Posture Recognition with a 3D Human Model

Bernard Boulay, François Brémond, Monique Thonnat

Posture Recognition with a 3D Human Model

B. Boulay
INRIA - France
Bernard.Boulay@sophia.inria.fr

F. Bremond
INRIA - France
Francois.Bremond@sophia.inria.fr

M. Thonnat
INRIA - France
Monique.Thonnat@sophia.inria.fr

ABSTRACT

This paper proposes an approach to recognise human postures in video sequences, which combines a 2D approach with a 3D human model. The 2D approach consists in projections of moving pixels on the reference axis. The 3D model is a realistic articulated human model which is used to obtain reference postures to compare with test postures. We are interested in a set of specific postures which are representative of typical applications in video interpretation. We give results for recognition of general (e.g. standing) and detailed (e.g standing with one arm up) postures. First results show the effectiveness of our approach for recognition of human posture.

INTRODUCTION

We are interested in recognising classical postures of people evolving in a scene using only one camera with a non-optimal view point. Human behaviour analysis is an important field for many applications such as 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. The recognition of posture is an ambitious and difficult goal because of the large variety of postures due to the high degree of freedom in the human body. Moreover, for a same posture people can have different appearances on an image (e.g. different clothes or different camera view points).
The next section summarises briefly previous work on human posture recognition. Then we present our approach and the results we have obtained. Finally we conclude on the accuracy of the approach.

PREVIOUS WORK

In this section, we are presenting techniques from the literature on human posture recognition algorithms based on non-intrusive vision techniques, rather than techniques using body markers. Previous work can be classify by considerating the type of human model (statistical model, 2D model, 3D model) used for posture recognition. The existing approaches can be classified in three categories [5]: 2D approaches with statistical models, 2D approaches with explicit models and 3D approaches.
The 2D approaches with statistical models enable to recognise postures without having to detect the different body parts [3]. The postures are usually described in statistical terms derived from low level features. Baumberg et al. [1] use salient points on the edge of the silhouette. Panini and Cucchiara [8] model postures with probabilistic maps by using horizontal and vertical projections.
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. [6] first determine the posture and orientation of the person. Second thanks to this information they select a 2D model and recognise the different body parts. The models are usually stick figures wrapped around with ribbons.
The 2D approaches are not resource demanding and they are well adapted for real time problems. However these approaches depend on the camera view point 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 compare and try to fit the obtained 3D model with image features. A 3D human model consists generally 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 less parameters even if it is the 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. But most of them need several cameras (Delamarre and Faugeras [4]), 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.
Since these 3D approaches use a 3D model they are more independent of the view point than the 2D approaches. However there are several drawbacks. 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 just detect few body parts and are limited to a predefined number of postures. Therefore these approaches usually try to recognise postures in optimal conditions: contrasted people observed by a fronto-camera. However most video understanding applications require the recognition of people postures in real situations observed by only one camera.

Few work address this issue. In [10], Zhao et al. use just one camera in realistic situations and a 3D human model to understand people behaviour. This computation is done by recognising postures of a walking and running person using an articulated dynamic human model.
We propose to generalise this method by proposing a new approach combining 2D and 3D techniques to recognise 8 types of postures in any possible orientation and using only one static camera observing the scene from a non-optimal view point.

APPROACH

Our work has been carried out in the framework of a video interpretation platform (Avanzi et al. [2]). This platform is first able to detect moving pixels in a video. The detection is made by subtracting the current image with the reference image to obtain a binary image. The moving pixels are grouped in connected regions, called blobs. These blobs are classified into predefined classes (e.g. vehicle, person) and are tracked all along the video. The final step consists in recognising the behaviours related to the tracked mobile objects. The posture recognition algorithm will help this last step by providing accurate information on people.

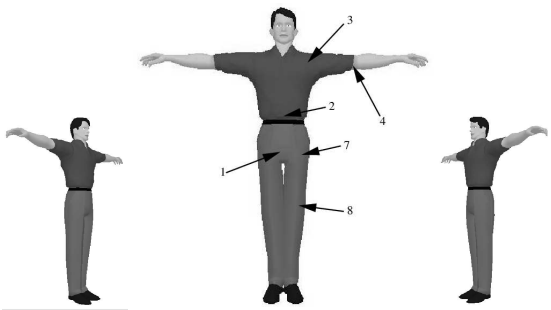

Figure 1: 3D hierarchical human model for the T-shape posture

Figure 2: 3D model of sitting on the floor posture

We have selected a set of specific postures which are representative of typical applications in video interpretation. These postures are classified in four general posture 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).

In our approach, we use a 3D hierarchical articulated human model which body parts were first defined in SimHuman (Vosinakis and Panayiotopoulos [9]). We propose to use 9 articulations as shown in figure 1: the abdomen(1), shoulders(2), elbows(2), hips(2), and knees(2). The pelvis is not considered like an articulation, it enables us to rotate all the body.
We have first defined for each posture of interest a specific set of 27 parameters. These parameters are the three Euler angles for each articulations. We can see, for example, the 3D model shown in figure 2 of a person sitting on the floor for a set of parameters. So for each posture of interest, we are able to generate a 3D model.
A simplified scheme of the approach is given in figure 3. For each detected person, the video interpretation platform first computes the 3D position using different information about the scene context (e.g. the camera parameters).

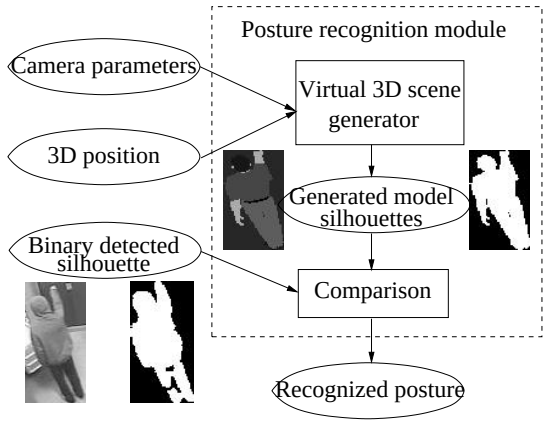

Figure 3: Simplified scheme showing the approach

Then the 3D human model silhouettes are obtained by projecting the corresponding 3D human model on the image plan using the 3D position of the person and a virtual camera which has the same characteristics (position, orientation and field of view) than the real camera.
The orientation of the 3D model is computed by scanning all possible rotations (based on a rotation step). A 0 degree orientation corresponds to a person facing the camera.
Important points of the approach is how the 3D models are positioned and how it is turned over in the virtual scene. This position and this 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 and the rotation axis of lying posture is the axis passing by the abdomen. The position of a person in standing, bending and sitting posture 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 center of gravity corresponding to the abdomen of the lying person.
We project the 3D model on an image for each reference posture which has been generated for all possible orientation. Then we compare horizontal and vertical (H. & V.) projections of theses images with the (H. & V.) projections of the detected person silhouette. The horizontal (resp. vertical) projection on the reference axis is obtained by counting the quantity of motion pixels corresponding to the detected person for each image row (resp. column).

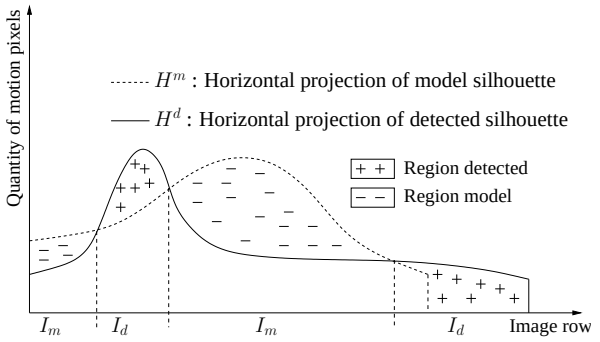

Figure 4: "Region detected" corresponds to the region where the horizontal projection of the detected silhouette is superior to the horizontal projection of the model silhouette, and inversely for the "region model".

The comparison is made by studying the non-overlapping areas (equations 1, 2, 3) of the projections (figure 4).
Let's define two ratios:

$$R_d(H) = \frac{\sum_{ir \in I_d} \left(H^d_{ir} - H^m_{ir}\right)^2}{\sum_{ir} \left(H^d_{ir}\right)^2} \quad (1)$$

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

$$R_m(H) = \frac{\sum_{ir \in I_m} \left(H^d_{ir} - H^m_{ir}\right)^2}{\sum_{ir} \left(H^m_{ir}\right)^2} \quad (2)$$

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:

$$\begin{aligned} dist(Sil_m, Sil_d) &= R_d(H) + R_m(H) \\ &\quad + R_d(V) + R_m(V) \end{aligned} \quad (3)$$

The posture model which gives the minimum distance is chosen for the posture of the studied person.

RESULTS

In this section we present how we have validated our posture recognition algorithm. We have recorded a set of videos taken in an office (figure 5) where different persons were evolving in different postures. The persons act the postures by turning on themselves in order to have all possible orientations. Each frame is processed independently by the posture recognition algorithm.

Figure 5: Example of studied image

To evaluate the approach the ground truth is acquired for each posture with the Viper software (VIdeo Performance Evaluation Resource) [7]. This software enables us to draw graphically the bounding box of people present in the images and to manually enter the different properties associated to 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 rotation step is the main parameter to be tuned for the posture recognition algorithm. We have chosen a 36 degree step because it gives the better ratio between recognition rate and computation time (about 2 images by second).
Table 1 shows the confusion matrix for the recognition

Ground Truth / Recognition	Standing	Sitting	Bending	Lying
Standing	**171**	5	0	0
Sitting	1	**102**	1	6
Bending	5	12	**42**	0
Lying	0	0	1	**289**
Detected/total	171/177	102/119	42/44	289/295
Success percentage	96	86	95	98

Table 1: Confusion matrix for general postures recognition

Ground Truth / Recognition	Standing	Standing with one arm up	T-shape	Sitting on a chair	Sitting on the fbor	Bending	Lying with spread legs	Lying with curled up legs
Standing	**44**	9	3	4	0	0	0	0
Standing with one arm up	15	**52**	22	1	0	0	0	0
T-shape	1	8	**12**	0	0	0	0	0
Sitting on a chair	0	0	1	**21**	14	0	0	2
Sitting on the floor	0	0	0	10	**57**	1	0	4
Bending	0	0	5	0	12	**42**	0	0
Lying with spread legs	0	0	0	0	0	0	**103**	59
Lying with curled up legs	0	0	0	0	0	1	20	**107**
Detected/total	44/60	52/74	12/43	21/36	57/83	42/44	103/123	107/172
Success percentage	73	70	28	58	69	95	84	62

Table 2: Confusion matrix for detailed postures recognition

of the 4 general postures. The obtained results are good (the good recognition is above 95%) and show the robustness of the recognition of general postures in all possible orientations. The two ambiguous postures are sitting and bending. This is due to the ambiguity problem : these two postures are visually ambiguous under certain points of view.

We give the confusion matrix for recognition of detailed postures in table 2. The postures are often mixed with another posture of the same category (e.g. sitting on the floor and sitting on a chair). A quick analysis of the results shows that 25% of the wrong recognition are due to problems of segmentation or 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). For the other 75 % the correct posture is the second choice. These cases correspond to the ambiguous cases. Thus the recognition algorithm is able to correctly recognise visually non-ambiguous postures. In many cases these results are sufficient to analyse behaviour because temporal coherency can resolve ambiguities when the posture becomes observable.

CONCLUSION

We have presented an approach to recognise human posture combining 2D and 3D techniques. The use of a 3D human model improves results by making the approach independent of the camera view point.

We have shown that the approach is efficient to discriminate the 4 general postures : standing, sitting, bending and lying from any view point. Except the cases where postures are visually ambiguous, this approach manages also to recognise detailed postures. For recognition in ambiguous situations, temporal coherency of the postures can be used with the help of person tracking information.

The algorithm is relatively fast (2 frames by second). This frame rate is sufficient since we only need to recognise posture on few key frames to help with the behaviour analysis.

First, we plan to determine a confidence value in recognised postures by using information on position, orientation and 3D human model. Second, we want to adapt the 3D model to the studied person in video (e.g. corpulence, clothes) to achieve a better detailed posture recognition.

References

[1] A. Baumberg, D. Hogg, 1995, "An Adaptive Eigenshape Model," British Machine Vision Conference.

[2] A. Avanzi, F. Bremond, C. Tornieri, M. Thonnat, [to be published], "Design and Assessment of an Intelligent Activity Monitoring Platform", EURASIP JASP IVS.

[3] B. Boulay, F. Bremond, M. Thonnat, 2003, "Human Posture Recognition in Video Sequence," VS-PETS.

[4] Q. Delamarre, O. Faugeras, 1999, "3D Articulated Models and Multi View Tracking with Silhouettes," ICCV, pp. 716-721.

[5] D. M. Gavrilla, 1999, "The Visual Analysis of Human Movement: A Survey," Computer Vision and Image Understanding, 73, pp. 82-98.

[6] I. Haritaoglu, D. Harwood, L. S. Davis, 1998, "Ghost: a Human Body Part Labelling System Using Silhouettes," ICPR.

[7] Language and Media Processing Laboratory, "Viper: Video Performance Evaluation Resource," *http://lamp.cfar.umd.edu/media/research/viper/*.

[8] L. Panini, R. Cucchiara, 2003, "A Machine Learning Approach for Human Posture Detection in Domotics Applications," ICIAP.

[9] S. Vosinakis, T. Panayiotopoulos, 2001, "Simhuman: a Platform for Real-Time Virtual Agents with Planning Capabilities," IVA'01.

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