Monitoring Activities of Daily Living (ADLs) of Elderly Based on 3D Key Human Postures

Nadia Zouba, Bernard Boulay, François Brémond, Monique Thonnat

▶ To cite this version:

Monitoring Activities of Daily Living (ADLs) of Elderly Based on 3D Key Human Postures

Nadia ZOUBA, Bernard BOULAY, Francois BREMOND and Monique THONNAT

INRIA Sophia Antipolis, PULSAR Team, 2004, route des Lucioles, BP93, 06902 Sophia Antipolis Cedex, France

Abstract. This paper presents a cognitive vision approach to recognize a set of interesting activities of daily living (ADLs) for elderly at home. The proposed approach is composed of a video analysis component and an activity recognition component.

A video analysis component contains person detection, person tracking and human posture recognition. A human posture recognition is composed of a set of postures models and a dedicated human posture recognition algorithm.

Activity recognition component contains a set of video event models and a dedicated video event recognition algorithm.

In this study, we collaborate with medical experts (gerontologists from Nice hospital) to define and model a set of scenarios related to the interesting activities of elderly. Some of these activities require to detect a fine description of human body such as postures. For this purpose, we propose ten 3D key human postures usefull to recognize a set of interesting human activities regardless of the environment. Using these 3D key human postures, we have modeled thirty four video events, simple ones such as "a person is standing" and composite ones such as "a person is feeling faint". We have also adapted a video event recognition algorithm to detect in real time some activities of interest by adding posture.

The novelty of our approach is the proposed 3D key postures and the set of activity models of elderly person living alone in her/his own home.

To validate our proposed models, we have performed a set of experiments in the Gerhome laboratory which is a realistic site reproducing the environment of a typical apartment. For these experiments, we have acquired and processed ten video sequences with one actor. The duration of each video sequence is about ten minutes and each video contains about 4800 frames.

1 Introduction

The elderly population is expected to grow dramatically over the next 20 years. The number of people requiring care will grow accordingly, while the number of people able to provide this care will decrease. Without receiving sufficient care, elderly are at risk of loosing their independence. Thus a system permitting elderly to live safely at home is more than needed. Medical professionals believe that one of the best ways to detect emerging physical and mental health problems, before it becomes critical - particularly for the elderly - is analyzing the human behavior and looking for changes in the activities of daily living (ADLs). Typical ADLs include sleeping, meal preparation, eating, housekeeping, bathing or showering, dressing, using the toilet, doing laundry, and managing medications. As a solution to this issue, we propose an approach which consists in the modeling of ten 3D key human postures useful to recognize some interesting activities of elderly. The recognition of these postures is based on using the human posture recognition algorithm proposed in [1] which can recognize in real time human postures with only one static camera regardless of its position.
In this paper, we focus on recognizing activities that elderly are able to do (e.g. ability of elderly person to reach and open a kitchen cupboard). The recognition of these interesting activities helps medical experts (gerontologists) to evaluate the degree of frailty of elderly by detecting changes in their behavior patterns. We also focus on detecting critical situations of elderly (e.g. feeling faint, falling down), which can indicate the presence of health disorders (physical and/or mental). The detection of these critical situations can enable early assistance of elderly.

In this paper, section 2 briefly reviews previous work on human posture recognition and activity recognition using video cameras. Section 3 describes our activity recognition approach. Results of the approach are reported in section 4. Finally, conclusion and future works are presented in section 5.

2 State of the Art

In this section we present firstly previous work on human posture recognition using 2D and 3D approaches, and secondly previous work on activity recognition.

2.1 Human Posture Recognition by Video Cameras

The vision techniques to determine human posture can be classified according to the type of model used (explicit, statistical, ...) and the dimensionality of the work space (2D or 3D). The **2D approaches** with **explicit models** [2] try to detect some body parts. They are sensitive to segmentation errors. The **2D approaches** with **statistical models** [3] are then proposed to handle the problems due to segmentation. These two 2D approaches are well adapted for

real time processing but they depend on the camera view point. The **3D approaches** can also be classified in **statistical** and **model** based techniques. They consist in computing the parameters of the 3D model, such as the model projection on the image plane fits with the input image (often the silhouette). Some approaches compare the contour of the input silhouette with one of the projected model. In [4], the authors propose a method to reconstruct human posture from un-calibrated monocular image sequences. The human body articulations are extracted and annotated manually on the first image of a video sequence, then image processing techniques (such as linear prediction or least square matching) are used to extract articulations from the other frames. The learning-based approaches avoid the need of an explicit 3D human body model. In [5], the authors propose a learning-based method for recovering 3D human body posture from single images and monocular image sequences. These 3D approaches are partially independent from the camera view point but they need to define many parameters to model the human posture.

In this study, we choose to use an **hybrid approach** described in [1]. This approach combines the advantages of the 2D and 3D approaches to recognize the entire human body postures in real-time. It is based on a 3D human model and is independent from the point of view of the camera and employs silhouette represented from 2D approaches to provide a real-time processing.

2.2 Activity Recognition

Previous activity detection research focused on analyzing individual human behaviors. **Rule-based methods** proposed in [6] have shown their merits in action analysis. Rule-based systems may have difficulties in defining precise rules for every behavior because some behaviors may consist of fuzzy concepts. **Statistical approaches**, from template models, linear models, to graphic models, have been used in human activity analysis. Yacoob and Black [7] used linear models to track cyclic human motion. Jebara and Pentland [8] employed conditional Expectation Maximization to model and predict actions. Aggarwal et. al. [9] has reviewed different methods for human motion tracking and recognition. The probabilistic and stochastic approaches include HMM (Hidden Markov Model) and NNs (Neuronal Networks). They are represented by graphs. Hidden Markov models [10] have been used for recognizing actions and activities, and illustrated their advantages in modeling temporal relationships between visual-audio events. Chomat and Crowley [11] proposed a probabilistic method for recognizing activities from local spatio-temporal appearance. Intille and Bobick [12] interpret actions using Bayesian networks among multiple agents. Bayesian networks can combine uncertain temporal information and compute the likelihood for the trajectory of a set of objects to be a multi-agent action. Recently, Jesse Hoey et al. [13] successfully used only cameras to assist person with dementia during handwashing. The system uses only video inputs, and combines a Bayesian sequential estimation framework for tracking hands and towel, with a decision using a partially observable Markov decision process. Most of these methods

mainly focus on a specific human activity and their description are not declarative and it is often difficult to understand how they work (especially for NNs). In consequence, it is relatively difficult to modify them or to add a priori knowledge. The **deterministic approaches** use a priori knowledge to model the events to recognize [14]. This knowledge usually corresponds to rules defined by experts from the application domain. These approaches are easy to understand but their expressiveness is limited, due to the fact that the variety of the real world is difficult to represent by logic. The **approaches based on constraint resolution** are able to recognize complex events involving multiple actors having complex temporal relationships.

In this work we have used the approach described in [15]. This approach is based on constraint resolution. It uses a declarative representation of events which are defined as a set of spatio-temporal and logical constraints. This technique is easy to understand since it is based on constraints which are defined in a declarative way.
The next section presents the approach for activity recognition we used in this paper.

3 The Proposed Activity Recognition Approach

3.1 Overview

The proposed approach is composed of (1) a video analysis component which contains person detection, person tracking and a human posture recognition, (2) an activity recognition component which contains a set of video event models and a video event recognition algorithm. A simplified scheme of the proposed approach is given in figure 1. Firstly, we present the video analysis component and secondly the activity recognition component.

3.2 Video Analysis

In this section we describe shortly person detection and tracking method. We also describe the used human posture recognition algorithm and we detail the proposed 3D posture models.

Person Detection and Tracking. For detecting and tracking person we use a set of vision algorithms coming from a video interpretation platform described in [16]. A first algorithm segments moving pixels in the video into a binary image by subtracting the current image with the reference image. The reference image is updated along the time to take into account changes in the scene (light, object displacement, shadows). The moving pixels are then grouped into connected regions, called blobs. A set of 3D features such as 3D position, width and height are computed for each blob. Then the blobs are classified into predefined
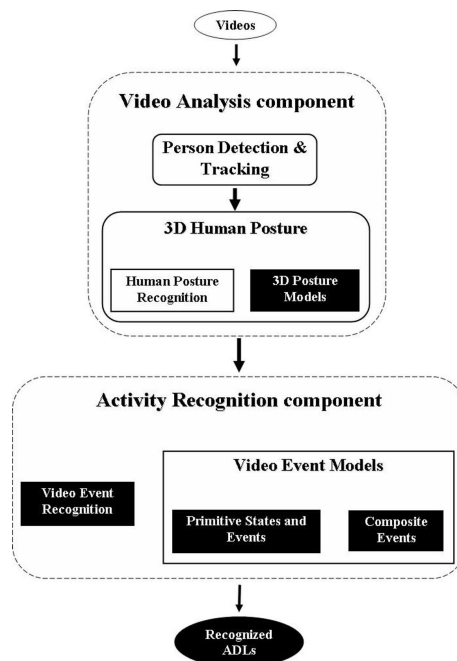
Fig. 1. The architecture for the proposed approach. The contribution of this paper is represented with black background

classes (e.g. person). After that the tracking task associates to each new classified blob a unique identifier and maintains it globally throughout the whole video. Figure 2 illustrates the detection, classification and tracking of a person in the experimental laboratory.

(a) Original image   **(b) Detection**   **(c) Classification**   **(d) Tracking**
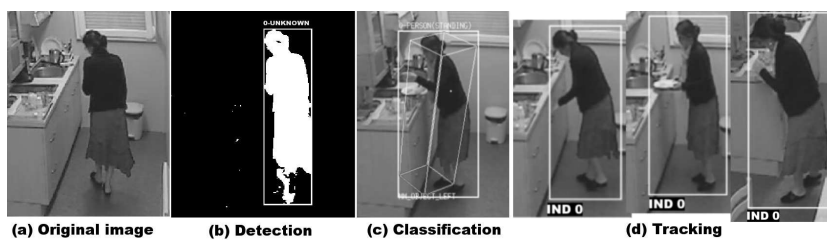
Fig. 2. Detection, classification and tracking of a person. (a) represents the original image. (b) the moving pixels are highlighted in white and clustered into a mobile object enclosed in an orange bounding box. (c) the mobile object is classified as a person. (d) shows the individual identifier (IND 0) and a colored box associated to the tracked person

3D Human Posture Recognition. In this section, we firstly present the human posture recognition algorithm and secondly the proposed 3D posture models.

– **Human Posture Recognition Algorithm:** We have used a human posture recognition algorithm [1] in order to recognize in real time a set of human postures once the person evolving in the scene is correctly detected. This algorithm determines the posture of the detected person using the detected silhouette and its 3D position. The human posture recognition algorithm is based on the combination between a set of 3D human model with a 2D approach. These 3D models are projected in a virtual scene observed by a virtual camera which has the same characteristics (position, orientation and field of view) than the real camera. The 3D human silhouettes are then extracted and compared with the detected silhouette using a 2D techniques (projection of the silhouette pixels on the horizontal and vertical axes, see figure 3).
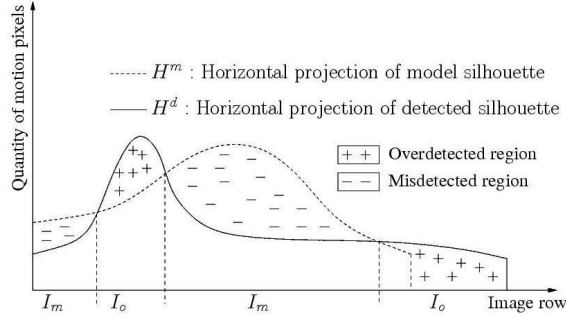
Fig. 3. Horizontal projection of model and detected silhouettes. I_o (resp. I_m) represents the overdetected (resp. misdetected) region

$$R_o(H) = \frac{\sum_{i \in I_o}(H_i^d - H_i^a)^2}{\sum_i (H_i^d)^2} \ , \ R_m(H) = \frac{\sum_{i \in I_m}(H_i^d - H_i^a)^2}{\sum_i (H_i^a)^2} \qquad (1)$$

$(\alpha_1, \alpha_2, \beta_1, \beta_2) \in [0,1]^4, \alpha_1 + \alpha_2 + \beta_1 + \beta_2 = 1$

The distance from the detected silhouette to the model silhouette is computed with the equation (2):

$$dist(S_a, S_d) = \alpha_1 R_o(H) + \beta_1 R_m(H) + \alpha_2 R_o(V) + \beta_2 R_m(V) \qquad (2)$$

The most similar extracted 3D silhouette corresponds to the current posture of the observed person. This algorithm is real time (about eight frames per

second), requires only a fix video camera and do not depend on the camera position.

– **3D Posture Models:** The **posture models** are based on a 3D geometrical human model. We propose ten 3D key human postures which are useful to recognize activities of interest. These postures are displayed in figure 4: standing (a), standing with arm up (b), standing with hands up (c), bending (d), sitting on a chair (e), sitting on the floor with outstretched legs (f), sitting on the floor with flexed legs (g), slumping (h), lying on the side with flexed legs (i), and lying on the back with outstretched legs (j). Each of these postures plays a significant role in the recognition of the targeted activities of daily living. For example, the posture "standing with arm up" is used to detect when a person reaches and opens kitchen cupboard and her/his ability to do it. The posture "standing with hands up" is used to detect when a person is carrying an object such as plates. These proposed human postures are not an exhaustive list but represent the key human postures taking part in everyday activities.
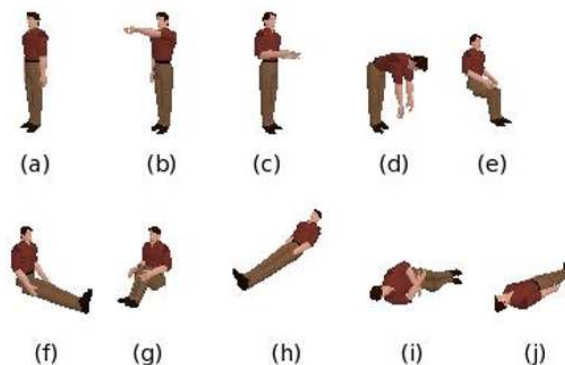
Fig. 4. The proposed 3D human postures

3.3 Activity Recognition

In this section, we firstly describe the video event recognition algorithm and secondly we present the proposed video event models .

Video Event Recognition Algorithm. The video event recognition algorithm detects which event is happening from a stream of observed persons

tracked by a vision component at each instant. The recognition process takes as input the a priori knowledge of the scene and the event models.

An event is composed of actors, sub-events and constraints. An actor can be a person tracked as a mobile object by the vision component or a static object of the observed environment like a chair. A person is represented by her/his characteristics: her/his position in the observed environment, width, velocity,... A static object of the environment is defined by a priori knowledge and can be either a zone of interest (e.g. the entrance zone) or a piece of equipment (a 3D object such as a table). A zone is represented by its vertices and a piece of equipment is represented by a 3D bounding box. The zones and the equipment constitute the scene context of the observed environment.

To recognize the pre-defined events at each instant, the algorithm verifies that the sub-events are recognized and the constraints are satisfied.

The video event recognition algorithm is based on the method described in [15]. We have adapted this algorithm to detect in real time some activities of interest by adding posture.

Video Event Models. The **event models** are defined using an event description language designed in a generic framework [16]. The video event model corresponds to the modeling of all the knowledge used by the system to detect video events occurring in the scene. The description of this knowledge is declarative and intuitive (in natural terms), so that the experts of the application domain can easily define and modify it. Four types of video events (called **components**) can be defined: primitive states, composite states, primitive events and composite events. A state describes a stable situation in time characterizing one or several physical objects (i.e. actors). A **primitive state** (e.g. a person is located inside a zone) corresponds to a perceptual property directly computed by the vision components. A **composite state** is a combination of primitive states. An event is an activity containing at least a change of state values between two consecutive times. A **primitive event** corresponds to a change of primitive state values (e.g. a person changes a zone). A **composite event** is a combination of primitive states and/or primitive events (e.g. preparing meal). As general model, a video event model is composed of five parts: "**physical objects**" involved in the event (e.g. person, equipment, zones of interest), "**components**" corresponding to the sub-events composing the event, "**forbidden components**" corresponding to the events which should not occur during the main event, "**constraints**" are conditions between the physical objects and/or the components (including symbolic, logical, spatial and temporal constraints including Allens interval algebra operators [17]), and "**alarms**" describe the actions to be taken when the event is recognized.

In the framework of homecare monitoring, in collaboration with gerontologists, we have modeled several primitive states, primitive events and composite events. First we are interesting in modeling event characteristic of critical situations such as falling down. Second, these events aim at detecting abnormal changes of behavior patterns such as depression. Given these objectives we have selected the activities that can be detected using video cameras. For instance, the detection of "**gas stove on**" when a person is doing a different activity for a long time is interesting but cannot be easily detected by only video cameras and requires

additional information such as the one provided by environmental sensors (e.g. gas consumption sensor). In this paper we are focusing only on video cameras and contributions with other environmental sensors for activity recognition belong to other ongoing work.

In this work, we have modeled **thirty four video events**. In particular, we have defined fourteen primitives states, four of them are related to the location of the person in the scene (e.g. inside kitchen, inside livingroom) and the ten remaining are related to the proposed 3D key human postures. We have defined also four primitive events related to the combination of these primitive states: **"standing up"** which represents a change state from sitting or slumping to standing, **"sitting down"** which represents a change state from standing, or bending to sitting on a chair, **"sitting up"** represents a change state from lying to sitting on the floor, and **"lying down"** which represents a change state from standing or sitting on the floor to lying. We have defined also six primitive events such as: stay in kitchen, stay in livingroom. These primitive states and events are used to define more composite events.

For this study, we have modeled ten composite events. In this paper, we present just two of them: **"feeling faint"** and **"falling down"**.

There are different visual definition for describing a person falling down. Thus, we have modeled the event "falling down" with three models:

Falling down 1: A change state from standing, bending, sitting on the floor(with flexed or outstretched legs) and lying (with flexed or outstretched legs).

Falling down 2: A change state from standing, and lying (with flexed or outstretched legs).

Falling down 3: A change state from standing, bending and lying (with flexed or outstretched legs).

The model of the "feeling faint" event is shown bellow. The "feeling faint" model contains three 3D human postures components, involves one person and additional constraints between these components.

```
CompositeEvent(PersonFeelingFaint,
PhysicalObjects( (p: Person) )
Components( (pStand: PrimitiveState Standing(p))
(pBend: PrimitiveState Bending(p))
(pSit: PrimitiveState Sitting_Outstretched_Legs(p)) )
Constraints( (pStand; pBend; pSit)
(pSit's Duration >=10))
Alarm(AText("Person is Feeling Faint")
AType("URGENT")) )
```

<div align="center">"Feeling faint" model</div>

The following text shows an example of the definition of the model "falling down 1".

```
CompositeEvent(PersonFallingDown1,
PhysicalObjects( (p: Person) )
Components( (pStand: PrimitiveState Standing(p))
```

```
(pBend: PrimitiveState Bending(p))
(pSit: PrimitiveState Sitting_Flexed_Legs(p))
(pLay: PrimitiveState Lying_Outstretched_Legs(p)) )
Constraints( (pSit before_meet p_Lay)
(pLay's Duration >=50))
Alarm(AText("Person is Falling Down")
AType("VERYURGENT")) )
```

<div align="center">"Falling down 1" model</div>

In this approach we have proposed ten 3D key human postures and thirty four video event models useful to recognize a set of ADLs of elderly living alone in her/his own home. In the next section we present experiments we have done in the Gerhome laboratory and the obtained results.

4 Results and Evaluation

This section describes and discusses the experimental results. First, we describe the experimental site we have used to validate our approach and models. Then we show and discuss the results of activity recognition.

4.1 Experimental Site

Developing and testing the impact of the activity monitoring solutions requires a realistic near-life environment in which training and evaluation can be performed. To attain this goal we have set up an experimental laboratory (Gerhome laboratory) to analyze and evaluate our approach. This laboratory is located in the CSTB (Centre Scientifique de Techniques du Batiment) at Sophia Antipolis. It looks like a typical apartment of an elderly person: $41m^2$ with entrance, livingroom, bedroom, bathroom, and kitchen. The kitchen includes an electric stove, microwave oven, fridge, cupboards, and drawers. 4 video cameras are installed in Gerhome laboratry. One video camera is installed in the kitchen, two video cameras are installed in the livingroom and the last one is installed in the bedroom to detect and track the person in the apartment and to recognize her/his postures. This laboratory plays an important role in research and system development in the domain of activity monitoring and of assisted living. Firstly, it is used to collect data from the different installed video cameras. Secondly, it is used as a demonstration platform in order to visualize the system results. Finally, it is used to assess and test the usability of the system with elderly. Currently, in this experiment, we have collected and processed data acquired by one video camera. The 3D visualization of Gerhome laboratory is illustrated in Figure 5.

4.2 Experimental Results

To validate our models, we have performed a set of human behaviors in the Gerhome laboratory. For this experiment, we have acquired ten videos with one

human actor. The duration of each video is about ten minutes and each video contains about 4800 frames (about eight frames per second).

For performance evaluation, we use classical metrics. When the system correctly claims that an activity occurs, a true positive (TP) is scored; a false positive (FP) is scored when an incorrect activity is claimed. If an activity occurs and the system does not report it, a false negative (FN) is scored. We then used the precision and sensitivity standard metrics to summarize the system effectiveness. Precision is the ratio TP/(TP + FP), and sensitivity is the ratio TP/(TP + FN).

The results of the recognition of the primitive states and events are presented in table 1. The primitive states "in the kitchen" and "in the livingroom" are

States and events	Ground truth	#TP	#FN	#FP	Precision	Sensitivity
In the kitchen	45	40	5	3	93%	88%
In the livingroom	35	32	3	5	86%	91%
Standing (a, b, c)	120	95	25	20	82%	79%
Sitting(e, f, g)	80	58	22	18	76%	72%
Slumping(h)	35	25	10	15	62%	71%
Lying (i, j)	6	4	2	2	66%	66%
Bending (d)	92	66	26	30	68%	71%
Standing up	57	36	21	6	85%	63%
Sitting down	65	41	24	8	83%	63%
Sitting up	6	4	2	1	80%	66%

Table 1. Results for recognition of a set of primitive states and events

well recognized by video cameras. The few errors in the recognition occur at the border between livingroom and kitchen. These errors are due to noise and shadow problems.

The preliminary results of the recognition of the different postures (a, b, c, e, f, g, h, i, j) are encouraging. The errors in the recognition of these postures occur when the system mixes the recognized postures (e.g. the bending posture instead the sitting one). These errors are due to the segmentation errors (shadow, light change, ...) and to object occlusions. To solve these errors, we plan to use temporal filtering in the posture recognition process.

We show in the figure 5 the recognition of the localization of the person inside livingroom and the recognition of the posture "sitting in the floor with outstretched legs". In the ten acquired videos, we have filmed one "falling down" event and two "feeling faint" events which have been correctly recognized.
Figure 6 and figure 7 show respectively the camera view and the 3D visualization of the recognition of the "feeling faint" event. Figure 8 and figure 9 show

Fig. 5. Recognition of two video events in Gerhome laboratory. (a) 3D visualization of the experimental site "Gerhome", (b) person is in the livingroom, (c) person is sitting in the floor with outstretched legs

Fig. 6. Recognition of the "feeling faint" event

Fig. 7. 3D visualization of the recognition of the "feeling faint" event

respectively the camera view and the 3D visualization of the recognition of the "falling down" event.

5 Conclusions and Future Works

In this paper we have described a cognitive vision approach to recognize a set of activities of interest of elderly at home by using ten 3D key human postures. This approach takes as input only video data and produces as output the set of

Fig. 8. Recognition of the "falling down" event

Fig. 9. 3D visualization of the recognition of the "falling down" event

the recognized activities.

The first contribution of this work consists in the identifying and modeling of the ten 3D key human postures. These key postures are useful in the recognition of a set of normal and abnormal activities of elderly living alone at home. The second contribution of this work is the modeling of some activities of interest of elderly.

The proposed approach is currently experimented on small datasets, but we will next validate the performance of our approach on larger datasets (long videos on a long term period with different persons). Currently 14 different video sequences (more than 56 hours) have been recorded with elederly people in Gerhome laboratory. We plan two other different studies with elderly (+ 70 years) which will take place during the next six months. The first study will be performed in a clinical center (e.g. hospital, nursing home), and the second one in a free living environment (home of elderly people). We also plan to study some other postures and take account gestures in order to detect finer activities (e.g. kneeling).
Moreover, we envisage to analyze which sensors in addition to video cameras are the best for monitoring most activities of daily living. Our ultimate aim is to determine the best set of sensors according to various criteria such as cost, number of house occupants and presence of pets.

We also envisage to facilitate incorporation of new sensors by developing a generic model of intelligent sensor and to add the data uncertainty and imprecision on sensor measurement analysis.

References

1. Boulay, B., Bremond, F., Thonnat, M.: Applying 3d human model in a posture recognition system. Pattern Recognition Letter. (2006.)
2. C. Wren, A. Azarbayejani, T.D., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence (1997) 780–785
3. Fujiyoshi, H., Lipton, A.J., Kanade, T.: Real-time human motion analysis by image skeletonisation. IEICE Trans. Inf. & Syst (January 2004)
4. Tao, Z., Ram, N.: Tracking multiple humans in complex situations. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9) (September 2004)
5. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE Transactions on Patten Analysis and Machine Intelligence **28**(1) (January 2006)
6. Ayers, D., Shah, M.: Monitoring human behavior from video taken in an office environment. In: Image and Vision Computing. (2001.)
7. Yacoob, Y., Black, M.J.: Parameterized modeling and recognition of activities. In: ICCV. (1998.)
8. Jebara, T., Pentland, A.: Action reaction learning: Analysis and synthesis of human behavior. In: IEEE Workshop on the Interpretation of Visual Motion. (1998.)
9. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. In: Computer Vision and Image Understanding. (1999.)
10. D. J. Moore, I.A.E., Hayes, M.H.: Exploiting human actions and object context for recognition tasks. In: ICCV. (1999.)
11. Chomat, O., Crowley, J.: Probabilistic recognition of activity using local appearance. In: International Conference on Computer Vision and Pattern Recognition (CVPR)., Vancouver, Canada. (June 1999.)
12. Intille, S., Bobick, A.: Recognizing planned, multi-person action. In: Computer Vision and Image Understanding. (2001.)
13. J. Hoey, A. V. Bertoldi, P.P., Mihailidis, A.: Assisting persons with dementia during handwashing using a partially observable markov decision process. In: International Conference on Computer Vision Systems (ICVS)., Germany. (March 2007.)
14. Ivanov, Y., Bobick, A.: Recognition of visual activities and interactions by stochastic parsing. In: IEEE Transactions on Patterns Analysis and Machine Intelligence. (2000)
15. Vu, V., Bremond, F., Thonnat, M.: Automatic video interpretation: A novel algorithm based for temporal scenario recognition. In: The Eighteenth International Joint Conference on Artificial Intelligence. (September 9-15 2003.)
16. Avanzi, A., Bremond, F., Tornieri, C., Thonnat, M.: Design and assessement of an intelligent activity monitoring platform. EURASIP Journal on Applied Signal Processing, Special Issue on 'Advances in Intelligent Vision Systems: Methods and Applications'. (August. 2005)
17. Allen, J.F.: Maintaining knowledge about temporal intervals. In: In Communications of the ACM. (1983)