Tracking of Human Body using Multiple Predictors

Rui M. Jesus[1], Arnaldo J. Abrantes[1], and Jorge S. Marques[2]

[1] Instituto Superior de Engenharia de Lisboa, Postfach 351-218317001,
Rua Conselheiro Emído Navarro, n$^{\underline{o}}$ 1, 1940-014 Lisboa, Portugal
{rmfj,aja}@isel.pt
http://www.deetc.isel.ipl.pt

[2] Instituto de Sistemas e Robótica, Instituto Superior Técnico,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
jsm@isr.ist.utl.pt

Abstract. The objective of this work is to track the human motion from a video sequence, assuming that the motion direction is parallel to the image plane. Tracking the human body is a difficult task because the human body may have unpredictable movements, the complex structure and motion of the human body cause self-occlusions and it is difficult to accurately detect anatomic points in images without using artificial marks. This paper describes a body tracking algorithm, not relying on the use of artificial marks. The proposed system is able to learn from previous experience, and therefore its performance improves during tracking operation. The ability of the tracking system to gradually adapt to a particular type of human motion is obtained by using on-line learned multi-predictors, defined in a supervised way using information provided by a human operator. Typically, the human operator is called often to correct model estimates during the first few cycles of the observed motion, but the rate of human interference decreases as time goes on. Experimental results are presented to illustrate the performance of the proposed tracking system.

1 Introduction

Tracking the human body from a video sequence is a challenging problem with applications in many areas, such as bio-mechanic and virtual reality [5]. Since motion analysis involves the estimation of the human body configuration in a large number of images, several researchers have doing important progresses in attempting to automate this operation [2, 7–9]. Despite the reported advances, there are still open issues and the research in this area is more and more active. The main difficulties of the problem are related with (i) the complex structure of the human body; (ii) the occurrence of self-occlusions; and (iii) the difficulty to accurately separate the human body from the image background.

One way to circumvent these difficulties is the use of a set of visual marks glued to the body [5]. This technique makes the tracking operation much easier but limits drastically the range of possible applications. Another possibility, not relying on the presence of artificial visual marks, is to perform the analysis of the human motion manually, using a graphical editor. This is however clearly inadequate for many applications because it leads to a very time consuming task when large video sequences are involved. This paper proposes an interactive system to track the human body without using visual marks and with learning ability that reduces the number of user interventions to a minimum. The learning methods are used to improve the prediction of the body position and shape in future frames. The paper is organized as follows: section 2 describes the system overview; section 3 describes the multiple predictors technique used in this study; section 4 presents the experimental results and section 5 concludes the paper.

2 System Overview

The objective of the system described in this paper is to track the human body from the analysis of a video sequence. In this context, the shape, motion and visual appearance are three different measurable body features, which need to be modelled in some way. In the sequel, the interactive tracking system proposed in this paper is based on three models: a 2D articulated model, a dynamic model and an appearance model.

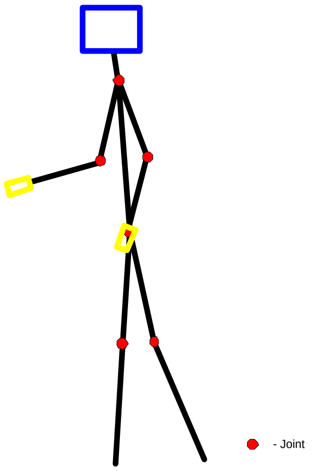


Fig. 1. Articulated model.

The 2D articulated model describes the body shape in the image and consists of 12 segments connected by 12 joints, as well as the head and the two hands (see fig. 1). The proposed model is closely related to the models used in Robotics to describe robotic manipulators [4].

Each segment of the articulated model represents a body part (neck, shoulders, torso, arms, forearms, hip, thighs and legs) which is assumed to be rigid and with known length. Each joint linking two segments is modelled by a rotation angle θ and by a translation vector t, which accounts for small displacements of the rotation center. The articulated model is therefore characterized by a set of parameters (joint angles and displacement vectors) and some of their derivatives.

The dynamic model rules the evolution of these parameters and is used to predict the model configuration. In simple cases, it could be defined by a stochastic linear equation. However, such linear dynamic model is inadequate to track fast or unpredictable human motions. Therefore, the proposed system adopts an hybrid dynamic model with two components, running in parallel: a trained linear stochastic equation; and a look-up table (dictionary) containing a list of exceptions (no linearly predictable model configurations).

The appearance model consists on a set of 1D and 2D RGB profiles, centered at specific points of the imaged human body. During tracking, those features are automatically detected in the image using template matching [1].

The proposed tracking system uses these three models to estimate the model configuration. In each frame, the algorithm performs the following five steps (see fig. 2):

- prediction - multiple predictors (the trained linear predictor and a set of predictors stored in the dictionary) are used, in parallel, to guess the position and shape of the body in the current frame;
- feature detection - for each predicted configuration, a set of visual features is obtained from the image, using template matching;
- filtering - this step updates each predicted configuration, using the corresponding visual features obtained in the previous step; this operation is performed using the equations of the extended Kalman filter [3];
- evaluation - each model estimate is evaluated by measuring its color matching relative to the appearance model;
- selection - selects the configuration with larger matching score.

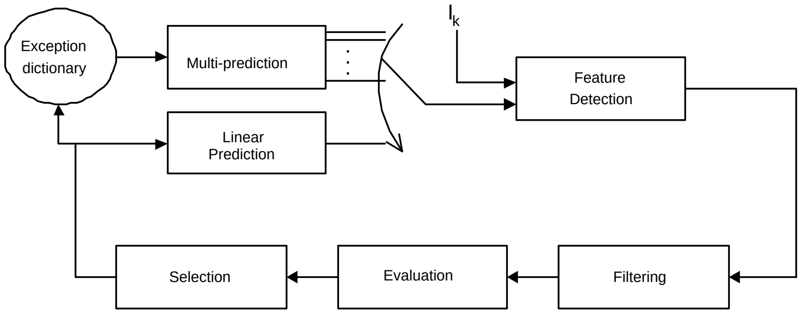


Fig. 2. Block diagram of the tracking system.

In this system the prediction step plays a central role. In fact, beyond its normal effect of enforcing the temporal coherence of the estimates, the predicted model is also used as starting point to localize the image features and to predict the occlusion of some body parts. In the sequel, only small prediction errors can be recovered by the filtering step.

In order to improve the quality of the prediction results, the system needs to incorporate, in some way, learning capabilities. Two learning phases were considered. The first phase consists of training the linear dynamic model used to describe the human motion. An improved motion model is obtained after this stage. However, in general this training model still fails when unpredictable movements occur. To solve this problem, the referred multi-prediction technique is used. This technique consists of creating a dictionary of exceptions containing the predictors for each position of the human body for which the automatic predictor failed before. Whenever one of these positions occurs in an image, the dictionary is automatically consulted and the multiple predictor technique is used.

3 Multiple Predictors

Detecting visual features and guessing which are the features occluded in the current frame are two related operations performed by the system which depend on the accuracy of the prediction stage. If prediction error is too large, the system may not be able to recover and have to be stopped by the operator. In order to improve the prediction results two types of dynamic models are used, simultaneously, by the system:

- a linear dynamic model, which enables the system to have a linear prediction of the model configuration in each new video frame;
- a look-up table (dictionary of exceptions), which enables the system to add on new predictors whenever a given model configuration is judged as an exception (in the sense that the next configuration may not be linearly predictable) and therefore it is added a new entry in the dictionary.

The linear dynamic model is defined by the following stochastic linear equation [6],

$$x_k = Ax_{k-1} + w_k \tag{1}$$

where x_k is the state vector containing the unknown model parameters and some of their derivatives, A is a square matrix characterizing the type of motion, and w_k is a white random vector with normal distribution, accounting for the random changes of motion parameters.

Matrix A is either specified by the human operator or estimated using a training algorithm. In this case, the training data is defined by a small number of previous model configurations obtained in a supervised way and A is estimated using the method of least squares,

$$\hat{A} = \arg\min_A \sum_k \|x_k - Ax_{k-1}\|^2 \tag{2}$$

This learning phase enables the linear predictor to adapt to the specific type of motion being tracked, increasing the performance of the tracking system.

Figure 3 shows three random sequences synthesized using equation (1), after performing the training of matrix A. They all correspond to realistic human motions of the same activity. The trained dynamic model is still able to cope with motion variability but now in a very controlled way.

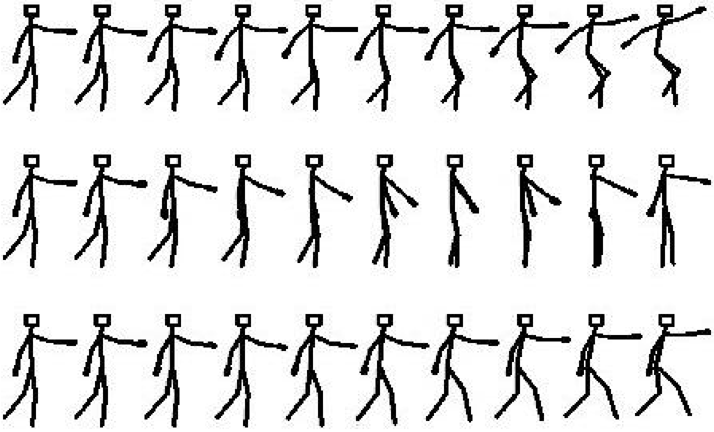


Fig. 3. Synthesized random sequences produced by the linear dynamic model after training.

Unfortunately, human motion is too complex, even in the case of performing a simple activity, to be dynamically modelled as linear. Typically, there are always some 'unexpected' movements (in the sense that they are not linearly predictable), which have to be considered separately as exceptions. Consequently, the tracking system considers a second learning phase, wherein a dictionary of exceptions is dynamically created while the video sequence is analyzed in a supervised way.

Every time automatic tracker fails, producing an erroneous estimate, the operator interferes performing the adequate corrections to the model. After that, the estimate obtained in the previous frame is used as a new entry into the dictionary, pointing to the edited configuration (or list of configurations, sharing the same entry). The dictionary of exceptions stores all model configurations where the linear predictor has failed as well as the corresponding solutions provided by the user. Table 1 shows how the data in the dictionary of exceptions is organized: the i-th entry contains a model configuration $x^{(i)}$ and the corresponding list of predictors, where $x_p^{(i,j)}$ is the j-th predictor of the i-th entry.

The dictionary of exceptions is consulted whenever one of his entries is close to the actual model configuration. More precisely, the predictors stored in the

Entries ($\hat{x}_{k-1}$)	Predictors ($\hat{x}_k^-$)
$x^{(1)}$	$x_p^{(1,1)} x_p^{(1,2)} \ldots\ x_p^{(1,n_1)}$
$x^{(2)}$	$x_p^{(2,1)} x_p^{(2,2)} \ldots\ x_p^{(2,n_2)}$
.	
.	
.	
$x^{(d)}$	$x_p^{(d,1)} x_p^{(d,2)} \ldots\ x_p^{(d,n_d)}$

Table 1. Structure of the dictionary of exceptions

i-th dictionary entry will be used to predict the model configuration at the k-th frame if the following two conditions are verified,

$$i = \arg\min_{n} \|\hat{x}_{k-1} - x^{(n)}\|, \tag{3}$$

and

$$\|\hat{x}_{k-1} - x^{(i)}\| < \gamma \tag{4}$$

where γ is a given threshold.

The tracking system uses the dictionary of exceptions as a multiple predictor generator. Whenever one of the exceptions occurs (one of the dictionary's entries), multiple predictors are automatically generated and used. At least two predictors are used when an exception is detected: the linear predictor obtained using equation (1),

$$\hat{x}_k^{(0)-} = A\hat{x}_{k-1} \tag{5}$$

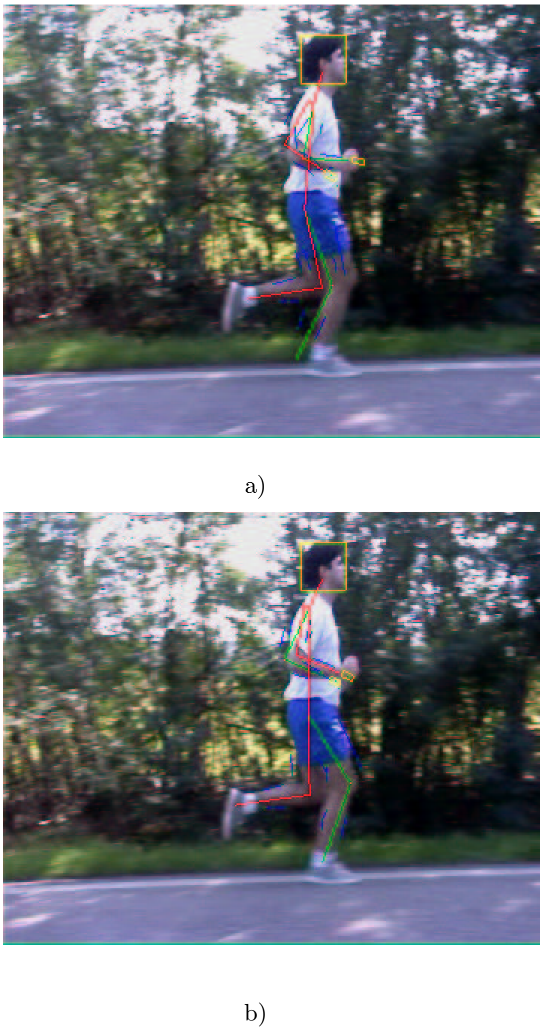
and the model configurations triggered by the exception,

$$\hat{x}_k^{(j)-} = x_p^{(i,j)} \tag{6}$$

where $j \in \{1, \ldots, n_i\}$.

The multiple predictors are used competitively and in parallel in two blocks of the tracking system (see fig. 2): feature detection and filtering. Different estimates (at least two) are usually obtained and a programmable criterion is needed to automatically select the best estimate. The criterion adopted is based on color histograms. For each estimate, a set of such histograms is evaluated, using small windows centered at the specific body points defined in the appearance model (see section 2). These histograms are compared (using L_1 metric) with the model histograms (stored in the appearance model) and the estimate obtaining the best global result is chosen.

Figure 4 shows an example of the use of two predictors to estimate the human body configuration. Figure 4a shows the linear predictor while in figure 4b it is observed the predictor provided by the dictionary.



a)

b)

Fig. 4. Example with 2 predictors: a) linear predictor; b) predictor provided by the dictionary of exceptions.

4 Experimental Results

The tracking system proposed in this paper was applied to four video sequences, corresponding to different activities (walking, cycling, writing in a board and running). The first three sequences were acquired with an analog color camera at 13 frames/seg and the last sequence (running) was acquired with a digital color camera at 25 frames/seg. The body movements are approximately periodic in three of the sequences: walking, cycling and running. In the other sequence (writing in a board) the body movements are not periodic and do not occur occlusions. In this case, the use of a simple dynamic model without training is enough. (figura a mostrar as 4 sequncias?...).

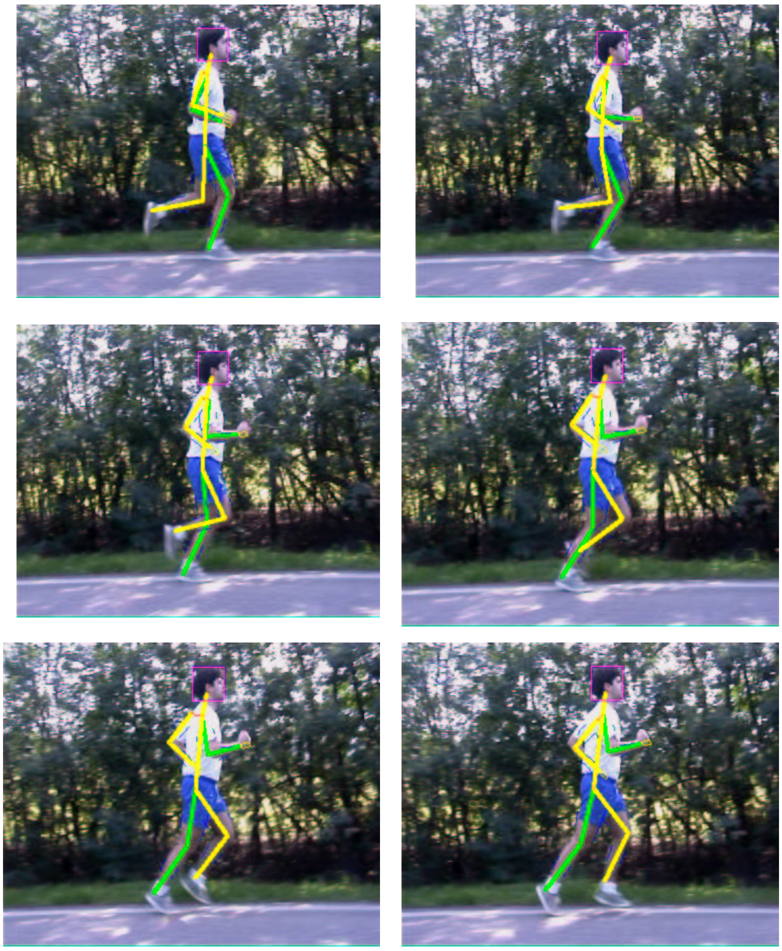
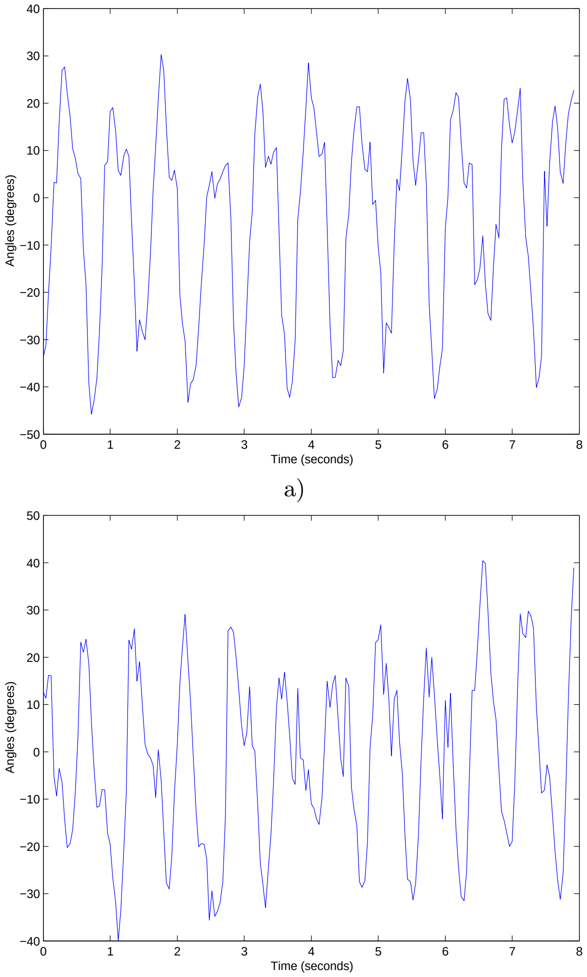


Fig. 5. Tracking results for the running sequence in six consecutive frames.

The automatic system managed to successfully track the three first sequences without human interventions during the second learning phase. The exception dictionary and the multi-prediction technique were needed only in the last sequence. Figure 5 shows six consecutive frames of this sequence showing the tracking results obtained with the interactive system proposed in this paper. This video sequence has 200 images. Only 9 of them were manually corrected. The multiple predictors technique was used 38 times, the linear predictor was chosen 8 times and the predictor given by the exception dictionary was chosen in the remaining times.



a)

b)

Fig. 6. Parameters evolution: a) right arm; b) left arm.

Figure 6 shows the evolution of the rotation angle associated with the arms. As expected, the evolution is periodic and the arms are in opposing phase. It should be stressed that a large number of self-occluded segments are present in this sequence. Furthermore the background (trees) is neither static nor homogeneous. The tracking algorithm manages to solve both difficulties well in most of the frames.

5 Conclusion

This paper describes a semi-automatic system to track the human body without artificial marks. The system has learning capability in the sense that the tracking performance improves during each experiment. Every time the user corrects the tracker output, the corrected model is stored in a dictionary. This information is then used automatically by the multi-prediction technique to correct similar cases in the future.

The main difficulties associated with the automatic operation of the tracker concern unpredictable motions and the presence of moving non-homogeneous background. The system is however able to deal with both of these difficulties well, most of time, as shown in the experiments described in the paper.

References

1. A. Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion.* Springer-Verlag London, 1998.
2. C. Bregler and J. Malik. Tracking people with twists and exponential maps. *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 1998.
3. R. Brown and P. Hwang. *Introduction Random Signals and Applied Kalman Filtering.* John Wiley and Sons, 1992.
4. J. Craig. *Introduction to Robotics Mechanics and Control.* Addison-Wesley, 1955.
5. D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
6. A. Gelb. *Applied Optimal Estimation.* MIT press, Cambridge, Mass, 1974.
7. D. Hogg. Model based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.
8. I. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *Internacional Journal of Computer Vision*, 30(3):191–218, 1998.
9. H. Sidenblabh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *European Conf. on Computer Vision*, 2000.