SCALABILITY IN HUMAN SHAPE ANALYSIS

Thomas Fourès and Philippe Joly

IRIT - Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 Toulouse Cedex 4, France
{Thomas.Foures,Philippe.Joly}@irit.fr

ABSTRACT

This paper proposes a new approach for the human motion analysis. The main contribution comes from the proposed representation of the human body. Most of already existing systems are based on a model. When this one is *a priori* known, it may not evolve automatically according to user needs, or to the detail level that is actually possible to extract, or to restrictions due to the processing time. In order to propose a more flexible system, a hierarchical representation of the human body is implemented. It aims at providing a multi-resolution description and results at different levels of accuracy. An explanation about the model construction and the method used to map it onto features extracted from an image sequence are presented. Relations between the different body limbs and some physical constraints are then integrated. The transition from a model level to the next one is also explained and results on frames coming from a video sequence give an illustration of the proposed strategy.

1. INTRODUCTION

Human motion analysis is a problem which has been addressed in different ways according to various expected goals. Methods using low ([1]) or high level features ([2]) such as optical flow have been proposed in the past. Those ones are most of the time dedicated to one specific task such as the recognition of a specific motion and then difficult to use in another context. The need to define a model of the human body appeared in order to provide a sharper and more flexible description. Our objectives are the creation of a system as generic and adaptable as possible. No knowledge about the performed motion or the subject posture is required. Only one view is employed. The assumption made is that the camera is static.
In previous works, the study of human motion is covered by a wide range of approaches. Works in 2D as well as in 3D can be found, with or without intrusive methods. Those ones are based on specific devices worn by the subject ([3]) and high level procedures such as motion recognition in different spaces ([4]) generally follows. Obviously, an intrusive approach can not be generalized outside a controlled environment, and so techniques coming from computer vision can be required. A distinction can be made between the requirements or not of a human model. Methods without any model rely generally on assumptions to ensure the features correspondence ([5], [1]). But they show difficulties to provide a sharp description of the subject posture. For this reason, human models are widely used. Their goal is to help the segmentation, tracking and pose recognition by introducing pieces of information about the human body. These models are based on different features: sticks connected by joints, edges ([6]), ribbons ([7]), silhouette ([8]), blobs ([9]), in 2D ([8], [7]) or in 3D space ([6], [9]). In this last case, models are a volumetric extension of the first ones. The main differences come from the way the features correspondence between the model and the ones extracted from the image is performed. But the common concept is to minimize a criterion between extracted features and the model. The prediction of the feature location in the next frame is frequently used (by Kalman or velocity constraints). In [10], regions of interest are estimated rather than prediction of feature locations. Methods employed are chosen according to features used and, in general, the complexity of the task increases with the complexity of the model. Thus, all of those models have to respect a trade-off between their parameters and the real capacities of extraction.

2. A HIERARCHICAL MODELING

2.1. Principles of a Refined Description

The principle is to have a multi-level description of the human body, defined in a hierarchical way. The first level is quite rough and down to the last in the hierarchy, the model becomes sharper. A comparison between this concept and a multi-resolution approach ([11]) can be made. In a multi-resolution context, only one model is used with different resolutions of the matching space. The idea is to refine the localization of the desired characteristic from one level to the next. In our approach, the resolution is a constant, the variable one being the proposed level of representation. The underlying concept is to provide first a quite general description of the subject pose. According to application's goals, this precision degree may or may not be sufficient. In the second case, the match can be refined according to the first one by using the

next level of the model. This one provides a sharper representation with more details. The same process is repeated until expected level of precision has been reached, in respect of processing time constraints or the impossibility to go further in the description.

2.2. Human Body Model

For practical implementation, we have used for now a description composed of three levels. The first one detects the head and the torso with the same element; each leg and arm is also associated to another segment. The next level refines this first decomposition by localizing more accurately the head, and the different parts of arms and legs. At last, motion of the hands and feet are localized (a graphical representation is provided by the result illustrations on Fig.3). This model has been cut according with human body segments. It is totally independent of the system itself and adaptable to other applications needs or object shapes. We have chosen rectangles to describe the human limbs because of their simplicity. Indeed, two parameters are required to characterize their size (length, height); and two other ones to describe their localization (one vertex and an orientation angle).

3. MODEL - SUBJECT MATCHING

This section provides a description of the different parts of the system used: the subject segmentation process, the features correspondence, the incorporation of the human body physical constraints, and the use of temporal information in the case of video content analysis.

3.1. Preprocessing

The first knowledge about subject posture is obtained trough its bounding box. In our model definition, the bounding box corresponds to the level 0. As in many systems, a simple background subtraction allows to retain only subject pixels. Background can be modeled by different approaches ([12]). Opening and closing operations allow noise removal and the obtaining of connected pixels which compose the subject shape. This characteristic is the abstraction level used in the next section.

3.2. Feature Correspondance

3.2.1. First Iteration

The matching of the proposed model is made at the region scale. The similarity measure is derived from the chamfer distance. The image is cut into search areas, one for each model element. Those ones are matched with pixels located within their own area. During the first iteration, as no *a priori* information is known about the subject posture, the output of the level 0 is cut according to the most probable location of different subject limbs in the image. The center of the bounding box corresponds to the search area of the head - torso set (for the first model level), the upper right and left sides correspond to the arms, and the lower right and left parts to the legs. This cutting can be considered as an initialization step and is used only with the first frames. Each model element is then matched within its own area independently from the others. A description of the matching algorithm used is given in [13]: it is a dichotomic search based on element symmetry properties, reducing the computational cost. The best position is selected by using the root mean square value m of pixels located under model elements ([11]).

3.2.2. Temporal and Physical Relations

Temporal relation between frames, when the matching is performed on a video, is used through the search area redefinition. A first matching on an element is made at time t within its search area which is an angular sector (Fig.1). From this result, a new search area is defined at time $t+1$ with two parameters: its origin $O(x, y)$ and its angle α which depends on the previous matching quality ([13]). A quite satisfying previous match implies a reduction of the search space. On the contrary, poor ones lead to increase α.

$$\alpha_{t+1} = f(m_{eval}) * \alpha_t \quad (1)$$

where f is a polynomial function applied on the value of the matching quality measure m_{eval}.

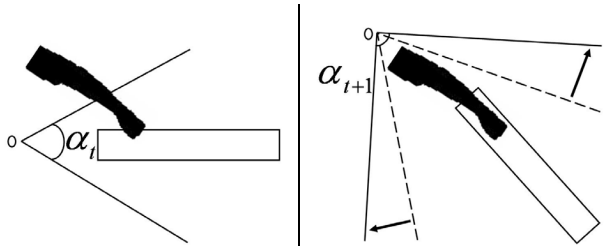


Fig. 1. Search area redefinition from iteration *t* to *t+1* allowing to match here the whole shape of an arm.

We define a priority relation between the different elements from a same level. Considering each limb independently, the head - torso unit is the one recovered with the most efficiency in many different configurations and is considered as the starting node of the human body model. Considering this, search space of each limb linked to the torso is reduced. Pixels previously matched with it are removed from the subject silhouette. Search areas are then redefined as explained earlier but can not contain any subject point being classified as torso. The same principle is applied in the refined level of the models: for example, pixels assigned to the lower arm are removed when searching the hand. The goal here is to limit effects of

occlusions or the attribution of the same pixels to different limbs. When this situation is truly encountered, the last limb in the hierarchy can be unmatched but this case is identified thanks to the matching quality measures.
In a similar way, pixels belonging to more than one search area of the same hierarchy level are removed. The proportion of concerned pixels is low but this process avoids them to be allocated to a same limb. What appears to be a lost of information is recovered by the "rigid" nature of the model elements: unless two search areas are totally superimposed, there will always be a subject part belonging to only one search area. On the contrary, that means that the limbs occlude themselves completely, which is already a piece of information about their localization.
Distances between the junction points of supposed linked elements of the model are also computed. Their integration within the matching process is made through a modification of the decision criteria *i.e.* the measure m:

$$m = \gamma m + (1 - \gamma)(1 - d_{norm}) \quad (2)$$

where d_{norm} is the normalized distance. We define d_{norm} equal to d/d_{max} where d is the measured distance between joints and d_{max} is the maximal distance allowed between two elements. Indeed, we consider that for a certain distance from the upper limb in the hierarchy, a matching can not be considered as reliable. Value for d_{max} is automatically defined as a ratio from the bounding box dimensions. Experimentally, γ has been set to 0.98. This value modifies m in order to privilege positions located near the element of higher priority in the model. But it does not constrain them only on the base of minimizing the distance. This process makes the system more flexible.

3.2.3. From Level 1 to the Next

Features correspondence for sharper levels of the model is made according to the ones obtained in the upper levels. Fig.2 shows the way new search areas are defined when going down in the model hierarchy. The limb is decomposed into new sub-rectangles. At each element corresponds a search area defined according to the one from which it is ensued. In our example $\alpha_1 = g_1(\alpha)$, $\alpha_2 = h_1(\alpha)$, $O_1(x, y) = g_2(O(x, y))$, and $O_2(x, y) = h_2(O(x, y))$. We have chosen in our experiments g_1,h_1,g_2 as the identity function to reduce the computational cost. The function h_2 implies a translational coefficient. Algorithms previously exposed to match model elements are quite similar to those used in upper levels. The major differences come from the search space which is quite reduced, both in terms of surface and angular values. One major advantage of the hierarchical modeling is that the model complexity in sharp levels is reduced thanks to the previous localization using coarse models. Even if all human body configurations are not recoverable with a first level, it provides an excellent starting point for a more precise model. The second hierarchy starts from the torso element and goes on to the extremities (hands, feet, and head elements). For each of them, joint points are defined in order to express and apply relational constraints as before.

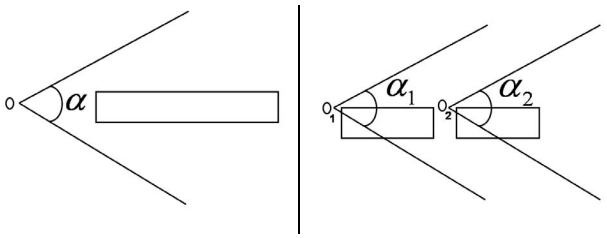


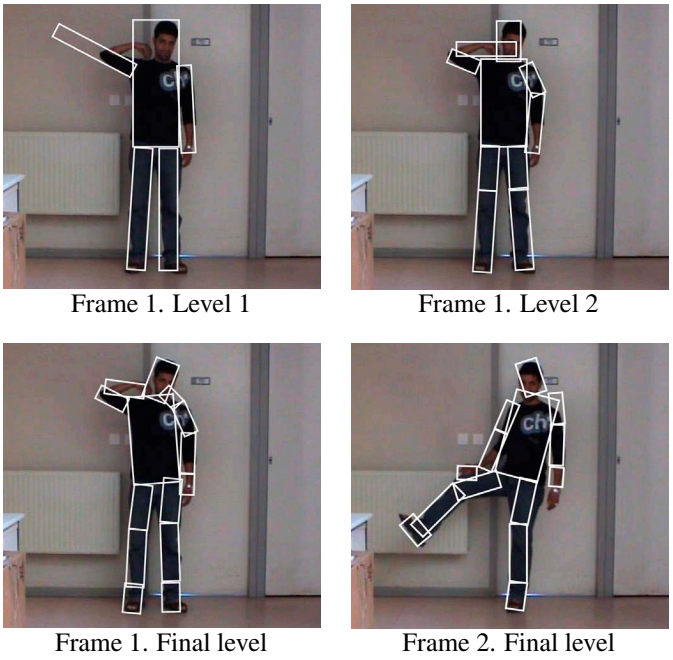
Fig. 2. Limb decomposition: new search area definition.

4. EXPERIMENTAL RESULTS

Methods presented in this paper have been experimented in various situations. Only one of them is provided here but results are generally quite stable from one to another. Fig.3 shows the three level matching on the same frame, and also the final result obtained with another example. The frame around the studied subject corresponds to model elements localization. We can see that the first matching is refined more and more down to the most accurate level of the model. We have tested our model on three kinds of situations. In the first scenario, the coarse level is matched, and then on the same frame, the second one is applied. We have noticed that the obtaining of the better matching for this level requires more than only one iteration. Indeed, because of the suppression of pixels belonging to more than one search area, some oscillations may happen. In all cases, they stop after 3 to 4 iterations. The same effect happens when the level three is applied: it also requires a few iterations to converge (3 to 4 again). This increases the global computational cost of the processing and we may wonder if the hierarchical approach is worth compared to a direct match with the model of the most accurate level. The result is that more iterations are required to converge (around 10) in the second case, and that the final matching is not always of the expected quality. This is illustrated in table 1: the second row corresponds to the number of required model elements matched to obtain the final result. The third row is the result accuracy computed as the percentage of silhouette pixels verified by the model. Both are mean values evaluated on eleven frames in different situations.
At last, we have experimented matching from a frame to the next one in a video sequence. The implementation of all the levels mainly depends on the motion velocity. When the studied subject performs no particularly fast motions, starting directly with the previous sharpest level provides the expected results. Obviously, if actual limb positions are too far from the previous ones, the matching process requires the use of a coarse model and works on a wider search area.

	One-level Model	Hierarch. Model
# required matching	152.8	108.4
Accuracy (%)	67.74	82.32

Table 1. Comparison of the hierarchical approach and a straight use of the sharpest model.



Frame 1. Level 1 Frame 1. Level 2

Frame 1. Final level Frame 2. Final level

Fig. 3. Refining of limbs localization by three model levels. Last frame is another example of final result.

5. CONCLUSION AND FORTHCOMING DEVELOPMENTS

In this paper, we described a human motion analysis system based on hierarchical modeling. This approach allows the obtaining of adopted postures at different levels of resolution, from coarse ones to the sharpest ones as shown by our experimental results. Thanks to a refining of the description at each step, our system adapts automatically (or according to specified needs) to the current resolution. This principle of hierarchical modeling can be applied generally to different other application fields. Our future developments will deal with the description of the motion itself for indexing and searching purposes. Until now, we have a succession of the subject postures and we will focus on the way to express them in terms of motion.

6. REFERENCES

[1] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, 1997, pp. 193–199.

[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and applications," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[3] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 814–827, 2003.

[4] L.W. Campbell and A.F Bobick, "Recognition of human body motion using phase space constraints," in *IEEE International Conference on Computer Vision (ICCV'95)*, 1995, pp. 624–630.

[5] R. Polana and R. Nelson, "Low level recognition of human motion," in *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994, pp. 77–82.

[6] K Rohr, *Human Movement Analysis Based on Explicit Motion Models*, chapter 8, pp. 171–198, Kluwer Academic Publishers, 1997.

[7] M.K. Leung and Y. Yang, "First sight: A human body outline labeling system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 4, pp. 359–377, 1995.

[8] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003.

[9] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[10] F. Lui, Y. Zhuang, Z. Luo, and Y. Pan, "A robust algorithm for video based human motion tracking," in *IEEE Pacific Rim Conference on Multimedia*, 2002, LNCS 2532, pp. 1161–1168.

[11] G. Borgefors, "Hierarchical chamfer matching: a parametric edge matching algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 849–865, 1988.

[12] A. Prati, I. Mikic, M.M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918–923, 2003.

[13] T. Fourès and P. Joly, "Defining search areas to localize limbs in body motion analysis," in *International Workshop on Adaptive Multimedia Retrieval (AMR 2003)*, 2003, LNCS 3094, pp. 147–163.