profile
   - sitting on the right side
   - sitting on the front side
   - sitting on the left side
   - upright profile:

B. Analyzing a sequence of images from the database

1) Primary classification between activities: by analyzing the parameters extracted from the silhouette (the parameters described in Section III.A) and taking into consideration the method described in [13] (based on

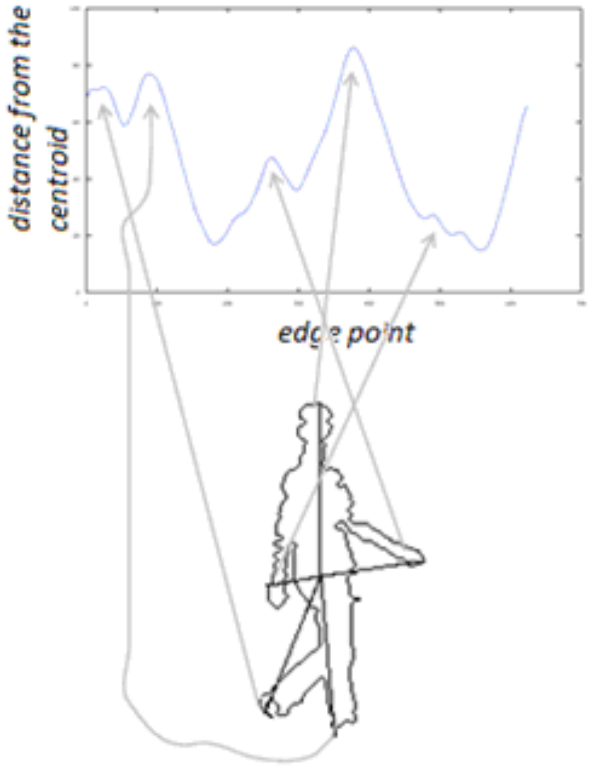

Figure 5. Example of skeletization

the aspect ratio and gradient distribution of the image), the activities can be a classified as follows:
   - *falling* (lying on the bed or falling on the floor): the aspect ratio is smaller than 1 and the vertical gradient becomes higher than the horizontal one (this is explained by the fact that, when the person is falling, the dynamics of the picture on the vertical axis becomes higher than the horizontal one);
   - *sitting or bending*: the aspect ratio remains between 1 and 2;
   - *standing up* (*standing, rotating or walking*): the aspect ratio is higher than 2 and the horizontal gradient is higher than the vertical one (the movement is mainly on the OX axes);
   - *standing* still in all the postures above. The aspect ratio and gradient parameters remain almost constant;

   In Fig. 6, the aspect ratio and gradient parameters for certain movements were plotted.

2) Basic activity recognition using Hidden Markov Models (HMM) For the evaluation of a set of images, a HMM was used - the observable states are the postures resulted from the ANN, as shown in Fig. 7.
   The parameters of the HMM (see Fig. 7) are retrieved

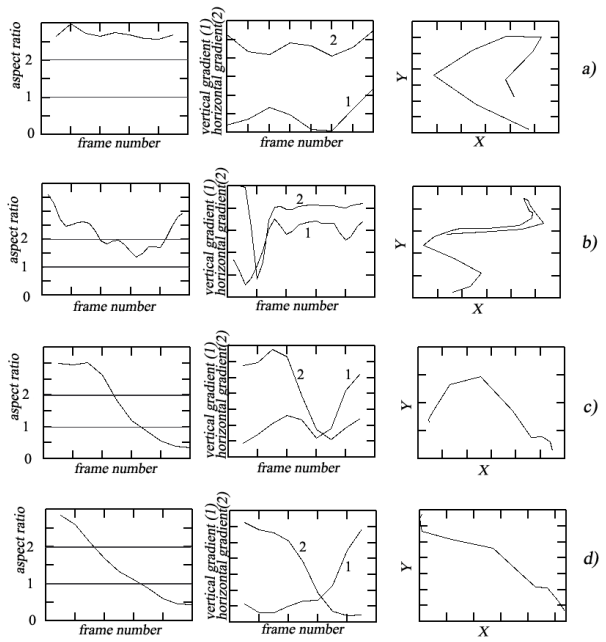

Figure 6. Parameters evolution (aspect ratio, gradient and 2D trajectory) over a sequence of images representing walking (a), sitting (b) and lying activities (c)

from the statistics made on the training sequences and from those made on the correctness of the ANN outputs (for the observation probabilities). Fig. 8 shows a HMM example: 180^0 rotation for states 1 to 5 and sitting down for states 5 to 9.

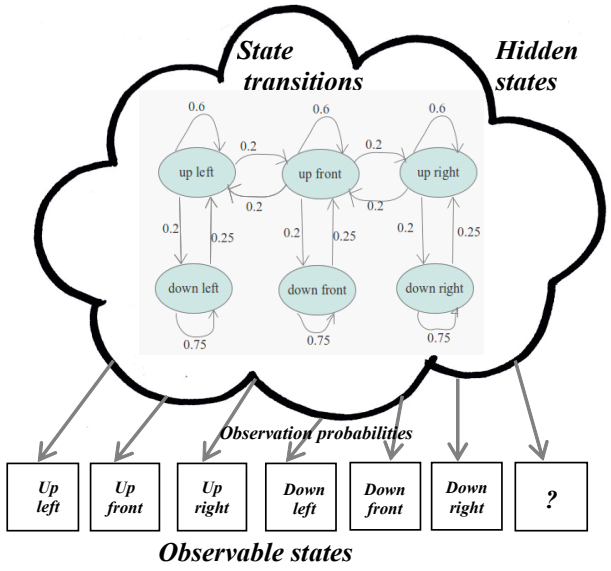

Figure 7. Transitions between hidden states and observable states for a sitting model HMM

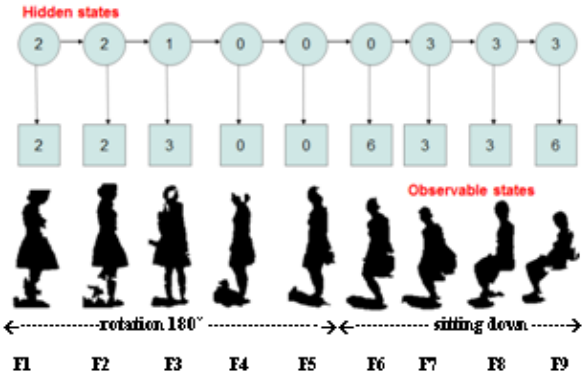

Figure 8. Transitions between hidden states and observable states for a sitting model HMM

In the observation sequence, the output of the ANN (observable states) is wrong for frames 3, 6, 9. The HMM network was used to solve the path decoding problem: to find the best state sequence for a given observation sequence. After applying the Viterbi algorithm on the HMM, a hidden state sequence consistent with the reality was obtained. The interpretation of the transitions is: rotation 180^0 for states 1 to 5 and sitting down for states 5 to 9.

- **Walking:** to decide if the person is walking, the angle between the legs was analyzed. For a walking activity, the angle may vary periodically among the value of 30, as shown in Fig. 9. This angle was computed by adding together the angle of the first extremity on the right side of the vertical axis, below the centroid, and the one on the left side. If the angle is higher than 70^0 with the vertical axis, it is considered that one of the legs is hidden and the angle is to 0^0.

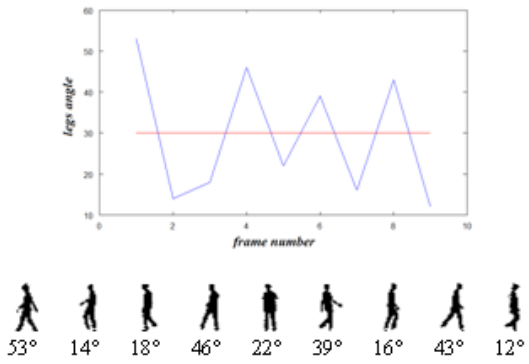

Figure 9. Legs angle variation during a walking movement

- **Bending:** in case of bending, the head/vertical axis angle was computed (the head is considered

the extremity above the centroid. Depending on the posture given by the ANN, it can be decided whether the head is on the right side of the vertical axis or on the left side. If the angles in the vector decrease, it means that the person is bending up. Otherwise the person is bending down (see Fig. 10).

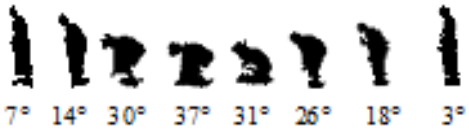

Figure 10. Head angles during a bending movement

- **Falling on the floor or lying on the bed:** for classifying a lying action (Fig. 11), the aspect ratio is analyzed. If the evolution of the aspect ratio of the sequence of images decreases from 2 to 0, it means that the person is falling down. No distinction between falling down on the floor and lying on the bed was made, but if the person faints, he or she will not move for a certain period of time, and the program will raise an alert. However, the case of abnormal reactions was not taken into consideration (such as an epileptic crisis, when the person falls down on the floor but still moves). In order to adjust the possible errors that appeared during the image analysis, the vector of numerical values was smoothed.

Figure 11. Falling down movement

3) High level activity recognition In order to analyze a longer set of images representing linked actions, the basic activity recognizer was used to evaluate sequences of a certain length, between a predefined window value. Also, to adjust the window, the evolution of the aspect ratio of the silhouette was taken into consideration. The output probability returned by the classifier was scaled to the number of images in the sequence. The activity with the highest probability was assigned to the sequence. A graphical representation of the parameter evolution for a more complex movement (that involves walking, sitting and lying) is shown in Fig. 12.

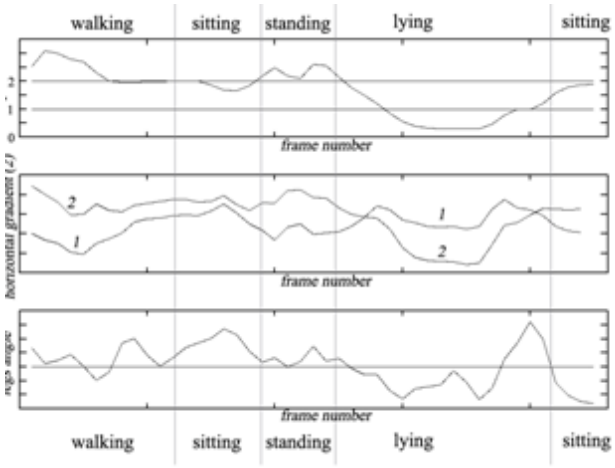

Figure 12. Aspect ratio, gradient and legs angle evolution for a sequence of imagest

IV. Experimental Results

In order to test the model, two standalone applications were used: one for single image frame processing (Octave framework was used) and other for the classification level (analyzing a sequences of frames from the database, see Section IIIB), implemented in Java. To simulate a continuous movement, the train and test image sets were captured with an ordinary camera, in continuous shooting mode. During the shooting, the parameters of the camera (exposure time and aperture) were kept unchanged, and the illumination conditions did not significantly chamge from the moment the background image was captured until the end of each shooting session. A total number of 870 images were taken, representing 50 moving scenarios. Information about each image frame was stored in a database (see Section IIIA). For each test, a sequence of consecutive images (one scenario) was selected from the database for activity classification. The application performs a non-real-time classification, by evaluating the probabilities of classification between the possible activities described in the previous section. The result of the classification was saved in a log file. However, a graphical interface simulating a real time evaluation was displayed. Supposing a continuous moving video with 20 frames per second, the simulations displays an average of 1.5 frames per second which are processed in less than 0.5 seconds each. However, the image processing is executed a priori in Octave. The source code and a demo of the application on a more complex movement with multiple activities can be found at [14].

A. Background subtraction

The accuracy of activity recognition depends on the quality of background subtraction and on the correctness of the ANN classification. Also, to obtain a good quality background subtraction, the thresholds and the precision

of the skeletization must be properly adjusted. The result of the background subtraction was good or satisfactory in all images taken in good conditions. However, from low quality images and images with multiple lightening focuses (with a lot of shadows) the background was not successfully extracted.

B. Artificial neural network

In order to perform the classification, an Artificial Neural Network trained on 200 images and tested on 200 images was use. The performance of the ANN had an efficiency of 72% in classifying the correct posture for new images (not used in the training process). After trying a large number of network configurations, the best results were obtained for a neural network with 900 input nodes, 6 output nodes and one hidden layer with 270 neurons. A result is considered to be wrong if no one of the outputs exceeds a probability higher than 50%, or if more than one output exceeds the 50% probability. The error in 28% of the cases may be explained by three factors:

- sometimes the posture of the person in the image is intermediate between two postures that the ANN was trained for (e.g. an intermediate posture between up-right profile and up front, i.e. a 45 degrees orientation);
- the low quality of the background extraction;
- the ANN cannot distinguish correctly between sitting on the front side and standing up front. This can be adjusted by taking into consideration the aspect ratio of the silhouette.

C. Hidden Markov Model

The parameters of the HMM were learnt from the statistics on the input sets. For example, to compute the transition probability from the hidden state up right to the observable state up front, ,it was counted the number of times when the output of the ANN classification was up front for an up right image input (this is in fact the probability of having a wrong classifications). The probabilities obtained using 200 test images are presented in Table I. A graphical representation of the HMM sitting model is provided in Fig. 7. For the implementation of the Artificial Neural Network and for the Hidden Markov Model, the Encog Java library [15] was used.

To quantify the performance of the HMM, 50 sequences of medium and high quality images describing simple activities were evaluated. 86% of them were correctly recognized. A detailed statistics on the results grouped by activity can are presented in Table II.

Evaluation errors may be expected any time, as long as the accuracy of the background subtraction is not good (the threshold was not properly chosen). Also, errors may appear if the supervised person is disproportional and the aspect ratio thresholds considered above do not correspond.

Table I
OBSERVATION PROBABILITIES: TRANSITION PROBABILITIES BETWEEN HIDDEN STATES (LINES) AND OBSERVABLE STATES (COLUMNS)

Hidden states	**Up left (0)**	**Up front (1)**	**Up right (2)**	**Deown left (3)**	**Down front (4)**	**Down right (5)**	**? (6)**
Up left	0.85	0.02	0.02	0.04	0	0.02	0.04
Up front	0.25	0.45	0.05	0.02	0	0.02	0.17
Up front	0.01	0.01	0.74	0.07	0	0	0.13
Down left	0	0	0.14	0	0	0.85	0
Down front	0.25	0.05	0.05	0.05	0.3	0.05	0.05
Down right	0	0	0.14	0	0	0.85	0

Table II
RESULTS OF THE CLASSIFICATION GROUPED BY ACTIVITY

Activity	**Number of scenarios**	**Number of successful classification**	**0%**
Standing/ Walking	10	8	80%
Sitting	10	8	80%
Rotating	10	10	100%
Bending	10	7	70%
Lying	10	10	100%

V. CONCLUSIONS AND FUTURE WORK

This paper presents a model for human activity recognition using a supervised learning model, applied to data retrieved by a single camera. Basic activities such as rotating, walking, sitting up/down, falling up/down are detected by the application in 86% of the tested scenarios. However, abnormal movements or gestures not taken into consideration in the training process are not identified.

In the future, this model will be integrated in a multi-agent system. Each agent will obtain its own information and the final activity will be computed using a negotiation method as described in [16]. Also we integrate in our solution another image classification method based on genetic algorithms as described in [17].

Important improvements of the current approach (using a single camera) will be achieved by adopting an adaptive background subtraction method. However, for a real life application, it is necessary to provide a real time computational system. This could be achieved by eliminating the most time consuming part of the computation - the image processing. A solution in this case would be to retrieve preprocessed data from the Kinect.

For a more accurate approach, the system will be improved in the future by automatically estimating the parameters of the HMM and consequently discovering new patterns in the human behavior and transforming the actual system in a self-learning system. This can be combined with a

more large range of features extracted from one image frame (information about interactions with objects, 3D positioning, and sound recording). However, this approach implies a more sophisticated and expensive smart environment.

ACKNOWLEDGMENT

The work has been co-founded by the FP7 project ERRIC: Empowering Romanian Research on Intelligent Information Technologies, No. 264207 and the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.
We would like to thank to all the people who were involved in the organization of the CASIA project - POSDRU/109/2.1/G/81772 and to all people who provided good conditions for working on this project.
Also, we would like to acknowledge all those who accepted to be taken photos for different activities input tests.

REFERENCES

[1] L. Chen, C. Nugent, J. Biswas, J. Hoey, "Activity Recognition in Pervasive Intelligent Environments", Chapter 1 , Activity Recognition: Approaches, Practices and Trends, ATLANTIS PRESS, 2011.

[2] V. Elangovan, V. Bandaru, A. Shirkhodaie, "Team Activity Analisys and Recognition Based on Kinect Depth Map and optical Imagery Techniques", Proceedings SPIE 8392, Signal Processing, Sensor Fusion, and Target Recognition XXI, 83920W (May 1, 2012), 2012, doi:10.1117/12.919946.

[3] H. Jiang, "3D Human Pose Reconstruction Using Millions of Exemplars", International Conference of Pattern Recognition 2010, IEEE 1051:4651/10.

[4] J.L. Boeheim, "Human Activity Recognition Using Limb Component Extraction", Rochester Institute of Technology, 2008.

[5] M. Piccardi, "Background subtraction techniques: a review," 2004 IEEE International Conference on Systems, Man and Cybernetics, vol.4, 2004, pp. 3099- 3104.

[6] E. Kim, S. Helal, D. Cook, "Human Activity Recognition and Pattern Discovery," IEEE Pervasive Computing", 2010, pp. 48-53.

[7] D. Cook, A. Crandall, B. Thomas, N. Krishnan, "CASAS: A Smart Home in a Box", Washington State University, U.S.A, 10.1109/MC.2012.328.

[8] E. Tapia, T. S. Intille, K. Larson, "Activity Recognition in the Home Using Simple and Ubiquitous Sensors," Pervasive Computing, vol. 3001, 2004, pp. 158-175.

[9] Ecaalyx project:
http://ecaalyx.org/

[10] B. Langmann, S.E. Ghobadi, K. Hartmann, O. Loffeld , "Multi-modal background subtraction using Gaussian mixture models," Paparoditis N., Pierrot-Deseilligny M., Mallet C., Tournaire O. (Eds), IAPRS 2010, Vol. XXXVIII, Part 3A

[11] E. Salvador,a A. Cavallaro, T. Ebrahimi, "Cast shadow segmentation using invariant color features," Computer Vision and Image Understanding, Volume 95 Issue 2, August 2004, pp 238 - 259.

[12] H. Fujiyoshi, A. Lipton, "Real-time human motion analysis by image skeletonization", Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision, WACV '98, 1998, pp. 15-21.

[13] V. Vishwakarma, C. Mandal and S. Sural , "Automatic Detection of Human Fall in Video", Proceeding PReMI'07 Proceedings of the 2nd international conference on Pattern recognition and machine intelligence, pp. 616-623.

[14] Source code:
http://www.interq.ro/mediawiki/index.php/Ai#Surse

[15] Encog:
http://www.heatonresearch.com/encog

[16] S. Radu, E. Kalisz, and A. M. Florea, "A model of automated negotiation based on agents profiles", Scalable Computing: Practice and Experience Journal, in press.

[17] I. Mocanu, E. Kalisz and L. Negreanu, "Genetic Algorithms Viewed as Anticipatory Systems," CASYS 2009, Liege, Belgium , AIP Conference Proceedings, Vol. 1303, DOI: 10.1063/1.3527157, 2010, pp. 207-215.