# A HHMM-Based Approach for Robust Fall Detection

Nicolas Thome
LIRIS - SAS Foxstream*
5, avenue Pierre Mendes France
69676 Bron Cedex France
Email: n.thome@foxstream.fr

Serge Miguet
LIRIS
5, avenue Pierre Mendes France
69676 Bron Cedex France
Email: Serge.Miguet@univ-lyon2.fr

***Abstract*— Automatic detection of a falling person in video sequences is an important part of future pervasive home monitoring systems. We propose here a robust method to achieve this goal. Motion is modeled by a Hierarchical Hidden Markov Model (HHMM) whose first layer states are related to the orientation of the tracked person. Finding a consistent way for robustly linking the observation vector to the human poses is the heart of our contribution. In that sense, we carefully study the relationship between angles in the 3D world and their projection onto the image plane. After performing an initial image metric rectification , we derive theoretical properties making it possible to bound the error angle introduced by the image formation process for a standing posture. This allows us to confidently identify other poses as "non-standing" ones, and thus to robustly analyze pose sequences against a given motion model. Several results illustrate the efficiency of the algorithm by pointing out its ability to accurately recognize a person falling down from another walking or sitting, as well as its capacity to run in an unspecified configuration.**

## I. Introduction

With the population growing older and the increasing number of people living alone, supportive home environments able to monitor automatically human activities are likely to widespread due to their promising ability helping elderly people living alone and reducing healthcare costs. At the moment home monitoring systems include personal embedded sensors, low-level sensors and video sensors. Embedded sensors such as fall detectors may produce false alarms. Moreover, they have to always be worn by the occupants which depends on the person's ability and willingness to do it. Simple remote sensors are broadcasted throughout the home and information is gathered to take a decision. However, these low-level data may produce crude data difficult to interpret. On the other hand, cameras offer semantic information. The main issue of using computer vision for home monitoring is related to the acceptability and privacy surrounding it. To circumvent such problems, we must clearly state that the video may be analyzed without any image is transmitted.



In this paper we propose a video analysis based method for monitoring human activities, with a particular interest to the problem of fall detection. The remainder of the paper is organized as follows. Section II proposes a state of the art of the existing methods for human behavior interpretation. Sections III and IV describe in more details our approach by developing how motion analysis is performed. Section V presents several results illustrating the efficiency of the proposed approach and showing its advantages with respect to previous works. Finally we conclude in section VI and propose some directions for future works.

## II. State Of The Art

Analyzing behavior with video, including a semantic description of human activity, corresponds to the high-level process in computer vision. Low-level part is usually composed of motion segmentation and feature extraction, as the intermediate level corresponds to tracking and is dedicated to link the previous modules. Approaches for analyzing video may be classified into two groups, depending on the use of a model. Non-Model-Based methods [1], [23] aim to automatically cluster different kinds of events. They are usually dedicated to detect "unusual events" by taking advantage of the fact that they are "difficult to describe but easy to recognize". Model-Based Methods explicitly describe a given type of movement. The first attempt to perform this task relies on building temporal templates [4]. Shortcomings of this approach are related to viewpoint and time variability dependence as well as sensitivity to noise in the observations. Alternatively, Hidden Markov Models (HMMs) [15] have been widely used for tackling simple behaviours such as gestures or gait recognition [12], [17], [21], and recently tested in the case of fall detection [20]. Due to the Markovian assumption, HMM are limited to model simple motions for one single human. Thus, other extensions to the basic HMM have been used such as the Coupled Hidden Markov Models (CHMMs) for modeling human behaviours and interactions [14], and variable length Markov models (VLMMs) to locally optimize the size of behaviour models [6]. Finally, requirements for scalable systems for high level understanding and semantically rich behaviour recognitions led to study frameworks that use

the inherent hierarchical structure of motion. In that sense, sophisticated stochastic methods have been used to model the combination between elementary behavioral pattern detected by the previous methods, leading to the highest level of the interpretation module. It has been accomplished by the development of Dynamic Bayesian Networks (DBN) [8], Abstract Hidden Markov Models (AHMM) [2] or Hierarchical Hidden Markov Models (HHMM) [13] and stochastic grammars (including Stochastic Context Free Grammar (SCFG) [22]). We now give more details on some recent works being the most related to our.

*Previous Work*

The Simbad project [16] uses infra-red sensors, and fall detection is performed by a neural network making use of vertical velocity as input. Nevertheless, the requirement to fast movements recognition may lead to a sensitivity to noise tending to send false alarms.

Nait-Charif and McKenna [7] propose a method for automatically extracting motion trajectory and providing human-readable summarization of activity and detection of unusual inactivity. Tracking is performed with an omnidirectional camera by a particle filter on the ellipse parameters describing human posture. Fall is detected as a deviation to usual activity. However, no information about the pose of the person or his motion dynamic is taken into account. The Ubisense project [10] proposes to classify human poses by computing orientation of the detected blob. But the posture sequences analysis is still missing for proper motion clustering.

Toreyin et al. [20] suggest a method for fall detection by making use of an HMM using both audio and video. For the vision part of the approach, the aspect ratio of the bounding box of the moving region detected with a standard camera is passed to the motion interpretation module. More precisely, its wavelet transform is used as input feature for the HMM. Using conjointly video and audio cues seems to be well founded. Defining HMM states in the frequency domain is interesting because it makes explicit use of feature motion. However, no study is proposed about the relevance of using bounding box aspect ratio, particularly its dependance to the relative position of the camera and the tracked people. It is clear for example that the aspect ratio observed in the image corresponding to a standing posture is quite different for a vertically-oriented optical axis than for an horizontally-oriented one. The problem remains for the wavelet coefficients, and a given sequence for a walking person might be confused with another one for a falling person under certain adverse acquisition conditions.

## III. Approach Overview

We choose to develop a model-based approach to monitor senior citizen activities. The reason is that we want to be able to detect unusual suspicious events. But their ranges are in fact quite limited and generally - the case of fall detection is a good example - not so ”hard to describe”. Thus, we use a HHMM as a generative model against which each current motion will be matched.

### A. Model Architecture

We use a simple kind of HHMM with two layers whose architecture is depicted in figure 1. The observation vector is composed of the image angle sequences, corresponding to the orientation of the detected human blob (see IV-A), after a metric rectification is applied. It provides a measurement of its degree of verticality, in the 3D world. The states of the first layer correspond to the human postures. In our falling detection application, two states are used, one corresponding to an upright standing pose and the other to a lying one. The first Layer motion features, that we call ”Behavioral Pattern” or ”elementary motion pattern”, correspond to movements whose temporal extent is relatively small. It is dedicated to detecting sudden changes such as falls. These behavioral patterns constitute the states for the second layer motion. This latter corresponds to a global motion that has a larger temporal extent and that we denote ”Behavior”. It is interpreted in the HHMM formalism as being the most probable motion model that explain a given sequence of elementary patterns. For the sake of clarity, we discriminate typographically a global motion from a behavioral pattern by using capital letters for the former and quotes for the latter. For example, a FALL is supposed to be composed of a sequence of several ”Is Walking” elementary pattern, followed by some ”Is Falling”, and finally many ”Is Lengthened”.

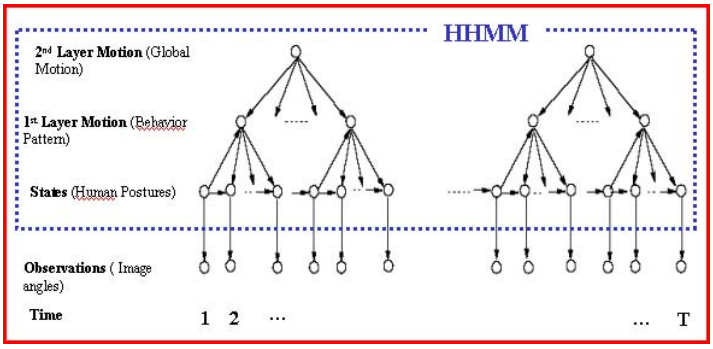


Fig. 1. HHMM Architecture

*1) Main Contribution:* Our contribution to the development of a robust motion analysis system breaks down into two points. Firstly, we propose a theoretical study of the relationship between angles in the 3D world and their projection onto the image plane. It makes it possible to optimize the link between the states (standing/lying poses) and the observation vectors (rectified blob angles). Secondly, the use of the HHMM is a robust way for interpreting motions, particularly in applications such as fall detection. Although the use of a sophisticated statistical model such as HHMM for simple motion analysis like fall detection may appear to be unnecessary, it actually provides a useful way to model a given motion as a composition of elementary behavioral patterns. Hierarchical architecture makes the model parameters easily human-readable as well as the proximity of the first layer states to human concept makes the model easy to interpret.

Moreover, the two levels of abstraction constitutes a powerful way to filter out possible false alarms that must occur using a single level model (results are shown in section V illustrating this idea). Finally, the increase of the time complexity due to the use of the HHMM is very small (it corresponds to less than 1% of the total computational costs).

Section IV gives decomposes as follows. We explain in IV-A how the observation vector is extracted from the image sequences. IV-B is the heart of the proposed approach, where the analysis between human posture and image angles is carried out. The two layer motion models are presented in IV-C and IV-D. Results illustrating the approach efficiency are presented in V. Finally Section V concludes the paper.

## IV. HHMM Description

### A. Determining Observation vector

The steps leading to the observation vector computation are the followings. First, human blobs are localized in the image and tracked over time (see IV-A.1). Then, the principal axis corresponding to each human blob is computed (IV-A.2). Finally, a rectification matrix is determined, dedicated to measure metric properties in a given plane in the image. This leads to the observation vector, corresponding to the angle between the vertical direction and the rectified principal axis.

*1) Isolating and Tracking Humans:* The first step of the system consists in applying a motion segmentation algorithm, leading to a binary map where moving and static pixels are labeled. This is achieved by modeling the background for each pixel by a mixture of Gaussians (first introduced in [18]). In addition, we make it possible to not assign the "moving" label to shadow pixels by using a color space invariant in luminance. Finally connected component analysis is applied to get sufficiently large regions where motion occurs. Then, a simple region-based tracking strategy is developed to match regions detected in one frame and in the subsequent ones. In particular, it enables us to detect regions corresponding to single humans. These regions are robustly tracked over time by using an articulated appearance model that constitutes a feature used to perform matching in difficult situations. For further details about this part of the approach, the reader can refer to [19].

*2) Feature Extraction:* Once a blob has consistently been tracked, the orientation of the principal axis is computed. This is traditionally performed by computing PCA on the silhouette or determining its minimal bounding rectangle. We use the latter method, because of its ability to better deal with non-convex, irregular and complex contour shapes, which is the case when tracking humans.

*3) Applying a Metric Rectification:* The proposed rectification is actually a preprocessing step that is performed offline. Otherwise, it is applied to the principal axis at this stage of the algorithm. Metric rectification of perspective images of planes has been recently widely studied by its ability to recover metric properties such as angle measurements in the world plane from a perspective image [9]. Representing points and lines in the plane with 3-D homogeneous coordinates, the homography $R$ mapping points on the image plane $X$ to points on the world plane $X'$ may be decomposed into a concatenation of three matrices, representing similarity, affine and pure projective transformation :

$$X' = RX = R_S R_A R_P \tag{1}$$

Thus, performing a metric rectification consists in determining $R_P$ and $R_A$. It discards perspective and affine effects from the image formation process, i.e the image of the plane is equivalent to the plane in the world up to a similarity (rotation, scaling and translation). Many approaches have been proposed for performing this kind of rectification. An exhaustive review of the method is beyond the scope of the paper and the reader can refer to [5]. In our context, we choose the following strategy, illustrated in figure 2. Two sets of horizontal and vertical lines on the world plane are manually selected on the image plane (see figure I.a)). This makes it possible to determine two orthogonal vanishing points by computing their intersection on the image plane in the least square sense. Thus the vanishing line is computed and affine rectification can be applied (see figure I.b)). To perform metric rectification, we additionally enforce two more conditions, illustrated in figure II.a). First, we impose horizontal and vertical lines to effectively be horizontal and vertical on the image plane. In figure II.a), blue and green line represent a given horizontal and vertical line of the previous sets. Red lines show the direction to which they are mapped. Finally, purple and yellow parts (in figure II.a)) of the two previous lines correspond to known distances (and thus a known aspect ratio) in the world plane. We then apply the needed correction in order to match the image aspect ratio to the real world one. Eventually, sub-figure II.b) shows the metrically rectified image.

We claim that our method is well-suited for performing the needed rectification in an indoor environment. Indeed, horizontal and vertical lines on the image are easily available in a room. Moreover our approach only requires a single pair of orthogonal directions and a known ratio in the world. It is much easier for example than requiring two pairs of orthogonal directions, which are difficult to find in real environments. The metric rectification is applied to the previously computed principal axis, and its angle with respect to the vertical direction constitutes the observation vector for the HHMM. This feature, that is simple and fast to compute, is additionally robust to occlusions. Indeed, someone standing is distinguishable from someone lying as long as the principal axis that is detected in the image corresponds to the head/feet direction, which allows large occlusions.

Finally, the angle is tracked itself over time. It consists in making the assumption that the angles in two subsequent frames must be relatively close from one another. It enables two major improvements in terms of robustness. First, it makes it possible to remove "outliers" (due to low-level process part errors) from the analysis loop. Moreover it makes it possible to disambiguate the principal axis determination in some adverse conditions. It occurs for example when the width and height of the computed minimal rectangle reach the same value. In

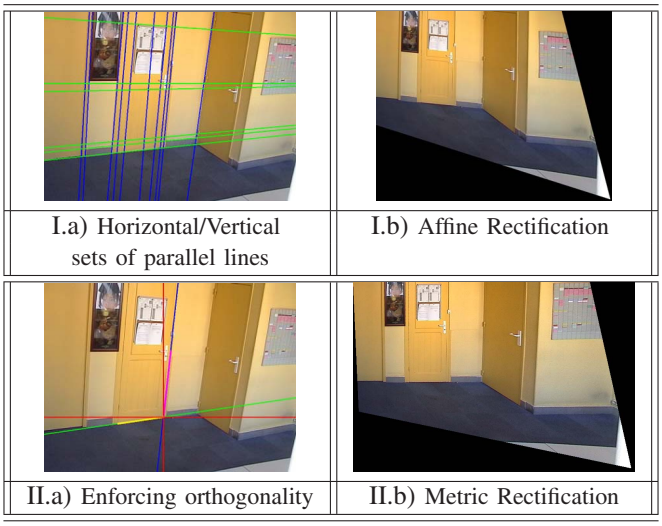


Fig. 2. Metric Rectification Performing.

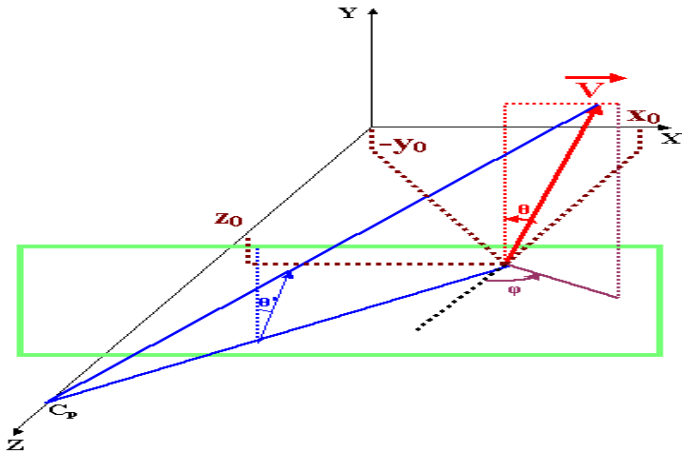


Fig. 3. Angle Projection

that case, the principal axis is likely to be instable, and switch to one orthogonal direction to the other. Tracking the angle as a feature makes it possible to overcome this shortcoming.

### B. States Definition

Orientation of the rectified principal axis on the image having been computed, the HHMM then requires to define states. We choose to use 3D human postures for the first layer states (standing and lying for the HHMM modeling a fall). As the rectification is performed on a plan $\Pi$, 3D angles between two directions included in $\Pi$ correspond to angles measured on the image. For our purpose, the key point is to choose $\Pi$ so that it contains the vertical direction. Thus, someone standing upright in the world appears to be standing on the image plane as well. Moreover, after applying the metric rectification, everything occurs as if the camera were in front of the plane $\Pi$ were the rectification took place. It makes it possible to derive theoretical relationships between the 3D angle corresponding to the head/feet direction in the world and the angle computed in the image.

*a) From Image Angles to 3D Poses:* We now give more details on how 3D human postures may be inferred from image angle measurements. After the metric rectification, the configuration corresponds to this illustrated in figure 3 (the wall where the rectification has been performed in figure 2 corresponds here to the $(0XY)$ plan). The red vector $\overrightarrow{V}$ represents the principal axis of the person in the 3D world. Its origin is located at $(x_0, -y_0, z_0)$ and corresponds to the feet of the person, and its extremity gives the position of the head. The $\theta$ angle corresponds to the 3D deviation from the vertical axis $(OY)$, and $\varphi$ is the angle between the $(OZ)$ axis and the projection of $\overrightarrow{V}$ on the $(OXZ)$ plane. The image plane intersects the $(OZ)$ axis at $z = d$, and the center of projection $Cp$ coordinates are $(0, 0, (d + f))$, where $f$ is the camera focal length. The $\theta'$ angle corresponds to the deviation from the vertical axis on the image plane. Assuming pinhole model for the camera, we get :

**Property 1.**

$$\tan(\theta') = \tan(\theta)\frac{(d + f - z_0)\sin(\varphi) + x_0\cos(\varphi)}{(d + f - z_0) - y_0\tan(\theta)\cos(\varphi)} \quad (2)$$

Equation 2 relates thus the angle $\theta$ in the 3D world to its projection $\theta'$ on the image. We can notice that $\theta'$ can take arbitrary values if $\theta$ is close to $\frac{\pi}{2}$. For example, if $\varphi = 0$ and $x_0 = 0$, $\theta' = 0$ whatever the value of $\theta$ : someone lying on the floor at the center of the room with $\varphi = 0$ will be detected as standing on the image. However, we claim that if someone is in an approximate standing posture in the 3D world (i.e $\theta$ few different from 0) then he will appear to be in an approximate upright standing pose as well on the image. This is expressed by the following property :

**Property 2.**

$$\textit{Provided that :} \begin{cases} z_0 < S \\ x_0 \in [-L; L] \\ \theta \in [0; \theta_{max}] \textit{ and } \theta_{max} < \arctan(\frac{d+f-S}{y_0}) \end{cases}$$

$$\tan(\theta') \in [\frac{-\tan(\theta_{max})(d + f + L)}{d + f - S - y_0\tan(\theta_{max})}; \frac{\tan(\theta_{max})(d + f + L)}{d + f - S - y_0\tan(\theta_{max})}] \quad (3)$$

Equation 3 makes it possible to bound the angular error introduced by the image formation process. $S$ and $L$ are positive thresholds, defining the intervals for $z_0$ and $x_0$ for which the property 2 is fulfilled. Qualitatively, the Property 2 may be reformulated as follows : if someone moves in a room in an *approximate upright standing position* ($\theta \in [0; \theta_{max}]$) and if they are *not too close to the camera* (meaning $z_0 < S$), then they will be seen in an *approximate standing posture* on the camera image plane as well. The assumption $z_0 < S$ is important for our purpose, and the value of the threshold $S$ directly determines the interval $[0; \theta_{max}]$ in which the property holds. The larger the distance, the larger the interval $[0; \theta_{max}]$ may be and the smaller the possible value ranges for $\theta'$ given by equation 3 are. However, $x_0 \in [-L; L]$ is not a very

restrictive additional assumption, since the camera field of view is intrinsically limited.

*b) Defining States:* Property 2 enables us to define states for our first layer HHMM. We actually take advantage of its contrapose, stating that if the angle measurement $\theta'$ on the image does not stay in the range defined by equation 3, then the 3D angle $\theta$ must not belong to $[0; \theta_{max}]$. Lying postures are thus defined in the image as those fulfilling :

$$|\tan(\theta')| > \frac{\tan(\theta_{max})(d+f+L)}{d+f-S-y_0\tan(\theta_{max})} \quad (4)$$

Let us define $\theta'_T$ by the following way :
$|\tan(\theta'_T)| = \frac{\tan(\theta_{max})(d+f+L)}{d+f-S-y_0\tan(\theta_{max})}$. It is the threshold that discriminate standing from lying postures based on the image measurements. Provided that the low level parts of the system leading to the computation of the angle are properly carried out, the property 2 is a proof that the states corresponding to lying poses are properly identified. Standing poses are thus defined as equation 4 is not satisfied (i.e $|\tan(\theta')| \leq |\tan(\theta'_T)|$). For the HHMM formalism, we define a conditional density probability observation for each of the two states. For the standing state, we use a single Gaussian for modeling the probability density function with 0 mean value. For the lying state, we use a mixture of two Gaussians, centered in $\frac{\pi}{2}$ and $-\frac{\pi}{2}$. The standard deviations are estimated so that the probability for standing and lying postures reach the same value at the interval bounds.

*c) Analyzing Labeling Errors:* Let us consider our model dedicated to fall detection. With respect to equation 4, we define "positive" and "negative" situations as lying and standing postures, respectively. For analyzing the relevance of the state definition, we use the following notations. A true "negative" corresponds to the detection of a standing posture in the image, as the 3D pose is indeed a standing one ($\theta \in [0; \theta_{max}]$). A false "negative" corresponds to the detection of a standing posture in the image, although the 3D pose is actually a lying one ($\theta \notin [0; \theta_{max}]$). True and false "positive" are similarly defined for detected lying poses. Equation 4 makes it possible to detect all negatives (100% true negatives). However, it will lead to some false positives in some adverse conditions : remember pathological case where $\varphi = 0$ and $x_0 = 0$. We propose to go further in the analysis and to determine, for positive situations, the conditions under which the lying poses are properly identified. We can intuitively think that the further $\varphi$ is from $\frac{\pi}{2}$, the larger will be the probability to detect a positive is. We derive the following property :

**Property 3.** *If $\theta = \frac{\pi}{2}$ and under property 2 assumptions,* $\tan(\theta') > \tan(\theta'_T) \Leftrightarrow (\tan(\varphi) > \tan(\theta'_T)\frac{y_0-x_0}{d+f-z_0}$ *or* $\tan(\varphi) < \tan(\theta'_T)\frac{-(y_0+x_0)}{d+f-z_0})$

Property 3 enables us to determine, for each space location of the feet $(x_0, y_0, z_0)$ and angle $\varphi$, if a lying posture in the world ($\theta$=$\frac{\pi}{2}$) is properly identified in the image thanks to $\theta'$ measurments. This Property may moreover be used to determine the needed value for $\varphi$, for each $(x_0, y_0, z_0)$, for which all false positive are removed. This is a vital knowledge, that can be used to reduce the false positive rate (see IV-B.0.e).

*d) Validation by Simulation:* We define the HHMM states in accordance to equation 4. As it has been proved that the false positive are by essence impossible, the score that remains to be computed is the false negative rate. Let us that denote $P_{FP}$ the false negative rate for $\theta = \frac{\Pi}{2}$ (i.e lying postures in the world). The following formula constitutes an upper bound for $P_{FP}$ :

$$P_{FP} = \frac{1}{M}\int\int\int\int_{\varphi=0}^{\frac{\pi}{2}} f(x_0, y_0, z_0, \varphi)dx_0\, dy_0\, dz_0\, d\varphi \quad (5)$$

The integration has to be performed for $(x_0, y_0, z_0)$ inside the field of view and $\varphi \in [0; 2\pi]$.
$f(x_0, y_0, z_0, \varphi) = 0$ if property 3 is satisfied and 1 otherwise (in that case we have a false positive at $(x_0, y_0, z_0, \varphi)$).
M is a normalization constant chosen so that $P_{FP} = 1$ if $f(x_0, y_0, z_0, \theta, \varphi) = 1\ \forall\ (x_0, y_0, z_0, \theta, \varphi)$.
However, analytically integrating equation 5 is quite complicated. We thus suggest to simulate the camera projection in a given configuration, and then to numerically evaluate $P_{FP}$. Beyond the computational facility that the simulation offers, its significance is justified as follows. This simulation makes it possible to get numerical values for all of the previous thresholds, and potentially to optimize them to a given configuration.
The first thing to do is to get camera calibration parameters. Taking advantage on the fact that we are interested in indoor environment, we choose to use three orthogonal vanishing points to calibrate camera. This kind of approach has been introduced by Caprile and Torre [3] for determining intrinsic parameters and recently extended by Fengjun et al. [11] for computing extrinsic parameters as well as for suggesting an approximate method applicable to outdoor environments. We use an algorithm similar to [11], by assuming unit aspect ratio and zero skew, which is common for standard CCD cameras. The calibration procedure provides principal point $(U_0, V_0)$, focal length $f$ as intrinsic parameters, height of the camera $H_C$, the three rotation angles $\alpha, \beta, \gamma$ as extrinsic parameters.
The calibration being performed, the camera is determined to be at height $H_c$. Thus $y_0 = H_c$ and the false positive rate does not depend on $y_0$ but only on $x_0$, $z_0$ and $\varphi$. The simulation results may thus be illustrated as presented in figure 4. This corresponds to a "from above" view, meaning that we look in the direction given by $-y$ axis on figure 3. We present two kinds of simulation : validation for the small angles corresponding to standing poses (i.e $\theta \in [0; \theta_{max}]$, Figures (a) and (b)) and for lying postures ($\theta = \frac{\pi}{2}$, Figures (c) and (d)). For the small angle validation, we then make $x$, $z$, $\theta$ and $\varphi$ vary in $[-L; L]$, $[0; S]$, $[0; \theta_{max}]$ and $[0; 2\Pi]$. $\theta_{max}$ is set to $\frac{\Pi}{10}$, defining the upright standing posture in the world. Concerning the study on a lying posture in the world, we keep the previous value and set $\theta$ to $\frac{\Pi}{2}$. The other values (given in millimeters) for the parameters are : $d = 4260mm$, $S = 2000mm$ , $L = 1395mm$. For each $x_0$, $z_0$, $\theta$ and $\varphi$, the goal is to get the corresponding value for $\theta'$ (thanks to equation

2), and to check if it is inside the interval provided by equation 3, that we denote $I_t = [-\theta'_T; \theta'_T]$ ($I_t = [-0.85; 0.85]$ with the current parameter values). The red horizontal line identifies $S$ value under which equation 3 can not be violated. The green horizontal line identifies, in the given configuration, the $S'$ value from which equation 3 begins to be violated. As seen in figure (a), $S'$ is significantly larger than $S$. We can then take advantage of this observation to reduce the size of the interval for $\theta'$ by tuning $\theta'$ until $S$=$S'$. This is shown in figure (b). The new interval for $\theta'$ is denoted $Id$ and is evaluated to be $Id = [-0.37; 0.37]$. We can notice that this simulation result leads again to a significantly smaller interval than the one resulting from the theoretical bound given by equation 3. It is of great interest for our purpose. In our configuration, we can now claim that each time an image angle presents a deviation larger than $0.37$ from the vertical line, the corresponding 3D angle in the world must present a deviation larger than $\frac{\Pi}{10}$. The validation for lying postures is illustrated in figures (c) and (d). The aim is to detect the times where $\theta'$ leads to detect a standing posture, while $\theta = \frac{\Pi}{2}$ (for that we use property 3). The computation may be performed using the theoretical interval $I_t = [-0.85; 0.85]$ (Figure (c)) or the just above determined one $I_d = [-0.37; 0.37]$ (Figure (d)). Additionally, these simulations allow to numerically compute the false negatives, i.e numerically integrating equation 5. We find about $38\%$ in Figure (c) conditions and $15\%$ in Figure (d) conditions . This concludes to prove the efficiency of the defined states to discriminate a standing pose from a lying one. Indeed, a lying posture in the image necessary means a similar one in the world ; and our system makes it possible to statistically detect $85\%$ of the overall possible falls, which is very good for a single-view computer vision based system.

*e) Discussion:* We estimate the discriminability of the image angle to separate standing from lying poses to be 85 %. This score has been established after camera calibration. Thus, it depends on the calibration parameters, but not on the position of the person in the scene. In that sense, the validation is more general than an experimentation in which standing/lying poses would have been manually labelled and compared to the computation of the first level states having image angle as input. Moreover, as Property 3 outputs the range value of $\varphi$ for which the algorithm fails in labelling the lying pose, what must be possible to do is to determine the optimal placement for an additional camera making all lying postures properly detected (100 % discriminability must be reachable by increasing the number of cameras). These ideas correspond to extensions of the proposed approach, but we claim that these improvements are made possible by the previous theoretical study carried out. Indeed, the derived properties enable us to grasp the meaning of the misclassified postures. We come back to this extension in section VI.

### C. First and Second Layer Motion Determination

The HHMM states correspond to the human poses in the scene. The first layer motion models, that we denote as *elementary behavioral pattern*, properly describe (or not) the

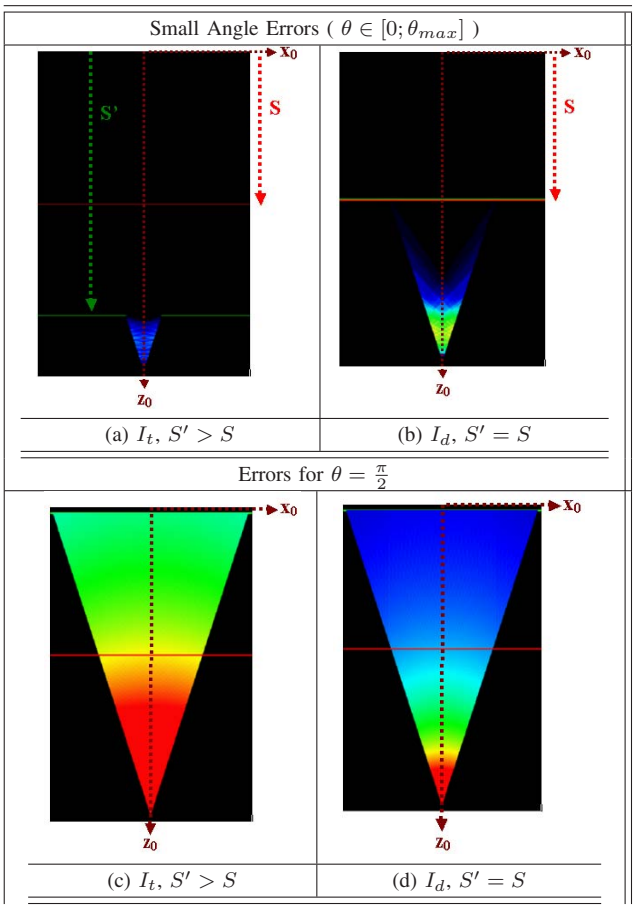


Fig. 4. Simulation Results

corresponding observations. The model ability to generate a given sequence of observations depends on its parameters (see IV-D). For the case of fall detection, we use the 3 followings elementary motion pattern : "Is Walking", "Is Falling" and "Is Lengthened". The motion analysis loop runs as follows. For each given observation sequence, we compute the most probable primitive event which may generate it, with the possible *unknown event* (by testing the most probable event again a given threshold). The best decoding state sequence (corresponding to 3D poses) may be determined by using Viterbi algorithm. The operation is recursively applied in the next level. In that case, the states correspond to the elementary pattern. We use two global motion model : FALL and WALK, and determine the one that best explains the sequence of primitive motions.

### D. Model Parameters Statement and Estimation

With the 2-layer hierarchical architecture proposed, the global motion model $\lambda$ is composed of the following set of parameters : $(\pi_j, A_j, B_j)$, where j corresponds to the layer index ($j \in 1, 2$) and i is the layer state index for layer j ($i \in [1, N_j]$, where $N_j$ is the state numbers of layer $j$). $\pi_j$ is the initial state probability vector for layer $j$. $A_j$ is the $N_j \times N_j$ state transition probabilities matrix. $B_j$ is the state relative observation probability. In the highest level, we

want our model to be able to represent a global motion as a combination of primitive events. We claim that the HHMM architecture with the image angle as observation leads to a very easy interpretable framework. For example, modeling an elementary fall in the first level can be accomplished by just requiring that this primitive event is able to generate a sequence of standing postures followed by lying ones. It is straightforward to manually set up the transition matrix for that purpose. The transition probability between a standing pose and a standing one must be quite small, the one between standing and lying must be quite large, the one between lying and lying must be very large and finally the one between lying and standing must be very small. Same kind of intuitive definitions are easily extended to the other primitive events as well as to the next hierarchy level. Of course, for more accurate modeling and recognition of complicated motions, EM Baum-Welch algorithm for parameters re-estimation might be used.

## V. Results

We present here some results illustrating the ability of the proposed model to analyze motion.
Figure 5 illustrates an example of using the HHMM. The different detected primitive events are shown at frame 40, 70, 90 and 180. Finally the global motion "FALL" (frame 180) is determined as being the most probable to be generated by a sequence of a walking-falling-lengthened motion patterns. Here, we point out the fact that the high level behavior recognition of the model is a great help for removing errors due to low-level part of the system. Especially when real time is required, motion segmentation is never perfect. Parts of the tracked object are likely to be missing when they look like the background. These segmentation errors are then propagated and may disturb the feature extraction (angle computation in our case) and then the elementary motion pattern as well. As we are detecting falls, being sudden changes, this noise sensitivity has actually no small probability to occur. Thus a system without a higher level validation would tend to wrongly detect the primitive "Is falling", sending false alarms. However, the global motion model enables us to overcome this problem. For detecting a "FALL" it actually requires that one (or more) "Is Falling" event is detected, that several "Is Walking" event were detected before and that a large number of "Is Lengthened" are detected after. This model has a much larger time extent, and wrongly detecting such an event because of noise is very improbable. The experimental validation part (V) gives quantitative results comforting this idea.

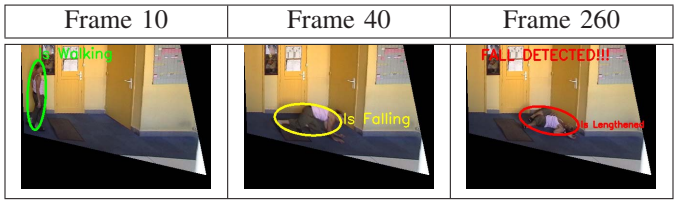


Fig. 5. Fall Detection with the HHMM

Figure 6 illustrates the case where the fall takes place in a direction relatively far from the rectified plane. It demonstrates the robustness of the approach, and the model ability to manage complex situations. Indeed, primitive events "Is Falling" and "Is Lengthened" are correctly recognized. This is a result of the lying state definition proposed in section IV-B.0.a.

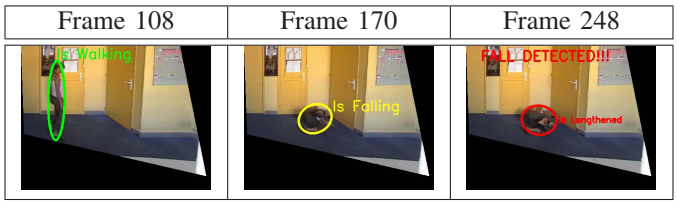


Fig. 6. Fall Detection with an arbitrary direction

Figure 7 proves the model ability to discriminate someone falling from someone just lying on the floor. We add a new primitive event called "Is Lying". With some examples of falls and some other examples of people lying, we train the first level HHMM parameters so that the example sets are properly clustered. The "Is Lying" and "Is Falling" models differ from their transition probability matrix, encoding the speed difference. As a result, the global motion "LENGHTENING" is recognized instead of "FALL" leading to the absence of alarm transmitted.

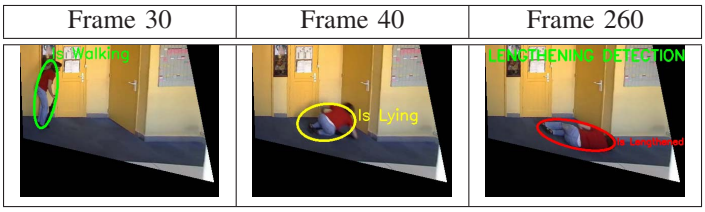


Fig. 7. The model captures motion speed

## Experimental Validation for Fall Detection

In order to validate the overall system performances, including segmentation, tracking and recognition, we apply the fall detection algorithm proposed to fifty cases of fall and fifty cases of walk. We try to use sequences that are relevant to evaluate the robustness of the system. Particularly, we take care in having different fall directions (angle $\varphi$ being uniformly distributed into $[0, 2\pi]$), in order to include examples that are a priori difficult to manage. Results are presented in table I. For real falling cases occurring we obtain a rate of 82 % correct detections (and then 18% of false negatives). For Non-Fall motions the system almost never send false alarms (98 % correct detections). As we can notice, the results obtained for the overall system are comparable to those corresponding only to the posture detection. It does not mean that the low level part of the system outputs perfect results, but rather demonstrates the capacity of the hierarchical motion model to filter out the low level errors.

Finally Figure 8 illustrates the use of the HHMM to recognize someone sitting. The HHMM structure has been

| Detection \ Truth | FALL | WALK |
|---|---|---|
| FALL | 41 | 1 |
| WALK | 9 | 59 |

TABLE I
EXPERIMENTAL RESULTS

enriched to be able to model a larger number of motions. First, a "seated" posture has been added, leading to a 3-state layer. We define it by partitioning the $\theta'$ space previously dedicated to the lying pose (see equation 4). This is entirely justified as the motion lies on a plane parallel to the one in which the metric rectification has been performed. Directions for improving this choice in a general framework is suggesting in section VI. For the first layer motion model, we add the primitive events "Is Seating" and "Is seated". Finally the global motion "SEAT" is inserted at the higher level of the hierarchy. The "SEAT" global motion detected at frame 260 is recognized as a sequence of the "Is Walking" primitive (Frame 45), the "Is Sitting" primitive (Frame 100) and the "Is Seated" primitive (Frame 235). This result proves the model capacity to analyze a large extent of different motions, with an unique simple feature as input.

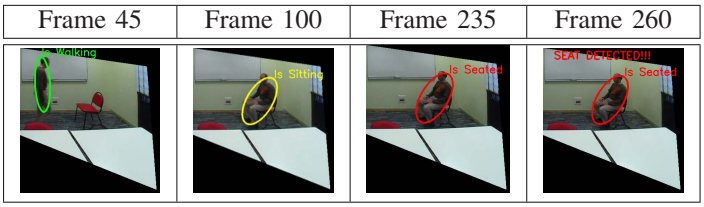


Fig. 8. Sitting motion Recognition

## VI. CONCLUSION AND FUTURE WORKS

We propose an efficient approach for activity recognition, principally dedicated to fall detection. Motion is modeled using an HHMM. The highest level corresponds to the description of a global motion. It is composed of a combination of primitive events at the previous level. These elementary motion pattern are generated by a sequence of states representing human 3D pose. Finally, the observation vector corresponds to the rectified angle between principal axis of the silhouette and the vertical direction. We claim that the model presents two major advantages compared to those traditionally used. First its hierarchical structure makes it possible to generate complex events by the concatenation of simple ones. Secondly its simplicity (unidimensional input) and interpretability (the states being directly related to world quantities) clearly facilitate its use and its computational requirement. There are many directions we can think about to improve the method. For the preprocessing steps, the rectification could be made automatic, by using Hough transform for detecting sets of parallel lines and then computing the orthogonal vanishing points. In order to improve the performance of the pose estimation (determining the first layer states from the image angle), two ways may be thought to reduce the false negative rate. The former is related to finding other features than the principal axis angle in static images so that the discriminability between standing/lying pose is improved. As stated in section IV-B.0.e, another direction were to take advantage of the relationship between image angle and posture to determine the optimal placement for additional cameras. It is a very interesting extension because it must be possible to reach 100 % recognition, making the detection more accurate and removing false negative. Moreover, the approach would not be excessively time consuming in that case because the complexity would be linear in the number of camera. Finally, using models able to manage interactions between people and objects may provide important cues for analyzing and recognizing complex activities of elderly people.

## REFERENCES


[1] S. Aoki, M. Onishi, A. Kojima, and K. Fukunaga. Learning and recognizing behavioral patterns using position and posture of human. *IEEE Conference on Cybernetics and Intelligent Systems*, 2004.
[2] H. H. Bui, S. Venkatesh, and G. West. Policy recognition in the abstract hidden markov model. *Journal of Artficial Intelligence Research*, 2002.
[3] B. Caprile and V. Torre. Using vanishing points for camera calibration. *International Journal of Computer Vision*, 1990.
[4] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. *CVPR*, 1997.
[5] D.Liebowitz. Camera calibration and reconstruction of geometry from images. *PhD thesis*, 2001.
[6] A. Galata, N. Johnson, and D. Hogg. Learning variable length markov models of behaviour. *Journal of Comp. Vision and Image Understanding*, 2001.
[7] H.Nait-Charif and S.McKenna. Activity summarisation and fall detection in a supportive home environment. *ICPR*, 2004.
[8] F. Jensen. An introduction to bayesian networks. *Springer*, 1996.
[9] D. Liebowitz and A. Zisserman. Metric rectification for perspective images of planes. *CVPR*, 1998.
[10] B. Lo, J. Wang, and G. Yang. From imaging networks to behavior profiling: Ubiquitous sensing for managed homecare of the elderly. *Pervasive*, 2005.
[11] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. *ICPR*, 2002.
[12] V. Nair and J. Clark. Automated visual surveillance using hidden markov models. *ICVI*, May 2002.
[13] N.Nguyen, , D.Phung, S.Venkatesh, and H.Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. *CVPR*, 2005.
[14] N. M. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *PAMI*, 2000.
[15] L. R. Rabiner. A tutorial on hidden markov models and selected speech recognition. *IEEE*, 1989.
[16] A. Sixsmith and N. Johnson. Simbad: Smart inactivity monitor using array-based detector. *Gerontechnology*, 2002.
[17] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. *International Workshop on Automatic Face and Gesture Recognition*, 1995.
[18] C. Stauffer and W. Grimson. Learning patterns of activity using real-ime tracking. *PAMI*, 2000.
[19] N. Thome and S. Miguet. A robust appearance model for tracking human motions. *AVSS*, 2005.
[20] B. U. Treyin, Y. Dedeoglu, and A. E. Cetin. Hmm based falling person detection using both audio and video. *ICCV*, 2005.
[21] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. *CVPR*, 1992.
[22] Y.Ivanov, , and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *PAMI*, 2000.
[23] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activities in video. *CVPR*, 2004.