Applying 3D Human Model in a Posture Recognition System

Bernard Boulay, François Bremond, Monique Thonnat

Applying 3D Human Model in a Posture Recognition System

Bernard Boulay *, Francois Brémond, Monique Thonnat

INRIA Sophia Antipolis, ORION group, 2004, route des Lucioles, BP93, 06902 Sophia Antipolis Cedex, France

Abstract

This paper proposes an approach to recognise human postures in video sequences, which combines a 2D approach with a 3D human model. The 3D model is a realistic articulated human model which is used to obtain reference postures to compare with test postures. Several 2D approaches using different silhouette representations are compared with each other: projections of moving pixels on the reference axis, Hu moments and skeletonisation. We are interested in a set of specific postures which are representative of typical video understanding applications. We describe results for recognition of general postures (e.g. standing) and detailed postures (e.g standing with one arm up) in ambiguous/optimal viewpoint with good/bad segmented silhouette to show the effectiveness of our approach.

Key words: Human posture, 3D human model, Vision and image processing, Silhouette, Horizontal and vertical projections, Hu moments, Image skeletonisation

1 Introduction

Human behaviour analysis is an important field for many video understanding applications such as intelligent video surveillance, aware house, augmented reality or intelligent user interfaces. The recognition of human posture is one step of the global process of analysing human behaviour. It is a difficult task

* Corresponding author. Present address: INRIA Sophia Antipolis, ORION group, 2004, route des Lucioles, BP93, 06902 Sophia Antipolis Cedex, France.Tel.: +33 (0)4 97 15 53 18; fax: +33 (0)4 92 38 79 39

Email addresses: Bernard.Boulay@sophia.inria.fr (Bernard Boulay), Francois.Bremond@sophia.inria.fr (Francois Brémond), Monique.Thonnat@sophia.inria.fr (Monique Thonnat).

because of the large variety of postures due to the high degree of freedom in the human body. Moreover, for the same posture, people can have different appearances in an image (e.g. different clothes or different camera view points). Two different types of technique can be considered to recognise human posture : intrusive and non-intrusive techniques. Intrusive techniques usually track body markers to recognise the posture of a person. Non-intrusive techniques observe a person with one or several cameras and use sophisticated vision algorithms. For video understanding applications the observed person is not always cooperative. So in this paper, we will focus in non-intrusive techniques to determine human posture. We are interested in recognising classical postures of people evolving in a scene using only one camera with a non-optimal viewpoint for video understanding applications. In the case of crime detection and prevention, the postures of the persons can provide important clues for understanding their activities.
We propose an approach combining a 3D human model with a 2D approach to recognise a set of specific postures (standing, sitting, bending and lying postures) which are useful for video understanding applications. The novelty of this approach consists in using a 3D human model for the posture recognition algorithm to be independent of the viewpoint. The 3D human model is embedded in the recognition process and used in real-time.
This paper is organised as follows: section 2 describes previous works on human posture recognition using non-intrusive techniques. Section 3 presents our human posture recognition approach. Then results and analysis are described in section 4. And finally, conclusions are presented in section 5.

2 Previous works

Previous works on human posture recognition algorithms based on non-intrusive vision techniques can be classified in three categories (Gavrilla, 1999) according to the type of human model used for posture recognition: 2D approaches with statistical models, 2D approaches with explicit models and 3D approaches.
The 2D approaches with explicit models need a 2D model and a priori knowledge on how people appear on the image. For example, Haritaoglu et al. (1998) determine first the general posture and orientation of a person by representing postures by average horizontal and vertical projections of the silhouette. Then thanks to this information they select a 2D model and recognise the different body parts. In the Pfinder project, Wren et al. (1997) determine the different parts of the body directly in the segmentation phase by using a multi-class statistical model of colour and shape to obtain a 2D representation of head and hands. Some techniques need hand initialisation of the body parts, as in Bregler and Malik (1998). The 2D models are usually stick figures wrapped

around with ribbons like in the Cardboard model (Ju et al., 1996). These approaches need to detect correctly all the body parts to achieve good posture recognition. They are generally very sensitive to segmentation errors.
To avoid these drawbacks, the 2D approaches with statistical models recognise postures without having to detect the different body parts (Boulay et al., 2003). The postures are usually described in statistical terms derived from low level features of the body silhouette. Baumberg and Hogg (1995) use salient points on the edge of the silhouette to describe the shape and achieve recognition by using an adaptive eigenshape model. Fujiyoshi et al. (2004) use skeletonisation to a silhouette representation. Panini and Cucchiara (2003) model postures with probabilistic maps by using horizontal and vertical projections of the silhouette.
The 2D approaches are not computationally expensive and they are well adapted for problems needing real-time processing. However these approaches depend on the camera viewpoint to obtain good results.
The 3D approaches search for the parameters defining the relations between the different parts of a 3D human model. Then, they try to fit the obtained 3D model to image features. A 3D human model generally consists of two components: a representation for the skeletal structure and a representation for the flesh surrounding it. The flesh can either be surface based (polygons) or volumetric (truncated cones). The volumetric model is the most frequent approach, because it requires few parameters even though it is less realistic. These approaches can work with one camera, and a priori knowledge in the form of a human model and constraints related to it. Kameda et al. (1993) use a contour based method and a surface based model. Each part of the model is taken up one by one and its rotation angles are determined based on the overlap relationship between the contour of the person silhouette and that of the projected model on the image plane. But most of the 3D approaches need several cameras (Delamarre and Faugeras (1999)), to resolve self-occlusion and posture ambiguities corresponding to situations where a person silhouette in a posture with a certain view point looks similar to silhouettes in other postures. Cohen and Li (2003) use four synchronous cameras to infer the body posture using a 3D visual-hull compared with one constructed from a set of silhouettes.
Since these 3D approaches use a 3D model they are less dependent on the viewpoint than the 2D approaches. However there are several drawbacks with them. First, these 3D approaches use a large number of parameters which are difficult to tune. A second drawback is the processing time. To recognise postures in real time the 3D approaches can only detect a few body parts and are limited to a predefined number of postures. Therefore these approaches usually try to recognise postures in optimal conditions: contrasted and isolated persons observed by a fronto-camera, or a set of cameras. However most video understanding applications require the recognition of people postures in real situations observed by only one camera under unconstrained conditions.
To our knowledge, few works address this issue. Zhao and Nevatia (2004) use

just one camera in realistic situations and a 3D human model to understand people behaviour. This is done by recognising postures of a walking or running person using an articulated dynamic human model. We propose to generalise this method by combining 2D and 3D techniques. Our approach is able to recognise 8 types of posture in any possible orientation, using only one static camera observing the scene from a non-optimal viewpoint.

3 The proposed posture recognition approach

3.1 Overview

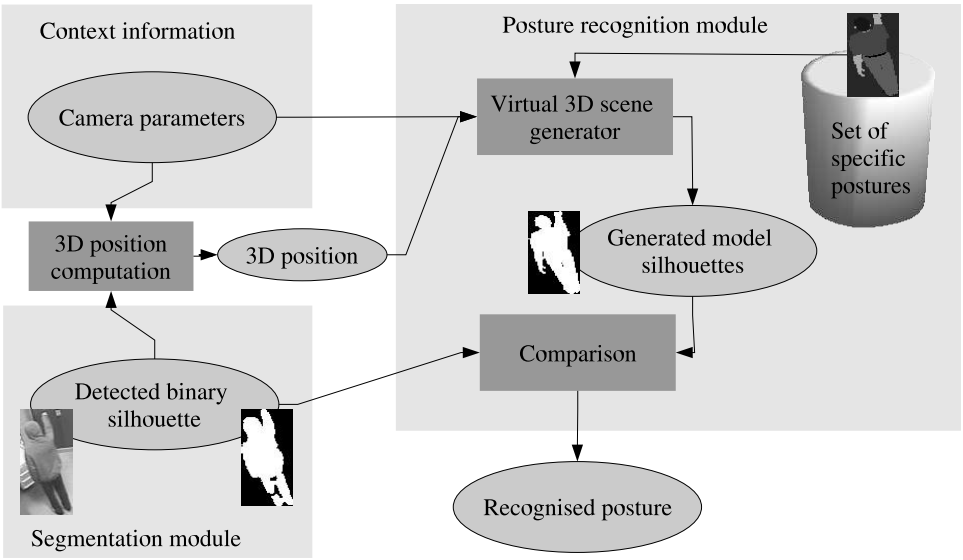

Fig. 1. Simplified scheme showing the posture recognition approach.

A simplified scheme of the approach is given in figure 1. The first step in posture recognition is to detect persons in videos. Our work is carried out in the framework of a video interpretation platform (Avanzi et al. (2005)). This platform detects moving pixels in a video with the segmentation module. The detection is done by subtracting the current image from the reference image to obtain a binary image. The reference image is periodically updated to take into account changes in the scene (light, object displacement,...). The moving pixels are then grouped into connected regions, called "blobs". A set of 3D features such as 3D position, width and height are computed for each blob. Then the blobs are classified, by using probabilistic distribution of the 3D features, into predefined classes (e.g. vehicle, person). The 3D positions of the persons are computed using the calibration matrix computed off-line.
The human posture recognition algorithm determines the posture of the detected person using the detected silhouette and its 3D position. The 3D human

model silhouettes are obtained by projecting the corresponding 3D human model on the image plane using the 3D position of the person and a virtual camera which has the same characteristics (position, orientation and field of view) as the real camera. Our dedicated 3D engine (virtual 3D scene generator) can animate and display a 3D human model. It also extracts the generated model silhouette. Finally, 3D human model silhouettes are compared with the detected silhouette to recognise the posture of the detected person. We describe the 3D human model in section 3.2. A description of the 3D engine is given in section 3.4.

3.2 The 3D human model

In our approach, we use a 3D hierarchical articulated human model for the body parts which was first proposed in SimHuman (Vosinakis and Panayiotopoulos, 2001). This 3D human model has been implemented using the Mesa library (http://www.mesa3d.org, website accessed: 9 January 2006). Mesa is a 3D graphics library with an API (Application Programming Interface) which is very similar to that of OpenGL (http://www.opengl.org, website accessed: 9 January 2006). We used Mesa because it is based on the C language and is well adapted to real time tasks. We propose to use 9 articulations as shown in figure 2 (quantity*articulation(*degree of freedom)): 1*abdomen(*3), 2*shoulders(*3), 2*elbows(*3), 2*hips(*3), and 2*knees(*1). The pelvis is not considered as an articulation, as it enables us to rotate and place the entire body in 3D space.

	Abdomen	Left shoulder	Left knee	Left elbow	Left hip	Right shoulder	Right knee	Right elbow	Right hip
α	-15/90	-45/45	0/120	-135/0	-90/30	-45/45	0/120	0/135	-90/30
β	-15/15	-135/45	0/0	0/135	-90/90	-45/135	0/0	-135/0	-90/90
γ	-30/30	-120/90	0/0	-135/0	-30/9	-90/120	0/0	0/135	-90/30

Table 1
Biomechanical constraints of our 3D human model (in degrees).

A posture is represented by a specific set of 23 parameters. These parameters are the three Euler angles for each articulation and must satisfy biomechanical constraints (see table 1). The case where all the angles values are equal to 0 corresponds to the T-shape posture. So for each posture of interest, we are able to generate a 3D model.

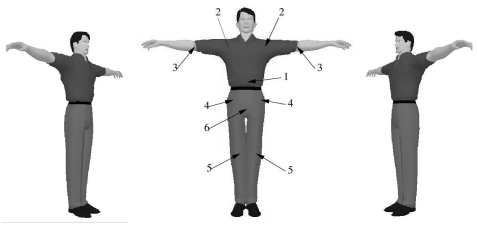

Fig. 2. 3D hierarchical human model for the T-shape posture (1: abdomen, 2: shoulders, 3: elbows, 4: hips, 5: knees, 6: pelvis).

3.3 Studied postures

We have selected a set of specific postures which are representative of typical applications in video understanding. These postures are classified in a hierarchical way. We have four general posture categories and eight detailed posture sub-categories:

- standing postures: standing with one arm up, standing with arms along the body and T-shape posture,
- sitting postures: sitting on a chair and sitting on the floor,
- bending posture,
- lying postures: lying with spread legs and lying with curled-up legs.

The parameters of our 3D human model are defined to represent each posture of interest.

3.4 Generated model silhouettes

First, the algorithm needs to generate 3D human model silhouettes for the set of specific postures. The virtual 3D scene generator engine takes as input the 23 parameters corresponding to a specific posture and places correctly the different parts of the body in 3D space. The 3D engine uses a Z-buffer to compute the projection of a 3D scene on an image plane when it is based on the Mesa library. The projection is then straightforward. The Z-buffer is an array used to store the maximum Z coordinate of any feature plotted at a given (X, Y) location on the image plane. The Z axis is perpendicular to the image plane with values increasing toward the viewer so that any point whose Z coordinate is less than the corresponding Z-buffer value will be hidden behind some features which have already been plotted. The 3D engine projects a realistic 3D model of a person in a scene (corresponding to the real scene) observed with a virtual camera. The virtual camera is defined with the same position, orientation and field of view as the real one (see figure 3). The orientation of the 3D model is computed by scanning all possible rotations (based on a ro-

tation step). A 0 degree orientation corresponds to a person facing the camera.

Fig. 3. Example of a studied image and its corresponding 3D virtual scene.

The key of this approach is how the 3D model is positioned and oriented in the virtual scene. This position and rotation axis depend on the posture type. The vertical rotation axis of standing and sitting posture is the vertical axis aligned with the head. The rotation axis of bending posture is the axis aligned with the person feet, the rotation axis of lying posture is the axis passing by the abdomen. The position of a person in standing, bending and sitting postures corresponds to the 3D coordinates of the middle point of the bottom of the bounding box of the mobile object. The position of a person in lying posture is the 3D coordinates of the moving region centre of gravity corresponding to the abdomen of the lying person.
Moreover, since the silhouettes are generated on-line, the 3D engine is only used if the detected person has a location different to the previous one. Storing off-line these silhouettes in a data base is not so efficient due to the variability of human model depending on its position, its size ... As such the processing time of the algorithm is greatly reduced. This treatment gives similar results to the ones when the silhouettes are generated at every frame. The processing time is 5 images per second compared to 2 images per second without this treatment.

3.5 Silhouette representation

Now, we have the silhouette of the detected person and the generated silhouettes of our 3D human model. The rest of the posture recognition task is to select an appropriate silhouette representation and measure to compare the detected silhouette to the generated ones.
Silhouettes can be described with statistical moments (Bobick and Davis, 2001) or horizontal and vertical projections (Boulay et al., 2005). Previous study shows that the silhouette contour also provides a good description for the binary silhouette (Fujiyoshi et al., 2004).

In the following, we describe three types of representations of a silhouette: the Hu moments, the skeletonisation and the horizontal and vertical projections.

3.5.1 The Hu moments

Shape representation by statistical moments is a classical technique in the literature (Bobick and Davis, 2001). These moments are based on 2D polynomial moments:

$$m_{pq} = \int \int x^p y^q \rho(x,y) dxdy$$

where ρ is equal to 1 for pixels belonging to the silhouette and 0 for the background. In order to make moments invariant to translations, the moments are centred :

$$\mu_{pq} = \int \int (x-\bar{x})^p (y-\bar{y})^q \rho(x,y) dxdy$$

where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$. Further the following moments are computed to make them invariant to scale.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\frac{p+q}{2}+1}}$$

Finally for these moments to be invariant to rotations the following seven Hu moments are computed:

$$\begin{aligned}
S_1 &= \eta_{20} + \eta_{02} \\
S_2 &= (\eta_{20} - \eta_{02})(\eta_{20} - \eta_{02}) + 4\eta_{11}\eta_{11} \\
S_3 &= (\eta_{30} - 3\eta_{12})(\eta_{30} - 3\eta_{12}) + (\eta_{03} - 3\eta_{21})(\eta_{03} - 3\eta_{21}) \\
S_4 &= (\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) + (\eta_{03} + \eta_{21})(\eta_{03} + \eta_{21}) \\
S_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) - 3(\eta_{03} + \eta_{21})(\eta_{03} + \eta_{21})] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{03} + \eta_{21})[3(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) - (\eta_{03} + \eta_{21})(\eta_{03} + \eta_{21})] \\
S_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) - (\eta_{03} + \eta_{21})(\eta_{03} + \eta_{21}) \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21})] \\
S_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) - 3(\eta_{21} + \eta_{03})(\eta_{21} + \eta_{03})] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{02})[3(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}) - (\eta_{21} + \eta_{03})(\eta_{21} + \eta_{03})]
\end{aligned}$$

The detected silhouette and the generated silhouettes are represented with their seven Hu moments.
To determine the posture of the detected person, Euclidean distance is used in the comparison of obtained moments. The generated silhouette which minimises the distance is chosen as the correct posture.

3.5.2 Skeletonisation

Another way to represent a silhouette is to study its contour. One way to extract salient points of the contour is by skeletonisation of the silhouette. There exist many

techniques to compute the silhouette skeleton such as thinning or distance transformation. But these techniques are computationally expensive. The method we use here is similar to Fujiyoshi et al. (2004).
The silhouette is dilated twice to remove small holes. Then an erosion is applied to smooth out any anomalies. The contour is obtained by using a border following algorithm. The centroid of the silhouette is determined based on statistical moments. The distances from the centroid to the contour points are calculated with Euclidean distance. Finally, the obtained curve is smoothed by using a linear smoothing filter before local maxima extraction. The local maxima correspond to the salient points of the contour. The skeleton is formed by connecting these maxima to the centroid. We choose a mean window algorithm to smooth the curve: the smoothed value of the curve is equal to the mean of the neighbour values. A window of 11 values is used in our tests. By varying the size of the window we can choose more or few salient points.
We propose a measure based on the distance between maxima to evaluate the similarity between two silhouettes. The skeleton points are centred on the centroid of the silhouette. Let us define SD, a set which contains the skeleton points of the detected silhouette, and SM_i, a set which contains the skeleton points of the model silhouette of the posture i. The measure between the two skeletons of the detected silhouette (SD) and the model of posture i (SM_i) is given by:

$$M_i = \sum_{pd \in SD} \min_{pm \in SM_i} (|pd, pm|) \tag{1}$$

where $|.,.|$ is the Euclidean distance. The posture that minimises this measure is chosen as the solution.

3.5.3 The horizontal and vertical projections

A usual way to represent a silhouette is its horizontal and vertical (H. & V.) projections (Haritaoglu et al., 1998),(Panini and Cucchiara, 2003), (Boulay et al., 2005). Once we have the binary silhouette of a person, we represent it by its horizontal and vertical (H. & V.) projections. The horizontal (vertical) projection on the reference axis is obtained by counting the quantity of motion pixels corresponding to the detected person for each image row (column).
We project the 3D model on an image for each reference posture which has been generated for all possible orientations. Then we compare (H. & V.) projections of these images with the (H. & V.) projections of the detected person's silhouette.
We propose a comparison based on the non-overlapping areas (equations 2, 3, 4) of the projections (figure 4).
Let us define two ratios:

$$R_o(H) = \frac{\sum_{ir \in I_o} (H^o_{ir} - H^m_{ir})^2}{\sum_{ir} (H^o_{ir})^2} \tag{2}$$

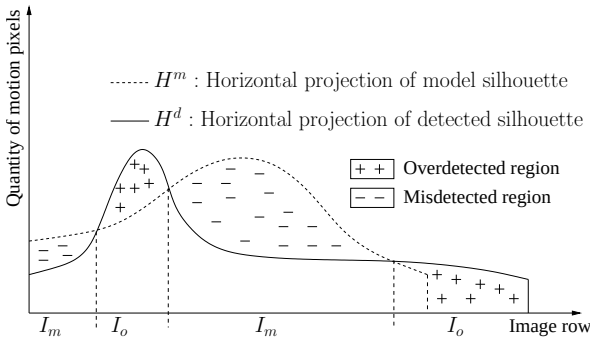

Fig. 4. "Overdetected region" corresponds to the region where the horizontal projection of the detected silhouette is superior to the horizontal projection of the model silhouette, and inversely for the "misdetected region".

which represents the sum of squared differences of the projections computed on the interval I_o, normalised by the sum of squared values of the horizontal projection of detected person (H^d), and

$$R_m(H) = \frac{\sum_{ir \in I_m} (H^o_{ir} - H^m_{ir})^2}{\sum_{ir} (H^m_{ir})^2} \tag{3}$$

which represents the sum of squared differences of the projections computed on the interval I_m, normalised by the sum of squared values of the horizontal projection of generated model (H^m).
The distance between the detected silhouette Sil_d and the model silhouette Sil_m is given by:

$$dist(Sil_m, Sil_d) = \frac{1}{4}(R_o(H) + R_m(H) + R_o(V) + R_m(V)) \tag{4}$$

This distance belongs to the range $[0, 1]$ whereby 0 corresponds to similar silhouettes. The posture model which gives the minimum distance is chosen for the posture of the studied person.

4 Results

In this section, we present how we have validated our posture recognition algorithm. We have recorded a set of video sequences taken in an office (figure 3) where a person was evolving in different postures. These video sequences have been realised with 3 different adults. The persons change the postures by turning around in order to have all possible orientations. Each frame is processed independently from the other frames by the posture recognition algorithm. The algorithm does not use information about the person posture computed in the previous frame.

4.1 Ground truth

To evaluate the approach, the ground truth is acquired for each posture with the Viper software (VIdeo Performance Evaluation Resource) (Mariano et al., 2002). This software enables us to draw the bounding box of people present in the images and to manually enter the different properties associated with each person such as "posture" and "orientation". The "posture" can be one of the 8 studied postures. The "orientation" is an approximation of the person orientation. It takes its value every 45 degrees. The ground truth is constructed as a set of 640 frames (171 standing, 118 sitting, 45 bending and 306 lying).

4.2 2D approach with learning stage

In order to show the advantage of using a 3D model rather than a classical 2D approach, we show results obtained using the approach described in (Boulay et al., 2003) with synthetic data. The approach consists in representing a posture by its (H. & V.) projections obtained through a learning stage. We adapt the approach by considering for each posture its orientation. Average (H. & V.) projections are computed for a range of orientations and a given posture. We use here an orientation range of 36 degrees. The comparison is made by using a sum of squared differences (SSD). Two different learning phases are used : one with the same point of view as the sequence test camera viewpoint, and other with a different point of view.

	3D methods			2D methods	
Approach	1	2	3	4	5
GPRR(%)	93	92	85	88	57

Table 2
The general postures recognition rates (GPRR) are given for different 2D and 3D approaches for synthetic data. 1: (H. & V.) projections, 2: Hu moments, 3: skeletonisation, 4: 2D projections with correct camera viewpoint used for the learning stage, 5: 2D projections with different camera view point used for the learning stage.

The results obtained are shown in table 2. The 2D method with a learning stage gives good recognition results for a point of view similar to the viewpoint of the test sequences (88%), whereas for a different point of view, the rate drops to 57%. The 3D method using (H. & V.) projections gives better results than the 2D methods. In the following, we will describe in more detail the results obtained with our human posture recognition approach using a 3D human model.

4.3 Rotation step

The rotation step is the main parameter to be tuned for the posture recognition algorithm. To determine the optimal value, we propose to use synthetic data. Syn-

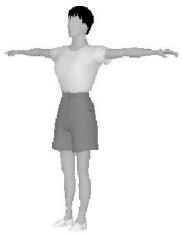

Fig. 5. 3D model used for synthetic data.

thetic data are generated with a 3D woman model (figure 5) that is different from the 3D man model used for recognition process. Synthetic data is used for both data generation and evaluation purpose (associated ground truth). The tests are realised on more than 3500 frames.

	General PRR (%)			Detailed PRR (%)			Comparison time for one silhouette (ms)			Silhouettes generation time (s)
Approach / Rotation step (degree)	1	2	3	1	2	3	1	2	3	
1	98	64	90	94	55	80	830	525	640	31
5	98	65	89	93	55	78	90	40	43	7
10	97	66	88	91	55	77	50	23	24	3
20	96	66	87	86	51	74	30	18	19	2
36	93	92	85	77	71	69	20	16	18	2
45	91	66	85	76	51	65	20	16	17	1
90	78	56	78	59	43	55	20	15	17	1

Table 3
Posture recognition rates (PRR) for different rotation step for synthetic data. 1: (H. & V.) projections, 2: Hu moments, 3: skeletonisation.

Table 3 describes the general and detailed posture recognition rate for different rotation steps with computation time information for the different silhouette representations. We remark that the (H. & V.) projections give the best recognition rates, followed by skeletonisation and Hu moments. We can also notice that 36 degrees is the optimal rotation step for the Hu moment representation. We have chosen a 36 degree step because it gives the best ratio between recognition rate and computation time combined with the (H. & V.) projections for the silhouette representation.

4.4 General posture recognition

We test our approach on real video for more than 600 frames with associated ground truth by treating each of them independently. Table 4 shows the recognition rate of general postures for the different silhouette representations. The rates obtained are equivalent to those obtained with synthetic data. The (H. & V.) projections give the best recognition. In what follows we study this representation in more depth.

	Standing	Sitting	Bending	Lying
(H. & V.) projections	96	86	87	92
Hu moments	68	73	27	35
Skeletonisation	93	68	88	65

Table 4
General postures recognition rates for the different approaches for real data.

Recognition \ Ground Truth	Standing	Sitting	Bending	Lying
Standing	**165**	5	2	0
Sitting	6	**102**	1	24
Bending	0	11	**39**	1
Lying	0	0	3	**281**
Detected/total	165/171	102/118	39/45	281/306
Success percentage	96	86	87	92

Table 5
Confusion matrix for general postures recognition for (H. & V.) projections.

Recognition \ Ground Truth	1	2	3	4	5	6	7	8
Standing (1)	**46**	6	0	5	0	0	0	0
Standing with one arm up (2)	9	**49**	11	0	0	2	0	0
T-shape (3)	2	11	**31**	0	0	0	0	0
Sitting on a chair (4)	3	0	3	**25**	8	0	0	7
Sitting on the floor (5)	0	0	0	4	**65**	1	0	17
Bending (6)	0	0	0	2	9	**39**	1	0
Lying with spread legs (7)	0	0	0	0	0	3	**102**	47
Lying with curled up legs (8)	0	0	0	0	0	0	23	**109**
Detected/total	46/60	49/66	31/45	25/36	65/82	39/45	102/126	109/180
Success percentage	76	74	69	69	79	87	81	60

Table 6
Confusion matrix for detailed postures recognition with (H. & V.) projections approach.

Table 5 shows the confusion matrix for the recognition of the 4 general postures. The obtained results are satisfactory (the rate of correct recognition is above 85%) and show the robustness of recognition of general postures in all possible orientations.
Sitting and bending are ambiguous postures because of the intrinsic ambiguity of these two postures under certain points of view. As we treat each frame independently we believe that using temporal information will partially solve this problem.

4.5 Detailed posture recognition

We give the confusion matrix for recognition of detailed postures in table 6. The quality of these results is also due to (and dependent on) the quality of the segmentation. In section 4.7 we study in more detail the influence of the errors of segmentation on the quality of posture recognition. The postures are often mixed with another posture of the same category (e.g. sitting on the floor and sitting on a chair). By considering cases of wrong recognition, we can see that the second choice is the correct posture in 75% of the cases. They correspond to ambiguous cases. Other cases of wrong recognition are due to the problem of segmentation. Finally, few errors are due to the fact that the 3D models represent a specific posture, and do not take into account the variability of the postures (e.g. for the standing posture with one arm up, the arm can be more or less up). We can notice that the proposed posture recognition approach can deal with a posture that is not of interest by recognising the most similar posture. In many cases these results are sufficient for behaviour because temporal coherency can resolve ambiguities when the posture becomes observable.

4.6 Ambiguous cases

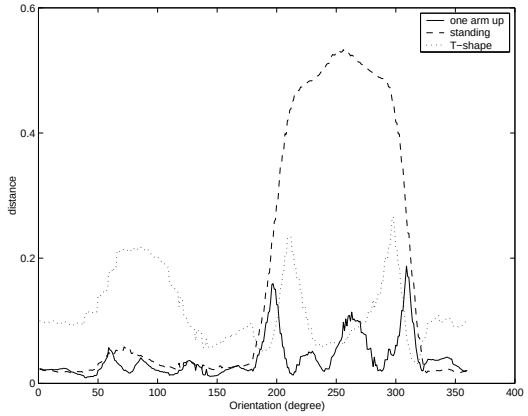

Fig. 6. The graphic shows distances obtained by comparing standing posture with one arm up (3D woman model) with all the standing postures (3D man model). (H. & V.) projections approach is used for different orientations in degree.

Silhouettes representative of different postures can have the same projection on the image for a certain point of view: defining an ambiguous case. These ambiguities are due to ambiguous viewpoints and person self-occlusion. These cases depend on the silhouette representation and the comparison measure. The percentage of ambiguity of a posture is determined by using synthetic data.
Figure 6 illustrates the ambiguity problem for standing with one arm up posture for (H. & V.) projections. This posture is similar to standing posture for many orientations. The graph can provide a confidence value for the recognition in function of the recognised posture and orientation. For example, during the interval [200,250],

we can recognise the standing with one arm up posture without ambiguity.

4.7 Segmentation errors

The segmentation of the person in a video is often not perfect (hole, missing body parts). The problem of little holes is solved partially with the silhouette representation based on (H. & V.) projections. Missing body parts are often the feet, the head or the hands (the extremities of the body).

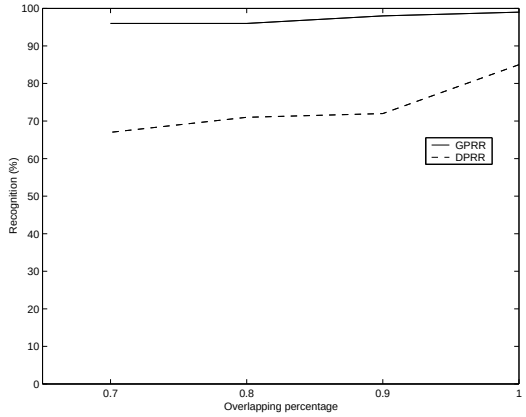

Fig. 7. General (GPRR) and detailed (DPRR) recognition rate for standing postures with different overlapping percentage.

To evaluate the incidence of missing body parts we have used a segmentation criterion based on bounding boxes. We compare the bounding box obtained by the platform with that of the ground truth. By comparing overlapping levels, we can take into account the segmentation quality during the evaluation phase. The levels correspond to the overlapping percentage of the bounding boxes (bbs) taking its value in the range [0,1] (0: the bbs are disconnected, 1: the bbs match perfectly). Figure 7 gives the general and detailed posture recognition rate for standing postures with different levels. Both recognitions are correct for all tested situations (with at least 70% of overlapping). The algorithm is able to recognise detailed postures even if the segmentation is not perfect.

5 Conclusion

We have presented an approach to recognise human posture combining 2D and 3D techniques. We believe that the use of a 3D human model is a key contribution for improving the results and making the approach independent of the camera viewpoint as shown in section 4.2.
We have compared different techniques and shown that, while using only one camera, the approach combining (H. & V.) projections with a 3D human model gives satisfying results. We have shown that the approach is effective in discriminating 4

general postures -standing, sitting, bending and lying- from any viewpoint. This approach can also recognise detailed postures, except for the cases where postures are visually ambiguous. The algorithm is able to deal with bad segmented silhouettes where body parts are missing. The algorithm is relatively fast (5 frames per second). This frame rate is sufficient since we only need to recognise posture on a few key frames to help further behaviour analysis. Moreover, even if only one instance of each posture of interest is taken into account (e.g. for the standing with one arm up posture, the arm can be more or less up) the approach recognises correctly the most similar posture among the set of possible postures.
For future work, we plan to improve our approach in two directions. On the one hand, the 3D human model must adapt itself automatically to the studied person (size, dressing, hair style, etc.). On the other hand, we need to include temporal information to maintain consistency and to reduce recognition errors. First, it will partially solve the ambiguity problem. Second, key postures will be detected in the video and the computation time will decrease. Finally, posture tracking will help human behaviour understanding.

Acknowledgement

This research is leaded in cooperation with STMicroelectronics Rousset under PS26/27 Smart Environment project financed by Conseil Régional Provence-Alpes-Côte d'Azur.

References

Avanzi, A., Bremond, F., Tornieri, C., Thonnat, M., 2005, Design and Assessment of an Intelligent Activity Monitoring Platform. EURASIP JASP IVS, pp. 2359-2374.

Baumberg, A., Hogg, D., 1995. An Adaptive Eigenshape Model. British Machine Vision Conference, vol. 1, pp. 87-96.

Bobick, A.F., Davis, J.W, 2001. The Recognition of Human Movement Using Temporal Templates. IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp.257-267.

Boulay, B., Bremond, F., Thonnat, M., 2003. Human Posture Recognition in Video Sequence. VS-PETS 2003, pp. 23-29.

Boulay, B., Bremond, F., Thonnat, M., 2005. Posture Recognition with a 3D Human Model. ICDP 2005, pp. 135-138.

Bregler, C., Malik, J., 1998, Tracking People with Twists and Exponential Maps. CVPR'98, pp. 8-15.

Cohen, I., Li, H., 2003. Inference of Human Postures by Classification of 3D Human Body Shape. IEEE International Workshop on Analysis and Modeling of Faces and Gestures, pp. 74-81.

Delamarre, Q., Faugeras, O., 1999. 3D Articulated Models and MultiView Tracking with Silhouettes. ICCV, pp. 716-721.

Fujiyoshi, H., Lipton, A.J., Kanade, T., 2004. Real-Time Human Motion Analysis by Image Skeletonization. IEICE Trans. Inf. & Syst., Vol. E87-D, pp. 113-120.

Gavrilla, D.M., 1999. The Visual Analysis of Human Movement: A Survey. CVIU, 73, pp. 82-98.

Haritaoglu, I., Harwood, D., Davis, L.S., 1998. Ghost: a Human Body Part Labelling System Using Silhouettes. ICPR, pp. 77-82.

Ju, S.X., Black, M.J., Yacoob, Y., 1996. Cardboard Model: a Parametrized Model of an Articulated Object from its Silhouette Image. 2nd International Conference on Automatic Face and Gesture Recognition, pp. 38-44.

Kameda, Y., Minoh, M., Ikeda, K., 1993. Three Dimensional Pose Estimation of an Articulated Object from its Silhouette Image. ACCV93, pp. 612-615.

Mariano, V.Y., Min, J., Park, J.H., Kasturi, Mihalcik, D., Doermann, D., Drayer, T., 2002. Performance Evaluation of Object Detection Algorithms. ICPR, pp.965-969.

Panini, L., Cucchiara, R., 2003. A Machine Learning Approach for Human Posture Detection in Domotics Applications. ICIAP, pp. 103-108.

Park, J.S., Oh, H.S., Chang, D.H., Lee, E.T., 2000. Human Posture Recognition Using curve Segments for Image Retrieval. Storage and Retrieval for Media Databases, pp. 2-11.

Vosinakis, S., Panayiotopoulos, T., 2001. Simhuman: a Platform for Real-Time Virtual Agents with Planning Capabilities. IVA'01, pp. 210-223.

Wren, C., Azarbayejani, A., Darrell, T., Pentland, A., 1997. Pfinder: Real Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 780-785.

Zhao, T., Nevatia, R., 2004. Tracking Multiple Humans in Complex Situations. PAMI, pp. 1208-1221.

Mesa, http://www.mesa3d.org/, Mesa is a 3D graphics library, website accessed: 9 January 2006.

OpenGl, http://www.opengl.org/, OpenGL is a trademark of Silicon Graphics Incorporated, website accessed: 9 January 2006.