Constraining Human Motion for Efficient Tracking with a Particle Filter

Ignasi Rius* and Carles Fernandez* and Mikhail Mozerov* and Jordi Gonzàlez+

* *Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*
E-mail:{irius, perno, mozerov}@cvc.uab.es

+ *Institut de Robòtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Spain*
E-mail:poal@cvc.uab.es

Abstract Particle filters are one of the most commonly used techniques for full-body human tracking. However, given the high-dimensionality of the involved models, the number of required particles make the problem computationally very expensive. To overcome this, we present an action-specific model of human postures which eases the process by guiding the prediction step of the particle filter, so only feasible human postures are considered. Thus, this model-based tracking approach samples from a first order motion model only those postures which are accepted by our action-specific model. In this manner, particles are propagated to locations in the search space with most a posteriori information avoiding particle wastage. We show that this scheme improves the efficiency and accuracy of the overall tracking approach.

Keywords: Motion Analysis and Recognition, Particle Filters

1 Introduction

Full-body 3D human tracking from a monocular image sequence is one of the most challenging problems from visual human motion analysis. However, the number of difficulties related to the problem are very large. Among others, the shape and appearance of a human body in 2D images may change drastically over time due to changing lighting conditions, loose fitting clothes and background clutter. Additionally, one must deal with 2D-3D projection ambiguities, and self and non-self occlusions of body parts. Hence, only a reduced number of DOF present in the model are directly observable from 2D images. Finally, the implied models are very high dimensional, non-linear, and may suffer from kinematic ambiguities and singularities. To overcome these issues, many approaches make use of Bayesian filtering techniques combined with carefully designed search strategies of the solution space [1, 6, 7, 9, 8]. When the involved distributions are non-Gaussian, the computation of model parameters over time can be approximated by means of a particle filter. This probabilistic framework can deal with multiple hypotheses, and brings a principled way to incorporate *a priori* knowledge about human motion into the tracking, so the solution space can be explored in a more efficient manner.

Particle filters supply a powerful tool for representing and propagating complex posterior distributions. However, the number of needed particles grows exponentially as the number of dimensions to be tracked does [4]. This fact is obvious in the human motion tracking case, due to the high DOF needed to represent human postures. For this reason, it is necessary to make particle filters more efficient. For example, the *annealed* particle filter aims to reduce the number of required samples by successively pruning less likely hypotheses [1]. Alternatively, it is possible to use efficient motion models which concentrate particles in areas of interest. In [7], Sidenbladh et al. learnt a dynamic model from a pre-recorded set of human motions, and predictions were made assuming a Gaussian distribution over subsequences of the learned motions. However, the model can only predict postures which were present in the motion database.

Likewise, we propose a posture-based human action space for modelling feasible postures within an action. This model is used to constrain human postures within the framework of a particle filter respon-

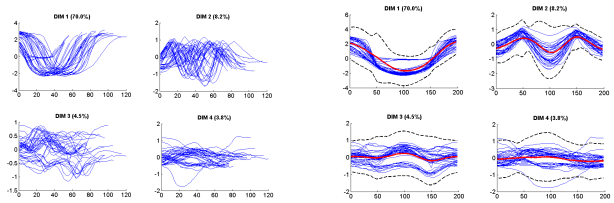
sible for tracking the human body motion. In such a recursive model-based tracking approach, human postures are projected forward by means of a dynamic model, and they are subsequently updated according to the measurements obtained from images. As a result, we must define both the dynamic model and the fitness function of human postures to images. In this work, predictions are made according to a dynamic model which focuses and constrains human postures only to a set of feasible postures within the performance of a particular action.

The remainder of this paper is organised as follows. In Section 2 we present the training of our action-specific model of human postures using real data acquired with a commercial Motion Capture system. This action model is used to determine whether a human posture belongs to a particular action or not. Section 3 introduces the tracking framework. We define a dynamic model based on a first order motion model constrained to the postures which are accepted by the action model. Moreover, we present a fitness function based on the overlapping area between the projection of the body state and the body region obtained from image segmentation. In Section 4 results of the tracking approach are presented for a performance not considered in the training set. Finally, Section 5 discusses the conclusions and future research.

2 Learning posture constraints

The 3D human body model used in this work is composed of 12 limbs with 3 DOF per joint expressed as relative angles in a 3D polar coordinate system. Using a commercial Motion Capture System, we acquired 45 performances, in average, of 9 different actions performed by 9 different actors. From the observed motion, we aim to automatically learn per each action, which human postures are feasible during the performance of that particular action. Towards this end, we first express all the training postures for action A in a lower dimensional representation called *aSpace* [2] which is computed as follows:

Let ϕ be a 36-dimensional vector representing a particular human posture, and Φ be a sequence of human postures, hereafter performance. Then, for a particular action A, we compute PCA over all the training performances Φ_j for that action. The resulting PCA-like space - called *aSpace* - will be denoted as Ω^A. The projections $\tilde{\Phi}_j$ on the *aSpace* of Φ_j constitute the lower dimensional version from the original data. Subsequently, we aim to characterize the *shape* of the training performances for action A within the *aSpace*. Since each performance $\tilde{\Phi}_j$ may be composed of a different number of postures and may exhibit different speeds, we use the method described in [5] for synchronizing all the performances from the training set. As a result, we obtain a synchronized version of the training set. Fig. 1 shows the first 4 dimensions of the *aSpace* from the non-synchronized (Fig. 1.(a)) and the synchronized (Fig. 1.(b)) versions of the training set for a bending action.



(a) (b)

Figure 1: Before (a) and after (b) synchronization of the training set using key-frames.

As a result, we can put in correspondence postures between different training performances. Therefore, we compute the synchronised mean performance $\hat{g}^A$, and the standard deviation σ_k^A for each k-th posture, using all the synchronised performances $\tilde{\Phi}_j$. In Fig.1.(b), we show the synchronised training performances (thin lines) and its mean performance (thick line) for a bending action. The dashed black line corresponds to three times the standard deviation computed from the mean. Finally, our action model is defined as:

$$\Gamma^A = (\Omega^A, \hat{g}^A, \sigma_k^A), \tag{1}$$

where Ω^A defines the *aSpace*, $\hat{g}^A$ stands for the synchronised version of the mean performance, and σ_k^A is the observed standard deviation.

The learnt action model will be used in the prediction step of the particle filter to probabilistically determine whether a posture belongs to action A or not. The probabilistic framework used to face the tracking problem is described in the next section.

3 Using the posture constraints

The Bayesian filter recursively estimates the state of the tracked object at each time step given the evidences (image data) up to that moment. It decomposes the problem in two differentiated steps, i.e. the *prediction* and *update* steps. The prediction step projects forward the model parameters to the next time step by means of a *dynamic model*. Then the update step makes use of a *likelihood* probability function in order to evaluate the fitness of the predictions to the evidences available at each moment.

Formally, within the Bayesian filtering framework, we formulate the computation of the *posterior* distribution $p(\phi_t|\mathbf{I}_t)$ of our model parameters over time as follows:

$$p(\phi_t|\mathbf{I}_t) \propto p(I_t|\phi_t) \int p(\phi_t|\phi_{t-1})\, p(\phi_{t-1}|\mathbf{I}_{t-1})\, d\phi_{t-1}\,, \tag{2}$$

where ϕ_t is a 36-dimensional vector from our body model representing a particular pose of the human body at time *t*, $\mathbf{I}_t$ is the image sequence up to time *t*, $p(I_t|\phi_t)$ is the *likelihood* of observing the image I_t given the parametrization ϕ_t of our model at time *t*, and finally $p(\phi_t|\phi_{t-1})$ is the *dynamic model*.

We use particle filtering techniques in order to approximate the true *posterior* pdf by means of a discrete weighted set of samples. Hence, whilst the likelihood function decides which particles are worth to propagate, the dynamic model is responsible for guiding the exploration of the space of solutions. The *posterior* $p(\phi_t|\mathbf{I}_t)$ represents all the current knowledge about the model state we have extracted from image measurements. We can estimate the state ϕ_t at a particular time step by computing the mean of the posterior pdf.

The number of samples -or *particles*- determines the accuracy and the speed of the tracker. However, the computational cost of particle filters mainly comes from the computation of the likelihood function from image measurements [10]. Additionally, the number of needed particles grows exponentially as the number of dimensions of the model to be tracked does [4]. Therefore, given the high-dimensionality present in human motion tracking, we need to design efficient search strategies to lower the number of particles needed. In other words, the dynamic model from the prediction step of the particle filter should be generic enough to track any motion, but specific enough to focus particles only to areas with high a posteriori information.

3.1 Constrained motion model

The action-specific posture model constitutes *a priori* knowledge on human motion which can be incorporated into the Bayesian tracking framework by means of the dynamic model $p(\phi_t|\phi_{t-1})$ from Eq. (2). We aim to define a dynamic model which samples only those postures which are feasible during the performance of a particular action A, based on a 1st order motion model. Thus, the prediction step of the particle filter is designed as a two-step process. First, we project forward the particle set $\{\phi^s_{t-1}\}$ following a 1st order motion model plus some Gaussian noise, i.e.,

$$\hat{\phi}^s_t = \phi^s_{t-1} + V_{t-1} + \eta(\sigma_\phi), \tag{3}$$

where ϕ^s_{t-1} denotes the particle s at time $t-1$, and $\hat{\phi}^s_t$ is the prediction for this particle. V_{t-1} is the velocity term computed at time at time $t-1$, and $\eta(\sigma_\phi)$ is a Gaussian diffusion term. To determine σ_ϕ, we used a constant velocity model to predict each performance of the training set. Then, σ_ϕ was computed as the standard deviation of the average error committed. Subsequently, we update the term V_t according to $V_t = \alpha V_{t-1} + (1-\alpha)(\phi_{t-1} - \phi_{t-2})$, where α is a learning coefficient, and ϕ_{t-1}, ϕ_{t-2} correspond to the estimated state of the human body at the two previous time steps.

Secondly, we filter those predictions $\hat{\phi}^s_t$ which are not accepted as feasible postures during the performance of the action A_i by our action-specific model. If a prediction $\hat{\phi}^s_t$ is rejected, we resample from Eq. (3) until a feasible posture is generated for this particle. Finally, the new set of predicted particles $\{\phi^s_t\}$ at time t is constituted by those predictions $\hat{\phi}^s_t$ which were accepted by the action model.

As a result, we reformulate Eq. (2) including the action model into the prediction step as

$$\begin{aligned} p(\phi_t|\mathbf{I}_t) \;\propto\; & p(I_t|\phi_t) \int p(\phi_t|\phi_{t-1}, \Gamma^A) \cdot \\ & p(\phi_{t-1}|\mathbf{I}_{t-1})\, d\phi_{t-1}. \end{aligned} \tag{4}$$

Now, by applying the Bayes' rule and assuming independence between ϕ_{t-1} and Γ^A, i.e. only current postures are constrained by the action model, we can further decompose Eq. (4) as

$$p(\phi_t|\mathbf{I}_t) \propto p(I_t|\phi_t) \int p(\phi_t|\phi_{t-1})\, p(\phi_t|\Gamma^A) \cdot p(\phi_{t-1}|\mathbf{I}_{t-1})\, d\phi_{t-1}, \quad (5)$$

where $p(\phi_t|\Gamma^A)$ is a function which determines whether a particular posture ϕ_t belongs to action A or not defined as follows:

$$p(\phi_t|\Gamma^A) = \begin{cases} 1 & if(|\tilde{\phi}_{t,d}, \hat{g}^A_{j,d}| < 2 \cdot \sigma^A_{j,d}), \\ & \forall d = 1..D \\ 0 & otherwise \end{cases}, \quad (6)$$

where $\tilde{\phi}_t = (\tilde{\phi}_{t,1}, ..., \tilde{\phi}_{t,D})^T$ is the projection of ϕ_t in the *D*-dimensional *aSpace*. $\hat{g}^A_j$ is the j-th posture from the mean performance computed for the action A which probabilistically matched $\tilde{\phi}_t$, i.e., we draw $\hat{g}^A_j$ from a Gaussian conditional distribution assuming that $\tilde{\phi}_t = \hat{g}^A_j + \eta(\Delta)$, where Δ is empirically determined from the training set. $\sigma^A_j = (\sigma^A_{j,1}, ..., \sigma^A_{j,D})$ stands for the learnt standard deviation of the j-th posture for the action A. Notice that the level of filtering depends on the number of dimensions D considered in the *aSpace* representation.

By defining this filtering method, we prune those predictions which are more distant than two times the learnt standard deviation from the matched posture of a particular action. As a result, our dynamic model predicts feasible human postures avoiding particle wastage on postures which are not likely to appear during the performance of a particular action.

3.2 Image Measurements

The *likelihood* function $p(I_t|\phi_t)$ computes how likely is to observe the image I_t given a human body posture ϕ_t. In this paper, we implemented a likelihood function based on the image region filled by the human body. Hence, the human body model has been fleshed out with 3D volumetric primitives consisting in 3D cylinders. As a result, we synthesise an image $\breve{I}_{\phi^s_t}$ of the region defined by a particular parametrization ϕ^s_t of the human body model. For simplicity and efficiency, we have simplified the 2D projections onto the image plane from the limbs' cylinders as rectangles.

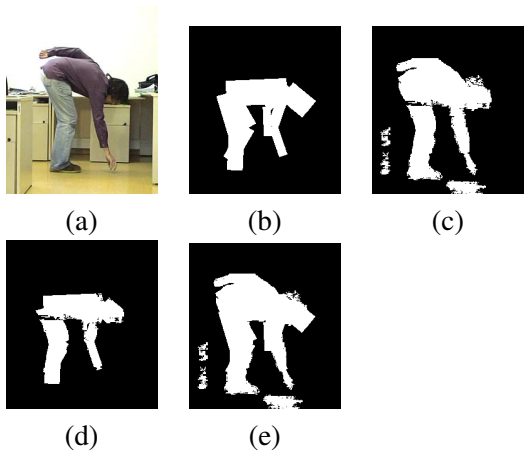


Figure 2: I_t (a), $\breve{I}_{\phi_t}$ (b), $\hat{I}_t$ (c), I^{OV}_{t,ϕ_t} (d) and I^{U}_{t,ϕ_t}(e) images from the likelihood computation. See text for details.

On the other hand, we extract the true region filled by the body in the current image I_t by applying a background subtraction algorithm from Horprasert et al. [3]. This pixel-wise algorithm needs to be trained with several background-only frames beforehand. Then, for each frame to be segmented, the algorithm computes for each pixel the normalised distortion on chromacity and brightness with respect to the learnt background model. Based on this values, each pixel is classified as background, foreground, shadow, or highlight. We denote the segmented body region image as $\hat{I}_t$. Finally, the *likelihood* is computed based on the overlapping area between the synthesised and the segmented images, i.e.,

$$p(I_t|\phi_t) \propto \frac{\sum_x \sum_y (I^{OV}_{t,\phi_t}(x,y))}{\sum_x \sum_y (I^{U}_{t,\phi_t}(x,y))}, \quad (7)$$

where I^{OV}_{t,ϕ_t} refers to the overlapping region between $\breve{I}_{\phi_t}$ and $\hat{I}_t$, I^{U}_{t,ϕ_t} is the union of both regions. The notation $I(x,y)$ is used to make reference to the pixel of I at column x, row y. As a result, we assign maximum weight to those postures whose synthesised image coincide totally with the segmented one, and lower values otherwise. Fig. 2 shows the images I_t (a), $\breve{I}_{\phi_t}$ (b), $\hat{I}_t$ (c), I^{OV}_{t,ϕ_t} (d) and I^{U}_{t,ϕ_t}(e) computed at a particular time t of the algorithm.

4 Experimental results

To test this work we used a training set of 40 performances of a bending action carried out by 9 different

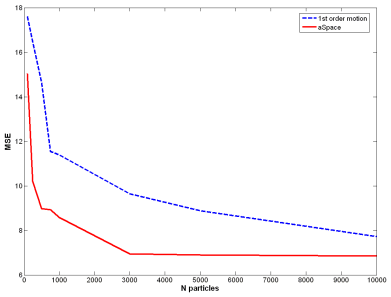
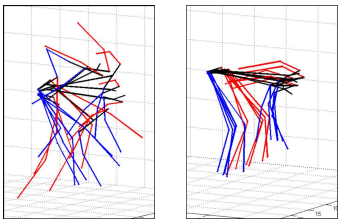


Figure 3: MSE obtained with both approaches.



(a) (b)

Figure 4: Predictions of the *aSpace* and 1st order motion approaches.

actors. However, the approach is easily extensible to other sets of actions. Hence, we have tested the tracking approach using a bending sequence not present in the training set, consisting in 86 frames from which we have 3D ground truth data available.

The number of D dimensions considered when building the *aSpace* representation determines the degree of adaptation of the action model to the training data. Hence, too low values for D result in a poor filtering effect, since too many particles with low a posteriori information will be accepted by the action model. On the other hand, too high values lead to overfitting to the training set, since the action model only accepts particles that are almost equal to postures used to learn the action model. To test this work, we used $D = 13$ dimensions which proved to achieve a good compromise between generality of the model and non-feasible postures rejection.

To test the effectiveness of the approach, we compared the results obtained using our action model against a first order motion model without any filtering method. We repeated the same experiment varying N from 100 to 10000 particles, with $D = 13$ and the learning coefficient of the velocity set to $\alpha = 0.5$.

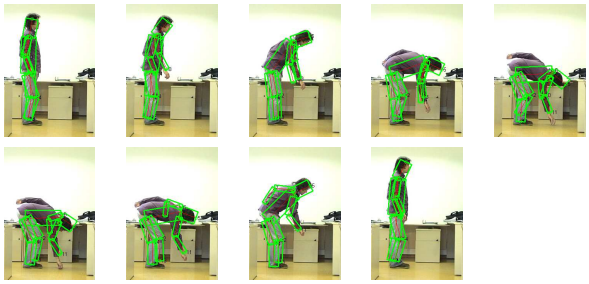


Figure 5: Estimated frames 1, 11, 21, 31, 41, 51, 61, 71 and 81.

In Fig.3 we show the obtained error for the *aSpace* filtering method (solid line) and the simple first order motion model (dashed line). The error was computed as the average Mean Square Error (MSE) of the relative angles between the final estimated postures -computed as the expectation of the posterior pdf- and the ground truth data from the sequence. We may observe that the action model overperforms the 1st order motion model in all the experiments. Furthermore, the error for the *aSpace* filtering method quickly stabilises around 7 at $N = 3000$ particles. One may observe that we obtain similar error measures using 2000 particles with the *aSpace* approach than 10000 particles without any filtering. Additionally, with very few particles -below 1000-, our approach quickly lowers the error and tends to stabilise, while the 1st order motion model approach gives very high error rates. Hence, our approach never totally looses the tracked object since it never produces non meaningful postures. This is depicted in Fig.4 where a frame of the tracked sequence is plotted with a randomly selected set of predicted postures projected over it for (a) the *aSpace* approach, and (b) the 1st order motion model approach. One may observe that the latter leads to unlikely and non feasible human postures for this action, while the *aSpace* filtering approach predicts natural and coherent human postures.

Finally, selected frames of the final estimated sequence are shown in Fig.5 for $N = 5000$ particles. We may observe, that the left arm is confused with the right arm in the first frames. This is an expected behaviour, since the right arm is totally occluding the left one, so the likelihood function gives us no clue for evaluating the proper arm position. However, in the second half of the sequence, the left arm tends to

its correct position since it becomes slightly visible in those frames, so the likelihood function is higher for postures covering the left arm. The ability to handle multiple hypothesis of the particle filtering framework is proved to be very suitable, since it can recover from a self-occluding situation where the likelihood function doesn't provide the right maxima.

5 Conclusion

We have presented an efficient tracking approach based on particle filtering for full-body human tracking, which makes use of an action model to guide the prediction step of the particle filter. Despite the use of a simple likelihood function, the space of possible solutions is explored in an efficient manner since only feasible human postures are generated by our dynamic model. We compared the overall error of our tracking approach against a first order motion model without filtering in the *aSpace*. Results point out that the action model approach drastically reduces the number of particles needed to track a 36 DOF human body model, thus reducing the high computational cost inherent to typical particle filter approaches. Moreover, given the PCA-like definition of the action space, the degree of dependence of the predictions to the training data set can be tuned by considering more or less dimensions when building the space.

Future work relies on extending this approach to a more general set of actions, so we can track any action and transitions between actions. Furthermore, the likelihood function needs to be improved in order to include other image-based cues like color or edges, so it provides more reliable information for evaluating the predicted poses. Moreover, we need to define a method for handling self-occlusions based on predicting which body parts are visible at each time step. Finally, it is possible to improve the action model by considering other formulations which may improve the pruning effect providing more accuracy and efficiency to the overall tracking process.

Acknowledgments: This work has been supported by the Generalitat de Catalunya Research Department, by the EC grant IST-027110 for the HERMES project and by the Spanish MEC under projects TIC2003-08865 and DPI-2004-5414. J. Gonzàlez acknowledges the support of a Juan de la Cierva postdoctoral fellowship from the Spanish MEC.

References

[1] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[2] J. Gonzàlez. *Human Sequence Evaluation: the Key-frame Approach*. PhD thesis, Universitat Autònoma de Barcelona, 2004.

[3] T. Horprasert, D. Harwood, and L.S. Davis. A robust background subtraction and shadow detection. In *Proc. Asian. Conf. on Comp. Vision*, January 2000.

[4] J. MacCormick and M. Isard. Partitioned sampling, articulated objects and interface-quality hand tracking. Dublin, 2000. ECCV'00.

[5] Mikhail Mozerov, Ignasi Rius, Xavier Roca, and Jordi Gonzàlez. 3D human motion sequences synchronization using dense matching algorithm. In *DAGM'2006*, Berlin, Germany, September 2006.

[6] H. Ning, T. Tan, L. Wang, and W. Hu. Kinematics-based tracking of human walking in monocular video sequences. *IVC*, 22:429–441, 2004.

[7] H. Sidenbladh, M.J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV (1)*, pages 784–800, 2002.

[8] R. Urtasun, D.J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *IEEE International Conference on Computer Vision (ICCV05)*, volume 1, pages 403–410, 2005.

[9] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *CVIU*, 74(3):174–192, June 1999.

[10] Y. Wu, J. Lin, and T.S. Huang. Analyzing and capturing articulated hand motion in image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1910–1922, 2005.