# Posture Recognition Based on Fuzzy Logic for Home Monitoring of the Elderly

Damien Brulin, Yannick Benezeth, and Estelle Courtial

***Abstract*—We propose in this paper a computer vision-based posture recognition method for home monitoring of the elderly. The proposed system performs human detection prior to the posture analysis; posture recognition is performed only on a human silhouette. The human detection approach has been designed to be robust to different environmental stimuli. Thus, posture is analyzed with simple and efficient features that are not designed to manage constraints related to the environment but only designed to describe human silhouettes. The posture recognition method, based on fuzzy logic, identifies four static postures and is robust to variation in the distance between the camera and the person, and to the person's morphology. With an accuracy of 74.29% of satisfactory posture recognition, this approach can detect emergency situations such as a fall within a health smart home.**

***Index Terms*—Decision support system, fall detection, posture recognition.**

## I. Introduction

With progress in medicine and the increase in life expectancy, population aging has become a crucial issue especially in the health field. One consequence of population aging is the increase in dependence leading to impaired autonomy. During the last two decades, the concept of the "Health Smart Home" (HSH) has been extensively investigated [1]–[3]. HSH proposes solutions to tackle the difficulties faced by the elderly or the disabled in everyday life by using Information and Communication Technologies. While the aim of some projects mainly concerns comfort and leisure, most smart homes aim to improve the monitoring of the elderly. To achieve this, several devices are used: movement sensors (infrared detectors), microphones [1], wearable sensors such as accelerometers [4], [5], sensory floors [2], etc.

However, the proposed solutions can be intrusive. An accelerometer or a gyrometer can be used to monitor physical activity or to detect a fast fall [5], but these devices have to be embedded directly on the person which is often hard to bear in everyday life, especially for the elderly. The sensory floor proposed in [2] is able to locate and to differentiate people but is not adapted to existing buildings. This is a drawback since the majority of seniors (90%) still live in their own home and desire to stay at home as long as possible.

Cameras are an interesting alternative solution, but this technology is still rarely used in HSH [6]. Computer vision-based techniques can be used both to detect a person and to determine its posture. Several studies have been reported concerning the detection of human presence [7]–[9], the 3-D location of the person [10], and posture recognition [11]–[13]. Computer vision-based human detection has been widely investigated in the last few decades and can be classified into three categories: 3-D approaches, 2-D approaches with or without an explicit shape model. As the first two categories require strong assumptions on the context scene and a nontrivial mathematical model, we focus here on the third category which detects people without locating body parts. A comprehensive review of these methods can be found in [9]. The main point of these approaches is the choice of representation, which can be global or local. Once the description of human appearance has been accomplished, a supervised learning algorithm is used to train a classification function. Some approaches use background subtraction to first detect regions of interest [14], but often make assumptions about the object to be detected. We propose in this paper an approach that takes advantage of both methods by using tools dedicated to object detection in still images in a video analysis framework.

The specific problem of posture recognition has been addressed by several studies. The authors in [12] proposed a method to recognize activities of daily life (standing up, sitting down) by determining a binary volume with a recognition rate of 90%. In [15], 3-D information was used to estimate the 3-D location of the person and to learn the person's visual appearance. In [13], 3-D information was also used by comparing a detected silhouette to a 3-D articulated human model. However, these approaches require information from multiple cameras or from several points of view and the computation complexity of some algorithms is incompatible with a real-time constraint. Methods with a single camera have also been reported. Belshaw *et al*. [16] and Lee and Mihailidis [17] propose methods based on a ceiling-mounted camera. This configuration greatly simplifies the analysis by limiting partial occlusions but it is not appropriate in our case as we need to monitor over a wide-angle scene such as a bedroom. The authors in [18] propose an asynchronous temporal contrast vision sensor which extracts, after background subtraction, the centroid of moving regions. If significant vertical velocity is computed, a fall can be distinguished from normal behavior but the subtraction is done assuming a static background model, thus reducing the recognition performance in a dynamic environment (lighting changes

Manuscript received June 6, 2011; revised October 15, 2011 and April 17, 2012; accepted July 7, 2012. Date of publication July 13, 2012; date of current version September 20, 2012. This work was supported by the French Ministry of Industry and local authorities, within the framework of the CAPTHOM project of the Competitiveness cluster $S^2E^2$ (www.s2e2.fr).

D. Brulin is with the IMS Laboratory, University of Bordeaux 1, 33405 Talence Cedex, France (e-mail: damien.brulin@gmail.com).

Y. Benezeth is with the Laboratory Le2i, UMR CNRS 6306, Université de Bourgogne, 21078 Dijon Cedex, France (e-mail: yannick.benezeth@gmail.com).

E. Courtial is with the PRISME Laboratory, University of Orléans, 45072 Orléans Cedex, France (e-mail: estelle.courtial@univ-orleans.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2012.2208757

for example). In [11], an approach was developed to detect static posture based on the belief theory and distance measurements relative to a reference posture at a given distance. Consequently, if this distance changes, a new definition of the reference posture is required.

The method proposed in this paper overcomes the reliance on the person/camera distance and is fast enough for a real-time application. The proposed system uses human detection prior to posture recognition. Posture analysis is only performed on a human silhouette. Human detection has been designed to be robust to different environmental stimuli. Consequently, the global system is more reliable. Partial occlusions, lighting variations, and the introduction or removal of static objects are handled by the human detection step. In this way, we can use simple and efficient features that are not designed to manage constraints related to the environment but only designed to describe the human silhouette.

The remainder of this paper is organized as follows. In Section II, we focus on the human detection method in which each step is detailed. Section III is devoted to the structure of the posture recognition process. Based on fuzzy logic, the principle of fuzzy set theory is first introduced and then followed by a case study of posture recognition. In Section IV, experimental results are presented both in a laboratory environment and in a home environment. The robustness and the efficiency of the proposed approach are tested. For the monitoring of the elderly, fall detection is particularly studied. Finally, conclusions and perspectives are presented in Section V.

## II. Visual Human Detection

The human detection method presented in this section is divided into three different steps: change detection, tracking, and classification. Further details can be found in [19].

### A. Change Detection

The objective of the change detection step is to simplify further processing by locating areas of interest in the image. As we are working with static cameras, we use background subtraction. With a model of the environment and an observation, we attempt to detect what has changed in the current frame. For our application, areas of interest are those in the image where there is a high probability of detecting a person.

*1) Background Model:* From a comparative study of various background subtraction methods [14], we chose to model each pixel of the background by a single Gaussian distribution. For each pixel $s$ at time $t$, the background model is composed of the mean vector $\boldsymbol{\mu}_{s,t} = \{\mu_{r,s,t}, \mu_{g,s,t}, \mu_{b,s,t}\}$ and the covariance matrix $\Sigma_{s,t}$, assumed to be diagonal. $\mu_{r,s,t}$, $\mu_{g,s,t}$ and $\mu_{b,s,t}$ correspond, respectively, to the mean of the red, green, and blue components of the pixel $s$ at time $t$. The Mahalanobis distance is used to compute the difference between the current image and the background model. Detection is achieved by thresholding the Mahalanobis distance and we obtain the foreground motion mask $\mathcal{X}$, where $\boldsymbol{\mathcal{X}}_{s,t} = 1$ if the pixel $s$ belongs to the foreground (or region of interest) or $\boldsymbol{\mathcal{X}}_{s,t} = 0$ if the pixel $s$ belongs to the background, i.e., the static part of the image.

*2) Update of Background Model:* In real applications, the scene is never completely static. The model must be sufficiently flexible to automatically adapt to various changes in the environment. In order to deal with slow variations in the illumination, caused, for example, by natural change of daylight, the model is updated as follows:

$$\boldsymbol{\mu}_{s,t+1} = (1-\alpha)\,\boldsymbol{\mu}_{s,t} + \alpha\,\boldsymbol{I}_{s,t} \tag{1}$$

and diagonal terms of the covariance matrix are updated with

$$\sigma^2_{k,s,t+1} = (1-\alpha)\sigma^2_{k,s,t} + \alpha(I_{k,s,t} - \mu_{k,s,t})^2 \tag{2}$$

where $k$ stands for the red, green, and blue components. The parameter $\alpha$ is related to the speed at which new observations are taken into account. Then, for sudden variations in illumination, if the percentage of active foreground pixels (*i.e.,* $\boldsymbol{\mathcal{X}}_{s,t} = 1$) is high (i.e., 70%), the background model is reinitialized with $\boldsymbol{\mu}_{s,t} = \boldsymbol{I}_{s,t}$. The background model is also reinitialized in order to take into account the removal or the introduction of "static" objects. Thanks to further classification steps described below, we are able to determine the nature of objects and their positions in previous frames. If an object, detected by the background subtraction, is static and is regarded as not being human during a preset number of images, the background model of its corresponding shape is reinitialized using the current image.

*3) Postprocessing:* Objects detected by background subtraction ideally correspond to a compact area with smooth borders. False detections are often randomly spread over the image and correspond to small clusters of isolated pixels. We use a set of morphological operations to remove isolated pixels and to fill holes in the foreground image. Then, foreground pixels are grouped into connected components (see Fig. 1).

Fig. 1. From left to right: input image, background subtraction result, and postprocessing result (one color represents one connected component).

### B. Tracking

Background subtraction provides a list of connected components in each image. These connected components may represent humans or other moving objects. Now, we wish to know the history of the connected components' displacements in the image plane. As one connected component potentially corresponds to one object, we track each object present in the scene independently by assigning it a label constant over time. The history makes it possible to time-smooth classification errors. As constraints concerning the algorithm complexity and the amount of memory used are very important for the present application, it does not seem suitable to use a complex model of each tracked object. Consequently, our tracking method is directly based on connected components. In order to deal with usual cases (e.g., when two distinct objects form only one connected component or when one object is represented by several connected components), we characterize each tracked object by a set of points of interest. These points of interest are tracked over frames. Background subtraction provides, at each time $t$, a list of detected


Fig. 2. Illustration of a tracking result with partial occlusion obtained with a 320 × 240 video. The first row corresponds to input images with interest points associated with each object (one color per object), and the second row corresponds to the tracking result with a label that remains constant over time for each object.

connected components and we have a list of tracked objects in previous frames. We, therefore, attempt to match these two lists. The position of the points of interest, regarding connected components, makes it possible to match tracked objects with the connected components that have been detected. The tracking of points of interest is carried out with Lucas and Kanade's method [20].

An example of a tracking result with partial occlusion is illustrated in Fig. 2. In column 3 of this example, the connected component represents two objects. Then, in column 4, when these two objects separate again, we are able to recognize object 2 in column 4 as the same object 2 as in column 2.

### C. Classification

An overview of existing classification methods can be found in [21]. Among these numerous methods, Viola and Jones' method [7] was chosen for our application because of its good performance and its relatively low computation cost. Thanks to its cascade architecture, false examples are quickly rejected and Haar-like features can be computed from integral images with just a few basic operations. We briefly describe Viola and Jones' method, and then we detail the part-based classification and the confidence index.

*1) Adaboost and Haar-Like Filters:* In order to recognize humans from any other moving objects or false detection of the background subtraction algorithm, Viola and Jones' method based on 14 Haar-like filters was used. A Haar-like feature $x_i$ considers adjacent rectangular regions at a specific location, sums up the pixel intensities in these regions, and calculates the difference between them.

Each feature $x_i$ is then fed to a simple one-threshold weak classifier $f_i$:

$$f_i = \begin{cases} +1, & \text{if } x_i \geq \tau_i \\ -1, & \text{if } x_i < \tau_i \end{cases} \tag{3}$$

where $+1$ corresponds to a human shape and $-1$ to a nonhuman shape. The threshold $\tau_i$ corresponds to the optimal threshold that minimizes the misclassification error of the weak classifier $f_i$ estimated during the training stage. Then, a more robust classifier is built with several weak classifiers trained with a boosting method [22]:

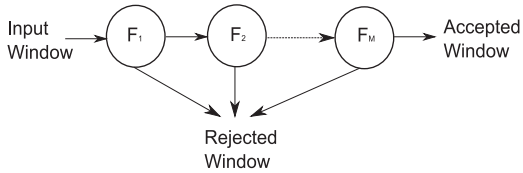


Fig. 3. Cascade of boosted classifiers.

$$F_j = \text{sign}(c_1 f_1 + c_2 f_2 + \cdots + c_n f_n). \tag{4}$$

Here, $F_j$ corresponds to the boosted classifier of the $j$th stage of the cascade where the $c_i$ coefficients correspond to weights determined during the training. A cascade of boosted classifiers is built (see Fig. 3). Each stage can reject or accept the input window. Whenever an input window passes through every stage, the algorithm labels it as a human shape.

An area of interest around the tracked object is defined. This area of interest is analyzed by the classifier with various positions and scales in a sliding window framework.

*2) Part-Based Classification:* In an indoor environment, partial occlusions are frequent. Thus, it is clearly insufficient to seek forms similar to the entire human body. The upper part of the body (head and shoulders) is often the only visible part. In practice, four classifiers are used:

1) the whole body;
2) the front/back view of the upper body;
3) the left view of the upper body;
4) the right view of the upper body.

The size of the classifier of the whole body is $12 \times 36$, composed of 27 boosted classifiers and 4457 weak classifiers. This classifier was trained on the well-known INRIA person dataset [8]. The size of classifiers of the upper body (front/back) is $20 \times 20$, composed of 23 stages and 1549 weak classifiers. Profile classifiers are $20 \times 20$ in size, made of 22 stages and 1109 weak classifiers.

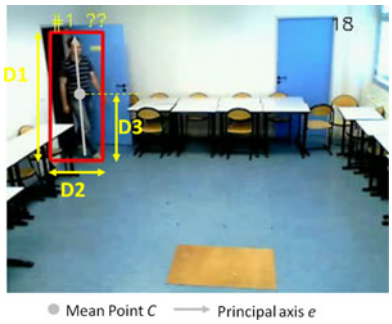


Fig. 4. Feature extracted from the visual human detection.

### D. Discussion

Management of the various thresholds and parameters used by this algorithm is very important. There are three kinds of parameters.

1) First, thresholding the Mahalanobis distance for background subtraction is very important and has a strong influence on the overall system. This parameter needs to be adjusted for each environment.
2) Then, $\alpha$, used in (1) and (2) to update the background model, may be set to a default value but the overall performance would be better if this parameter is modified based on prior knowledge of the environment. If the environment is in an office where people remain static for a while, this threshold should be low and inversely, in a corridor, this threshold may have a high value.
3) Then, other parameters can be set to a default value which will be valid for all kinds of environments.

## III. Posture Recognition

Once a human has been detected in the observed scene, posture recognition can be considered. The posture recognition process is divided into two steps: a feature extraction and a fuzzy logic system (FLS). Each of these steps is detailed below.

### A. Feature Extraction

Depending on the posture of the person, both the principal axis and the volume occupied by the human body change: a seated person is smaller than a standing one, or the principal axis is vertical in standing position while it is horizontal in lying position. The determination of these numerical parameters, used for posture recognition, constitutes the first step of our approach. Five parameters were identified, each of them computed from data extracted from the visual human detection.

The segmentation step, performed with background subtraction, delivers a binary image that highlights changes (see Fig. 1). We use the foreground mask to compute the first principal axis $e$ of the human body thanks to a principal component analysis method [23]. The main idea is to determine the best projection representing data in the least-square sense. We also extract the mean point $C$ which can be considered as the gravity center. Then, the bounding box provides information about the 2-D space occupied by the person: $D1$ height of the box (comparable to the distance head/feet), $D2$ width of the box, and $D3$ distance between $C$ and the bottom side of the box (see Fig. 4).

Finally, the relevant inputs of the FLS are defined by these four parameters:

1) $r_1 = D3/D1$;
2) $r_2 = D2/D1$;
3) $e_u$, orthogonal projection of $e$ onto the axis $u$;
4) $e_v$, orthogonal projection of $e$ onto the axis $v$.

The division between distances, $r_1$ and $r_2$, allow us to improve the robustness of the recognition with regard to the distance between the person and the camera but also to the person's morphology. Contrary to [11], the recognition system does not need to be configured for each environment or each person.

### B. FLS

The concept of fuzzy set theory was first introduced by Zadeh [24] and can be related to the work by Sugeno on fuzzy integrals [25]. Fuzzy logic merges data that can be imprecise and helps the classification of a system when the limits between classes are not really clear.

An FLS can be viewed as a mapping from inputs to outputs and can be expressed as $y = f(x)$. A FLS is composed of four main blocks, namely the fuzzifier, the rules, the inference system, and the defuzzifier. Generalities concerning the design of the different FLS blocks are addressed in the sequel and illustrated through a case study. We introduce a numerical input vector $V = (r_1, r_2, e_u, e_v)^T = (0.28, 2.8, 0.42, 0, 79)^T$, corresponding to values obtained for a person lying on the ground.

*1) Fuzzifier:* It constitutes the first step and maps numerical data into fuzzy sets. Indeed, each input or output is described by one or more fuzzy subsets defined by a membership function. Contrary to classical logic where a variable belongs wholly to a unique subset, fuzzy logic uses the concept of partial membership, i.e., a variable can belong, in different degrees, to more than one subset. The membership function of a fuzzy subset $A$ in a universe $U$ is defined as follows:

$$\forall x \in U, f_A(x) \in [0; 1]. \tag{5}$$

To define the membership functions, scenarios of everyday life were captured by a static camera. During these scenarios, all the selected postures were repeated in order to visualize, after data processing of the image sequences, the variations in the four parameters presented previously. Fig. 5 (a) and (b) illustrates the variations of the input $r_2$ during two scenarios. As can be seen in scenario 1, variations of $r_2$ are weaker or stronger depending on the posture of the person. In the second scenario, the person is lying on the ground perpendicular to the camera axis, which explains the high value of $r_2$ during this period.

The number of fuzzy subsets for each input depends on the sensitivity of each parameter with regard to the posture. With regard to the $r_2$ input, two fuzzy subsets are defined from scenario 1 corresponding to low and medium values. From scenario 2, a third subset is added for higher values. Fig. 6 shows the fuzzy subsets of the input $r_2$ and the fuzzy subsets of the input $e_u$. According to the value of the considered input vector $V$, $r_2$ is high with a degree of $0.6$ and $e_u$ is medium with a degree of $1$.

*2) Rules:* The rules used by the inference system can be obtained either by experts or extracted from numerical data. Generally, rules correspond to a set of IF–THEN propositions

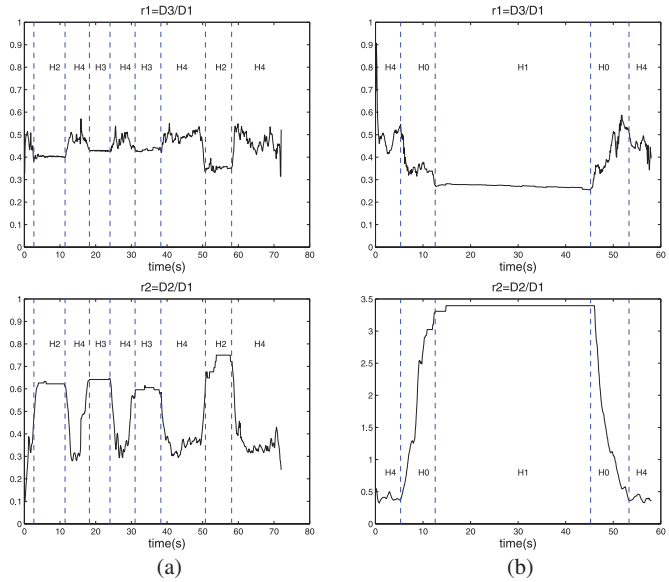


Fig. 5. Variations of $r_1$ and $r_2$ parameters during two scenarios (a) and (b) with different postures (H1: lying, H2: squatting, H3: sitting, H4: standing, and H0: undetermined).

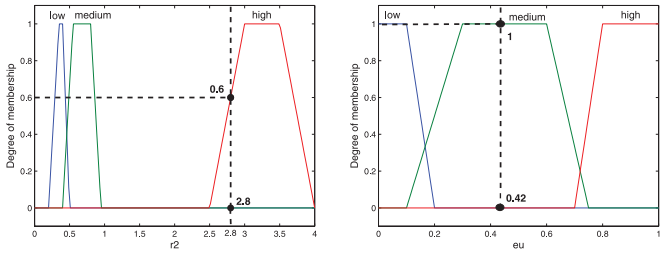


Fig. 6. Membership functions of the $r_2$ and $e_u$ inputs.

$P_i$ $(1 \leq i \leq n)$:

$$\text{IF } (x \in A_i) \text{ AND } (y \in B_i) \text{ THEN } (z \in C_i). \tag{6}$$

The logical connections (AND, OR) and implication operation (THEN) can be defined in different ways. If variables are linked by a logical function AND, the usual approach consists in considering only the variable with the lowest degree of membership. Different fuzzy implications exist to link the resulting antecedent membership function to the consequent membership function [26]. Mamdani [27] or Larsen [28] implications are usually used in an FLS. The former was chosen in our application as it offers good robustness and simplified calculations. In Mandani implication, the min operator is used to define both AND and THEN:

$$f_{S_i}(z) = \min(\min(f_{A_i}(x), f_{B_i}(y)), f_{C_i}(z)). \tag{7}$$

Scenarios and curves of variations are also used to identify combinations between two or more parameters each time a posture is realized. For our application of posture recognition, eight rules were defined as illustrated in Table I.

*3) Inference System:* If more than one rule is handled, it is necessary to define an aggregation operator $\wedge$ to synthesize the solutions $f_{S_n}$ of each rule $P_n$ and to obtain the final membership function $f_S$:

$$f_S = \wedge(f_{S_1}, f_{S_2}, \ldots, f_{S_n}). \tag{8}$$

Rules that are not concerned by the observation, i.e., when $f_{S_j} = 0$ for the rule $j$, do not have to influence the synthesis. The aggregation operator $\wedge$ has to consider 0 as a neutral element. The operator $\max$ is generally used.

According to the considered input vector $V$, only two rules are concerned:

1) $P_2$: $r_1$ is mid-low with a degree of 0.3, $e_u$ is medium with a degree of 1 and $e_v$ is high with a degree of 0.9. The membership function of posture $f_{S_2}$ is presented in Fig. 7(a) $(\min(0.3, 1, 0.9) = 0.3)$.
2) $P_6$: $r_1$ is low with a degree of 1 and $r_2$ is high with a degree of 0.6. The membership function of posture $f_{S_6}$ is presented in Fig. 7(b) $(\min(1, 0.6) = 0.6)$.

The final membership function is obtained by aggregation of $f_{S_2}$ and $f_{S_6}$ thanks to the operator $\max$ and is illustrated in Fig. 7(c).

*4) Defuzzifier:* The final step of an FLS converts the fuzzy set into numerical data. Once again several approaches can be considered to perform this step called defuzzification [29]. Among them, methods of maxima and the gravity center method can be distinguished. In the first approach, the solution is the abscissa of the maxima of the membership function solution (if several solutions exist, we can choose the mean, the largest, or the smallest solution). The second approach computes the abscissa of the gravity center of the membership function solution. For our application, the method of maxima, namely the MOM method (Mean of Max) and the gravity center method have similar computational time. Since the MOM method has slightly better results than the gravity center, we decided to choose this one for the evaluation of the process.

For the input vector $V$, the posture solution is *lying* [see Fig. 7(c)].

TABLE I
RULES OF THE INFERENCE SYSTEM

| N° | | Inputs | | | | | Output |
|---|---|---|---|---|---|---|---|
| | | $r_1$ | $r_2$ | $e_u$ | $e_v$ | | |
| $P_1$ | IF | | | high | low | THEN | Lying |
| $P_2$ | IF | mid-low | | medium | high | THEN | Lying |
| $P_3$ | IF | mid-high | | medium | high | THEN | Sitting |
| $P_4$ | IF | | medium | low | high | THEN | Squatting |
| $P_5$ | IF | | low | low | high | THEN | Standing |
| $P_6$ | IF | low | high | | | THEN | Lying |
| $P_7$ | IF | high | low | | | THEN | Standing |
| $P_8$ | IF | mid-high | medium | | | THEN | Sitting |

## IV. APPLICATION TO HOME MONITORING

To test the robustness and the performance of the proposed approach, 62 sequences of images were taken in different environments. These image sequences can be divided in three sets: the first one was used to evaluate human detection in a dynamic environment (29 sequences); the second set, recorded in a laboratory environment, was dedicated to testing and adjusting the FLS (15 sequences); and the last one was recorded in order to assess posture recognition in a home environment (18 sequences). Each recorded image sequence lasted between 2 and 10 min and was captured by a single perspective camera

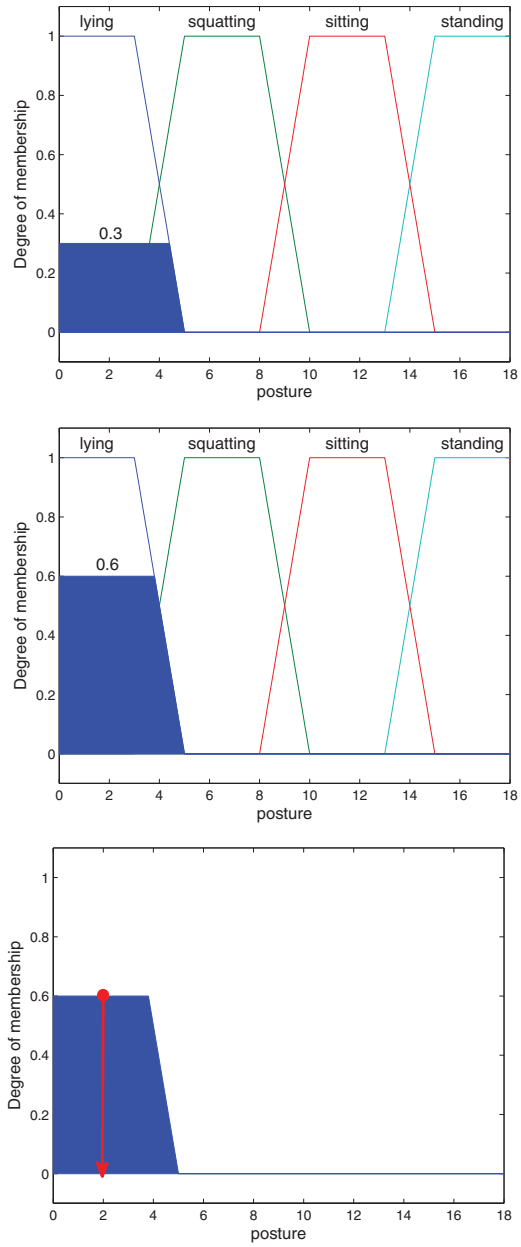


Fig. 7. Illustration of the inference system: (a) solution of the rule $P_2$, (b) solution of the rule $P_6$, and (c) membership function solution.

(Logitech Quickcam Pro 5000) placed at a height of 2 m . The 2-D signal delivered by the camera is an image of 320 × 240 pixels.

### A. Algorithm Simulation and Evaluation

*1) Evaluation of Human Detection:* The dataset used to evaluate human detection corresponds to: ($S_1$) 14 scenarios of normal activities carried out in different places (office or dining room); ($S_2$) 7 scenarios of unusual activities (agitation); and ($S_3$) 8 scenarios with false detection stimuli (lighting variations, moving objects). The union of $S_1$, $S_2$, and $S_3$ is henceforth called $S_H$. The videos were manually annotated. A complete description of these videos can be found in [19]. The proposed approach is compared to the original Viola and Jones' *et al.* [7] detection system in which the search space is reduced with background subtraction. To present the results, we use the maximum value of the *f-score* defined by

$$f\text{-}score = \frac{2.\text{Precision}\cdot\text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where *Recall* and *Precision* are determined by

$$\text{Recall} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false negatives}} \tag{10}$$

$$\text{Precision} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}}. \tag{11}$$

TABLE II
*f*-SCORE CORRESPONDING TO HUMAN DETECTION RESULTS OBTAINED ON VARIOUS DATASETS

| | Viola[7]+BS | Proposed method |
|---|---|---|
| ($S_1$) | 0.89 | **0.98** |
| ($S_2$) | 0.85 | **0.99** |
| ($S_3$) | 0.88 | **0.92** |
| ($S_H$) | 0.83 | **0.97** |

Table II presents the results obtained for each dataset. The proposed approach clearly improves the results obtained with the combination of background subtraction (called BS in Table II) and the Viola and Jones' detection method. By using tracking in order to obtain a history of displacements, we are able to time-smooth detection results since we continue to detect a person even if the classifier cannot recognize a human due for example to unusual postures. The proposed method presents a detection rate of 97% for a false detection rate of approximately 3%, underlining the advantages of combining background subtraction and tracking.

For a global evaluation of the human detection approach, the reader is referred to [19].

*2) Evaluation of Posture Recognition:* The datasets used to evaluate posture recognition correspond to training set $S_4$ with 15 scenarios (7667 images) carried out in a laboratory environment; and test set $S_5$ with 18 scenarios (7965 images) carried out in a home environment with perturbations (lighting variations, shadow, occlusions, pets, etc.). The videos were manually annotated. During the video sequences, the four postures were adopted several times by persons whose morphology is different. Our FLS approach is compared with a 1-nearest neighbor algorithm (1-NN) in which the training examples are chosen among images of dataset $S_4$. To present the results, we determine the confusion matrix in order to compute the overall and average accuracy. The overall accuracy represents the proportion of the total number of images that are correctly classified (sum of the diagonal elements divided by the total number of images of the dataset). The average accuracy is calculated as the sum of the producer's accuracies divided by the number of classes considered in our study. For each row, the producer's accuracy is the fraction of correctly classified images, i.e., the diagonal element, with regard to all images of that reference class, i.e., the row total. The columns of the confusion matrix represent the classification results and the lines the real postures adopted by the person (for example, in Table III, 60 images are classified as "sitting," whereas the real posture is "standing").

TABLE III
($S_4$) CONFUSION MATRIX WITH FLS APPROACH (TOTAL NUMBER OF IMAGES: 7687)

| FLS | Classification | | | | | | |
|---|---|---|---|---|---|---|---|
| Reference | Undetermined | Lying | Squatting | Sitting | Standing | Row Total | Producer's accuracy |
| Undetermined | 62 | 103 | 195 | 247 | 147 | 754 | 8.20% |
| Lying | 31 | 2075 | 163 | 227 | 0 | 2496 | **83.13%** |
| Squatting | 0 | 0 | 313 | 56 | 8 | 377 | **83.02%** |
| Sitting | 0 | 2 | 31 | 602 | 0 | 635 | **94.80%** |
| Standing | 46 | 15 | 153 | 60 | 3151 | 3425 | **92.00%** |
| Column Total | 139 | 2195 | 855 | 1192 | 3306 | 7687 | |

TABLE IV
($S_4$) CONFUSION MATRIX WITH 1-NN ALGORITHM (TOTAL NUMBER OF IMAGES: 7687)

| FLS | Classification | | | | | | |
|---|---|---|---|---|---|---|---|
| Reference | Undetermined | Lying | Squatting | Sitting | Standing | Row Total | Producer's accuracy |
| Undetermined | 492 | 77 | 12 | 25 | 148 | 754 | 65.25% |
| Lying | 566 | 1779 | 6 | 145 | 0 | 2496 | **71.27%** |
| Squatting | 126 | 0 | 247 | 0 | 4 | 377 | **65.52%** |
| Sitting | 168 | 211 | 0 | 248 | 8 | 635 | **39.06%** |
| Standing | 93 | 0 | 3 | 26 | 3303 | 3425 | **96.44%** |
| Column Total | 1445 | 2067 | 268 | 44 | 3463 | 7687 | |

TABLE V
($S_5$) CONFUSION MATRIX WITH FLS APPROACH (TOTAL NUMBER OF IMAGES: 7965)

| FLS | Classification | | | | | | |
|---|---|---|---|---|---|---|---|
| Reference | Undetermined | Lying | Squatting | Sitting | Standing | Row Total | Producer's accuracy |
| Undetermined | 229 | 45 | 200 | 225 | 103 | 802 | 28.55% |
| Lying | 45 | 989 | 13 | 229 | 0 | 1276 | **77.51%** |
| Squatting | 18 | 19 | 430 | 209 | 13 | 689 | **62.41%** |
| Sitting | 3 | 0 | 297 | 747 | 0 | 1047 | **71.35%** |
| Standing | 54 | 20 | 534 | 21 | 3522 | 4151 | **84.85%** |
| Column Total | 349 | 1073 | 1474 | 1431 | 3638 | 7965 | |

TABLE VI
($S_5$) CONFUSION MATRIX WITH 1-NN ALGORITHM (TOTAL NUMBER OF IMAGES: 7965)

| FLS | Classification | | | | | | |
|---|---|---|---|---|---|---|---|
| Reference | Undetermined | Lying | Squatting | Sitting | Standing | Row Total | Producer's accuracy |
| Undetermined | 448 | 133 | 4 | 34 | 183 | 802 | 55.86% |
| Lying | 243 | 1003 | 0 | 23 | 7 | 1276 | **78.61%** |
| Squatting | 368 | 25 | 248 | 6 | 42 | 689 | **35.99%** |
| Sitting | 436 | 95 | 19 | 267 | 230 | 1047 | **25.50%** |
| Standing | 331 | 1 | 25 | 47 | 3747 | 4151 | **90.27%** |
| Column Total | 1826 | 1257 | 296 | 377 | 4209 | 7965 | |

*Dataset* $S_4$ Tables III and IV show the recognition rates obtained with, respectively, the FLS approach and the 1-NN algorithm for dataset $S_4$.

The overall accuracy of our approach (80.69%) is slightly better than the overall accuracy of the 1-NN algorithm, while the average accuracy is better than the average accuracy of the 1-NN approach [Table VII]. However, some situations of conflict exist between postures, in particular between the sitting and squatting postures. This can be explained by the fact that these two postures are more complex to differentiate than the standing and lying postures. Furthermore, when the person is far from the camera, it is more difficult to differentiate them.

*Dataset* $S_5$ Tables V and VI show the recognition rates obtained for each posture for dataset $S_5$, respectively, with the FLS approach and the 1-NN algorithm.

The results show a satisfying global performance for each approach; the FLS remains the best approach with a better average accuracy [Table VIII]. We can notice that the FLS approach presents higher accuracies for the sitting and squatting postures compared with the 1-NN algorithm. There is still confusion between the sitting and squatting posture which can be explained by the fact that each time a person sits or squats, he/she does not act the same way. Besides, we can notice that there is more conflict between squatting and standing. This is particularly due to the cases when the person walks behind an obstacle like a chair. Indeed, only the trunk is detected by the human detection method, so the size of the bounding box is quite similar to the size of the box when a person is squatting.

TABLE VII
OVERALL AND AVERAGE ACCURACY ON DATASET $S_4$

| | FLS | 1-NN |
|---|---|---|
| Overall accuracy | **80.69%** | 78.95% |
| Average accuracy | **72.24%** | 67.51% |

TABLE VIII
OVERALL AND AVERAGE ACCURACY ON DATASET $S_5$

| | FLS | 1-NN |
|---|---|---|
| Overall accuracy | **74.29%** | 71.73% |
| Average accuracy | **64.93%** | 57.25% |

Errors can also be explained by limitations due to the visual system: the camera is sensitive to light variations and shadows, which can lead to poor human detection in particular concerning the size of the bounding box. As our approach depends on the dimensions of the box and the segmentation step, variations in

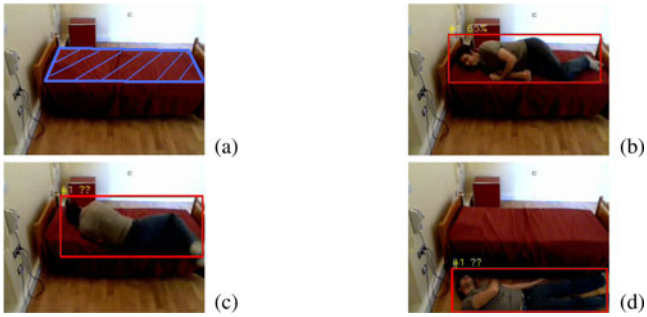


Fig. 8. Example of fall detection from a bed. (a) *A priori* knowledge of bed location, (b) normal situation, (c) fall, (d) person lying on the ground.

these two results can affect the recognition of the human posture. The shape of membership functions and inference rules could be adjusted to be more robust to these variations.

Fig. 10, at the end of this paper, illustrates some examples of postures and the results obtained by the visual human detection method and the posture recognition approach. To resume, the global performances of the visual human detection and the posture recognition system are quite satisfactory in a real environment. In the following, the approach is tested in a real environment for the monitoring of a human.

*3) Computational Costs of the Algorithm:* For the human detection step, computational time is mainly used by the classification stage (85%), since four classifiers are used to analyze the regions of interest. Background subtraction, which performs pixel-based processing, accounts for 11% of the computation time, while tracking uses only 2% (the remaining 2% concern all other low-level processing). Concerning the posture recognition, the computational time is about 1.5 ms and does not extend the time treatment of each frame. The average processing time is about 15 frames/s on a standard laptop (2 GHz and 2 GB of RAM) with 320 $\times$ 240 color images.

### *B. Monitoring of the Elderly in an HSH*

An HSH built in the "Bellevue" rest home in Bourges (France) was placed at our disposal. Since falls are the main risk for the elderly living at home, the experiments conducted targeted this issue. In order to detect the fall as soon as possible, the posture recognition system was used to determine if a person was lying down, and then additional information was used to differentiate a normal situation from an unusual situation. Additional information can be the location of the person, the position of the bed [see Fig. 8(a)], or temporal information (time, duration) for example. So depending on the values of these data, the system decides if a fall has occurred thanks to simple IF–THEN logic rules.

Knowledge of the duration of the posture makes it possible, first, to filter out the cases of incorrect posture recognition that can occur during a short period due to lighting variations, for example. Over 1 s, the posture is taken into consideration, in particular for the lying posture, the system considers that the person may have fallen and sets off the message "alertness." If the person is still lying after a preset threshold (e.g., 10 s), the message changes from "alertness" to "alert," and an emergency call is triggered by the system.

*Remark*: A sitting posture can also be an alert situation if it lasts too long in the case of a faint or a heart attack. Time, duration and daily living habits will be relevant information to discriminate such situations.


Fig. 9. Decision system for elderly monitoring: the first image represents an alertness situation just after a fall; the second image represents the same scene, 10 s later, classified as an alert situation.

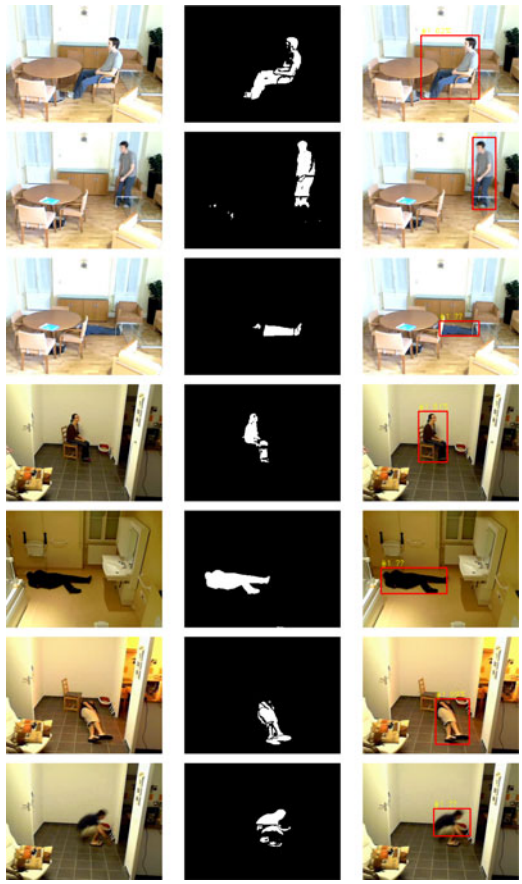

Fig. 10. Examples of well-classified and misclassified posture. From top to bottom: sitting, standing, lying, sitting, lying, squatting, and sitting.

Fig. 8 illustrates different situations observed during one scenario recorded for fall detection. Fig. 8(b) corresponds to a normal situation, whereas the fall is illustrated by Fig. 8(c) and (d).

Finally, a visualization interface collects all the information delivered by the visual human detection and the posture recognition systems (see Fig. 9). The interface is able to tell if someone

is present in the observed scene, and if the human presence is confirmed, the 2-D location and the posture of the person are displayed. Thus, from this information and the duration of the posture, the interface warns about a potential emergency situation.

## V. Conclusion

We have presented in this paper a computer vision-based approach for home monitoring of the elderly. Four static postures are identified and, therefore, emergency situations, such as a fall within an HSH, can be detected with the proposed method. The system performs human detection prior to posture analysis; posture recognition is only performed on a human silhouette. The human detection approach has been designed to be robust to different environmental stimuli. Thus, posture is analyzed with simple and efficient features that are not designed to manage constraints related to the environment but only designed to describe human silhouettes. The use of division between distances and fuzzy logic ensures the robustness of the approach with regard to the distance between the camera and the person, and to the person's morphology. The main advantages of the approach are first the low computation time required making a real-time application possible and, second, the use of a single camera without calibration. The next step is thus to test the performance of the proposed approach in real time and with longer sequences in order to see if it affects the human detection and/or the posture recognition. Furthermore, the displacements and postures of the elderly are different from those of young people, so the membership functions should certainly be adapted and the rules be refined.

## Acknowledgment

The authors would like to thank all their partners involved in the CAPTHOM project.

## References

[1] A. Fleury, M. Vacher, H. Glasson, J.-F. Serignat, and N. Noury, "Data fusion in health smart home: Preliminary individual evaluation of two families of sensors," presented at the Proc. 6th Int. Conf. Int. Soc. Gerontechnol., Pisa, Italy, 2008.

[2] W. H. Liau, C. L. Wu, and L. C. Fu, "Inhabitants tracking system in a cluttered home environment via floor load sensors," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 10–20, Jan. 2008.

[3] M. Chan, E. Campo, D. Esteve, and J.-Y. Fourniols, "Smart homes—Current features and futures perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[4] U. Anliker, J. A. Ward, P. Lukowicz, G. Troster, F. F. Dolveck, M. Baer, F. Keita, E. B. Schenker, F. Catarsi, L. Coluccini, A. Belardinelli, D. Shklarski, M. Alon, E. Hirt, R. R. Schmid, and M. M. Vuskovic, "Amon: A wearable multiparameter medical monitoring and alert system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 4, pp. 415–427, Dec. 2004.

[5] Q. Li, J. A. Stankovic, M. Hanson, A. Barth, and J. Lach, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in *Proc. Body Sensor Netw.*, 2009, pp. 138–143.

[6] A. Mihailidis, B. Carmichael, and J. Boger, "The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 3, pp. 238–247, Sep. 2004.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Comput. Vis. Pattern Recognit.*, vol. 1, pp. 511–518, 2001.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

[9] B. Schiele, M. Andriluka, N. Majer, S. Roth, and C. Wojek, "Visual people detection—Different models, comparison and discussion," presented at the Proc. IEEE Int. Conf. Robot. Autom.—Workshop People Detection Tracking, Kobe, Japan, 2009.

[10] S. Lee and Y. Kay, "An accurate estimation of 3D position and orientation of a moving object for robot stereo vision: Kalman filter approach," in *Proc. IEEE Int. Conf. Robot. Autom.*, Cincinnati, OH, May 1990, pp. 414–419.

[11] V. Girondel, A. Caplier, and L. Bonnaud, "A belief theory-based static posture recognition system for real-time video surveillance applications," in *Proc. Int. Conf. Adv. Video Signal Based Surveillance*, Como, Italy, Sep. 2005, pp. 10–15.

[12] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," *Pattern Recognit. Lett.*, vol. 29, pp. 81–89, 2008.

[13] B. Boulay and F. Brémont, M. Thonnat, "Applying 3D human model in a posture recognition system," *Pattern Recognit. Lett.*, vol. 27, no. 15, pp. 1788–1796, 2006.

[14] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly implemented background subtraction algorithms," in *Proc. 19th Int. Conf. Pattern Recognit.*, Tampa, FL, Dec. 2008, pp. 1–4.

[15] J. M. Canas, S. Marugan, M. Marron, and J. C. Garcia, "Visual fall detection for intelligent spaces," presented at the Proc. 6th IEEE Int. Symp. Int. Signal Process., Budapest, Hungary, 2009.

[16] M. Belshaw, B. Taati, D. Giesbrecht, and A. Mihailidis, "Intelligent vision-based fall detection system: Preliminary results from a real-world deployment," presented at the Proc. Rehabil. Eng. Assistive Technol. Soc. North Amer., Toronto, ON, Canada, Jun. 2011.

[17] T. Lee and A. Mihailidis, "An intelligent emergency response system: Preliminary development and testing of automated fall detection," *J. Telemed Telecare*, vol. 11, no. 4, pp. 194–198, 2005.

[18] Z. Fu, E. Culurciello, P. Lichtsteiner, and T. Delbruck, "Fall detection using an address-event temporal contrast vision sensor," in *Proc. IEEE Int. Symp. Circuits Syst.*, Seattle, WA, 2008, pp. 424–427.

[19] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, "Vision-based system for human detection and tracking in indoor environment," *Int. J. Soc. Robot.*, vol. 2, no. 1, pp. 41–52, 2010.

[20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, Vancouver, BC, Canada, 1981, pp. 674–679.

[21] M. Enzweiler and D.-M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[22] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Proc. MSRI Workshop Nonlinear Estimation, Classification*, 2002, pp. 149–172.

[23] J. Shlens. (2009, Apr.). A tutorial on principal component analysis [Online]. Available: http://www.snl.salk.edu/ shlens/pca.pdf

[24] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets Syst.*, vol. 1, pp. 3–28, 1978.

[25] M. Sugeno, "Theory of fuzzy integrals and its application" Ph.D. dissertation, Tokyo Inst. Technol., Tokyo, Japan, 1974.

[26] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*. Norwell, MA: Kluwer, 2001.

[27] E. H. Mamdani, "Applications of fuzzy set theory to control systems: A survey," in *Fuzzy Automata and Decision Processes*. Amsterdam, The Netherlands: North Holland, 1977, pp. 1–13.

[28] P. M. Larsen, "Industrial applications of fuzzy logic control," *Int. J. Man-Mach. Stud.*, vol. 12, no. 1, pp. 3–10, 1980.

[29] W. V. Leekwijck and E. Kerre, "Defuzzification: Criteria and classification," *Fuzzy Sets Syst.*, vol. 108, no. 2, pp. 159–178, 1999.

Authors' photographs and biographies not available at the time of publication.