A 3D Dynamic Model of Human Actions for Probabilistic Image Tracking

Ignasi Rius, Daniel Rowe, Jordi Gonzàlez, and Xavier Roca

Centre de Visió per Computador/Department of Computer Science
Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
irius@cvc.uab.es

Abstract. In this paper we present a method suitable to be used for human tracking as a *temporal prior* in a particle filtering framework such as CONDENSATION [5]. This method is for predicting feasible human postures given a reduced set of previous postures and will drastically reduce the number of particles needed to track a generic high-articulated object. Given a sequence of preceding postures, this example-driven transition model probabilistically matches the most likely postures from a database of human actions. Each action of the database is defined within a PCA-like space called *UaSpace* suitable to perform the probabilistic match when searching for similar sequences. So different, but feasible postures of the database become the new predicted poses.

1 Introduction

The analysis of motion in image sequences involving humans has become a great interest area in computer vision because of the wide amount of promising applications it brings, i.e. automatic surveillance, sports performance analysis, advanced interfaces, augmented reality and motion synthesis among others. This challenging domain is referred as *Human Sequence Evaluation* (HSE) in the framework presented by Gonzàlez in [3], and provides a general scheme for producing useful human motion descriptions from images suitable to be used for such applications.

The HSE framework divides the task of evaluating sequences of images involving human motion in several layers or modules, each one encapsulating different domains of knowledge. Hence, the interpretation of human motion is treated as a transformation process from level to level. We focus on the transformation process between the 3D human body configurations from 2D image sequences. This tracking and reconstruction task of articulated 3D human motion is a key point of HSE and has become a wide research topic in the last years [8].

Among others, one critical issue is the high dimensionality and the non-linearity of the articulated rigid objects to be tracked. For instance, if we consider a 3D body model of 12 joints with 3 Degrees of Freedom (DOF) per joint, it results in a model with 36 DOF, which means that our tracking algorithm must estimate at least 36 parameters at each time step. So several optimization techniques are usually applied.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3522, pp. 529–536, 2005.

The remainder of this paper is organized as follows. Section 2 explains the probabilistic framework used to face the tracking problem. Section 3 describes the human action model employed in this work. Section 4 focuses on the problem of the probabilistic search within the space of actions. Section 5 shows some experimental results, and section 6 concludes this paper.

2 Probabilistic Tracking Framework

The objective of visual tracking is to estimate the parameters of our model ϕ_t at time t given the sequence of images $\boldsymbol{I_t}$ up to that moment. In other words, we need to compute the posterior *probability density function* (pdf) $p(\phi_t|\boldsymbol{I}_t)$ over the parameters ϕ_t of the model to be tracked at time t. Thus, using the Bayes' rule, we formulate the computation of our model parameters over time as [2]:

$$p(\phi_t|\boldsymbol{I_t}) = k\, p(I_t|\phi_t) \int p(\phi_t|\phi_{t-1})\, p(\phi_{t-1}|\boldsymbol{I_{t-1}})\, dt\,, \tag{1}$$

where ϕ_t represents the pose of the human body at time t, $\boldsymbol{I_t}$ is the image sequence up to time t, k is a normalizing factor, $p(I_t|\phi_t)$ 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 *temporal prior*, or dynamic model in this work.

The recursive Bayesian filter provides the theoretical optimal solution. It decomposes the problem in two differentiated steps, i.e. *prediction* and *update*. On the prediction step, a dynamic model is used to derive the prior pdf at time t from the already computed posterior pdf at time *t-1*. On the update step, the *likelihood* function is used to compute the posterior pdf at time *t*.

Unfortunately, Eq.(1) relies on an integral which cannot be analytically calculated unless strong assumptions about Gaussianity and linearity on the involved distributions are made. Instead, we can approximate the true posterior distribution $p(\phi_t|\boldsymbol{I_t})$ by means of a *particle filter* [1, 5]. Particle filtering is based on Monte Carlo Simulation, thus, our posterior distribution at time t is represented by a set of samples or particles that in our case define a particular human body posture. Each particle has its own probability of being propagated over time depending on how likely is its corresponding body posture to be found on the image I_t. If a particle is selected to be propagated at time t, a transition model or *dynamic model* is used to predict the new location in the parameter space at time *t+1*, i.e. the new particle at the following time step.

This Bayesian model-based tracking approach brings us a principled way for considering multiple hypotheses about the human body posture, and allows us to integrate prior knowledge about the non-linear human dynamics into the tracking making it more robust and efficient.

Since the dimensionality of the parameters space is very large in 3D human motion tracking, a large number of particles may be needed to successfully track our model parameters over time. However, the number of particles grow exponentially with the model dimensionality [6]. To overcome this, we need an appropriate dynamic model in order to reduce the number of particles needed

to make the tracking task possible. This *temporal prior* should capture the behaviour of human motion accurate enough to predict only new feasible postures, but generic enough to be able to track any actor and any human motion.

The aim of this work is to present a temporal prior derived from [7], which is suitable to be used by the particle filter. Hence, the proposed model will propagate the parameters of our human body model over time while reducing the number of particles required to track a 3D human body model during a performance. The goal is focused in generating only the most plausible body postures within the performance of a particular action, rather than attempting to randomly propagate the parameters of a generic, high-articulated object.

3 Human Action Modeling Using *p-actions*

Our method learns the implicit probabilistic model of 3D human motion by using an example-based approach. Our dynamic model will use a database of learnt actions in order to predict the most suitable future body poses given a reduced set of the history of estimated poses. We perform a probabilistic search within a PCA-like space, called *UaSpace* [3] , which is built from a training set of human motions acquired with a commercial Motion Capture system.

In this work we use the human action model and the human action space defined in [4], called *p-action* and *aSpace* respectively. We show how to employ this action model to develop a dynamic model suitable to be used for human posture prediction which focuses and restricts the search space to those postures with highest likelihood values in factored sampling techniques.

An action will be represented as a sequence of postures, so a proper body model is required, which is learnt from examples. The training data has been acquired using a commercial Motion Capture system. A set of 19 reflective markers were placed on several characteristic points of the subject's body. The body model employed is composed of twelve rigid body parts (hip, torso, shoulder, neck, two thighs, two legs, two arms and two forearms) and fifteen joints. These joints are structured in a hierarchical manner, where the root is located at the hip. We represent the human body by 37 parameters which describe the relative elevation and orientation of each limb which are natural to be used for limb movement description. See [3, 4] for further details.

As a result, the training data set for each action $\boldsymbol{A}_i$ is composed of r_i sequences $\boldsymbol{A}_i = \{\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_{r_i}\}$, each one corresponding to a cycle or a performance of the action to be modeled.

Thus, we define the complete set of human postures for an action $\boldsymbol{A}_i$ as:

$$\boldsymbol{A}_i = \{x_1, x_2, ..., x_{f_i}\}, \tag{2}$$

where each x_j of dimensionality $n \times 1$ stands for the 37 values of the human body model described previously and f_i refers to the overall number of training postures for this particular action $\boldsymbol{A}_i$.

Then, we perform a Principal Component Analysis (PCA) on the training set $\boldsymbol{A}_i$, and compute its *aSpace* as defined in [4]. Afterwards, for each performance

$\boldsymbol{H}_j$, we consider its projections within the *aSpace* of the captured postures as the control values for an interpolating curve $\mathbf{g_j}(p)$, which is computed using a standard cubic-spline interpolation algorithm. The parameter p refers to the temporal variation of the posture, which is normalized for each performance, that is, $p \in [0, 1]$. This process is repeated for each performance of the learning set, and a mean manifold $\mathbf{g}(p)$ is obtained by interpolating between the means of $\mathbf{g_j}(p)$ for each index p.

After, a key-frame set $\boldsymbol{K}$ is found for each action by using the Mahalanobis distance, and the final human action model is represented as a parametric manifold $\mathbf{f}(p)$, called *p-action*, which is built by interpolation between those key-frames.

For our purposes, we need a common space where all the *p-actions* can be represented. We denote this space as the Universal *aSpace* or *UaSpace*, and is defined in the same fashion as the single *aSpace* for each action, but using all the postures from all the performances from all the actions of our database. After applying PCA, the first $b^U = 15$ eigenvectors are chosen to determine the 95% of the variance, and will constitute the basis of the space $\mathbf{\Omega^U}$ where all the *p-actions* will be represented.

Finally, an action $\boldsymbol{A_i}$ is modeled within the *UaSpace* as:

$$\mathbf{\Gamma}^{A_i} = (\mathbf{\Omega^U}, \boldsymbol{K}^{A_i}, \mathbf{f}^{A_i}), \tag{3}$$

where $\mathbf{\Omega^U}$ defines the eigenvectors and the eigenvalues of the *UaSpace*, and $\boldsymbol{K}^{A_i}$, $\mathbf{f}^{A_i}$ correspond to the key-frames and the parametric manifold that defines the *p-action*, respectively.

Closer points between different manifolds correspond to similar human postures of several actions. In fact, the distance between two points in the *UaSpace* can be considered as a measure of similarity between human postures.

4 Probabilistic Dynamic Model

Multiple hypotheses can be generated by considering different dynamical models. We consider the human action model $\mathbf{\Gamma}^{A_i}$ defined before as the basis for those dynamical models which can help to generate new samples over time within a probabilistic framework. As postures can be shared among different actions (such as in sitting, squatting and tumbling, for example), we need a probabilistic model which can deal with multiple hypotheses while predicting new postures. Fortunately, the *UaSpace* provides the framework where multiple motion models can be learnt and recognized.

The goal of a dynamic model is to predict new body postures ϕ_{t+1} at time $t + 1$ given the history of the observed motion Φ_t from time $t - d$ to time t. In our approach, the motion database used to build the dynamic model is derived from all the *p-actions* represented within the *UaSpace* described in the previous section. In order to obtain a set of body postures from each parametric manifold, each cubic-spline $\mathbf{f}^{A_i}(p)$ is sampled at a constant rate considering that $p \in [0, 1]$.

We denote each projected human posture of dimension b^U within the *UaSpace* as ψ_i, and $\Psi_i = [\psi_i^T, ..., \psi_{i-d}^T]^T$ refers to the $(d \times b^U)$-dimensional vector containing all the postures in the database from location $i-d$ to location i, i.e. the history of motion of the last d postures. In a similar fashion, let ϕ_t be the estimated posture at time t in the tracking framework described in section 2, and $\Phi_t = [\phi_t^T, ..., \phi_{t-d}^T]^T$ the estimated sequence from time $t-d$ to time t.

To perform the probabilistic tracking using the particle filtering approach, our final goal is to generate new particles at the prediction step, i.e. to draw samples ϕ_t^s from the dynamic model $p(\phi_t|\Phi_{t-1})$. Following the approach described by Sidenbladh in [7], we can rewrite this distribution as:

$$p(\phi_t|\Phi_{t-1}) = p(\phi_t|\Psi_{i-1})p(\Psi_{i-1}|\Phi_{t-1}), \tag{4}$$

where $p(\phi_t|\Psi_{i-1})$ is defined as 1 if $\phi_t = \psi_i$, or 0 otherwise.

Thus, sampling from the prior $p(\phi_t|\Phi_{t-1})$ corresponds to sampling from the distribution $p(\Psi_{i-1}|\Phi_{t-1})$.This can be seen as performing a probabilistic search of the estimated motion Φ_t with a stored sequence Ψ_i from the database. Assuming that sequences of estimated postures follow a Gaussian distribution around matching sequences on the database, i.e.:

$$\Psi_i = \Phi_t + \eta(\Delta_d), \tag{5}$$

the matching probability is given by

$$p(\Psi_i|\Phi_t) = k\, e^{-\frac{1}{2}(\Psi_i - \Phi_t)^T \Delta_d^{-1} (\Psi_i - \Phi_t)}, \tag{6}$$

where k is a normalizing factor.

The covariance matrix Δ_d is defined by calculating the covariance Δ of all the postures ψ_i from the database, and storing d copies of Δ along the diagonal of the $d \cdot b^U \times d \cdot b^U$ covariance matrix Δ_d. By doing this, we give the same importance to each posture when matching the sequences, see [7] for details.

Thus, the dynamic model will estimate feasible human postures for tracking by searching only for the most likely stored postures from the database, and adding an empirically determined Gaussian noise term to them. Since this is a probabilistic model, we can generate n new different particles ϕ_t^s at each time step by sampling n times from the distribution $p(\phi_t|\Phi_{t-1})$ defined using the learnt *p-actions* from the database.

5 Experimental Results

The dynamic model has been trained with 9 different basic actions (*aRun, aWalk, aBend, aSit, aJump, aSkip, aSquat, aTumble and aKick*) considering near 100 postures for each action, by sampling the parametric manifolds $\mathbf{f}^{A_i}(p)$ that represent each action A_i at a constant rate with a sampling step of 0.01, $p \in [0, 1]$.

The testing set consisted in 5 performances per action, each one performed by 9 different actors. This results in 45 performances of all the actions which were not included in the training set for the calculation of the *p-actions.*

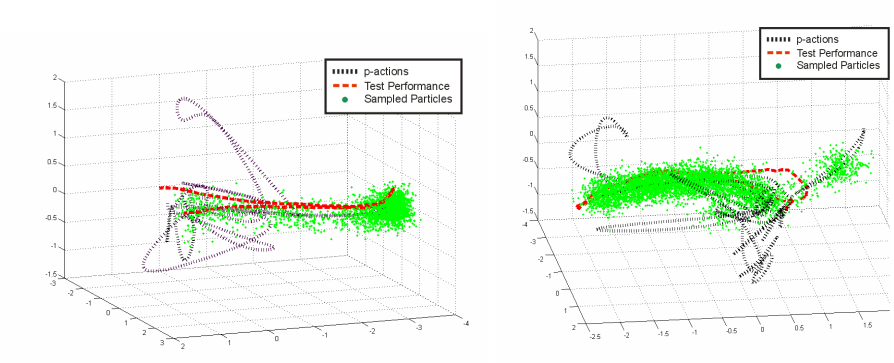
Table 1. Confusion Matrix in percentages.

Action	aRun	aWalk	aBend	aSit	aJump	aSkip	aSquat	aTumble	aKick
aRun	97	0	0	0	0	0	0	0	3
aWalk	0	72	0	1	1	23	1	1	1
aBend	0	9	83	2	3	0	1	1	1
aSit	0	21	3	65	0	4	4	3	0
aJump	8	3	0	1	70	7	1	1	8
aSkip	0	10	0	0	0	85	0	0	5
aSquat	5	0	4	0	1	0	90	0	0
aTumble	0	0	0	2	2	0	0	95	1
aKick	1	11	1	2	13	21	2	1	48

In order to explore the coverage of the search space performed by our dynamic model, we generated all the possible motion histories of length d ($d = 10$) for each test performance, and sampled 100 new postures or particles per each motion history following the procedure described above. After doing this for all the test performances, the confusion matrix shown in Table 1 was generated, where each row indicates the class, or *p-action* of the tested subsequence, and each column corresponds to the class of the sampled particle using a minimum Mahalanobis distance criteria.

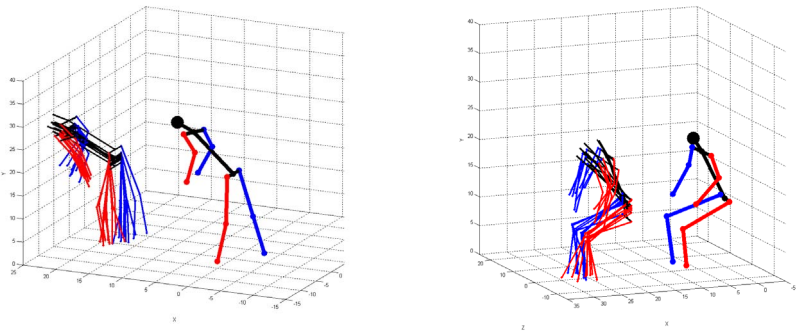
This table shows that our predictions are not too focused on an specific action, but still cover the truly performed action well enough. These results reflect the fact that some actions share a lot of similar postures between each other, especially at the beginning and at the end of the performances. This situation is very well handled by our dynamic model, since it is able to throw multiple hypotheses when the given subsequence is very similar in several actions, so we do not restrict the searching space to any of them. These hypotheses will be propagated over time by the particle filter until some of them become very unlikely over time. For instance, looking at Table 1, we observe that the action of *aWalk* has a lot of similarities with the action of *aSkip*. In the a *aSkip* action a subject starts walking, and after some frames it passes over some obstacle.Thus, the two actions share a lot of postures, especially at the beginning and at the end. Therefore, multiple hypotheses on what is the agent doing must be thrown on that situations, which is fulfilled by our dynamic model. We can find a similar situation between the *aBend* and the *aWalk* actions, and between the *aJump*, *aRun* and *aKick*. The table also shows that most of the actions only share a few postures, or none at all. So, this result is useful for establishing relationships between the involved actions. Further study needs to be done in order to determine similarities between parts of the same action, and not the action as a whole, in order to analyse the predictions made by the dynamic model.

In Fig 1.(a) the first 3 dimensions of the *UaSpace* are drawn together with a *aBend* test performance (dashed line). We have generated particles up to the middle of the performance by our dynamic model and plotted them on the *UaSpace* as single dots. We can observe that the predictions made at the



(a) (b)

Fig. 1. Sampling from the dynamic model within the *UaSpace*. See text for details.



(a) (b)

Fig. 2. Predicted human postures for the *aBend* and aSit actions. See text for details.

beginning of the action are split mainly between the bending and other actions such as *aWalk*, *aJump* and *aSit*. But, as the performance goes over time, almost all the predictions are concentrated along the bending *p-action*, since it becomes very different to the other actions. A similar situation for a *aSit* test performance is shown in Fig.1(b).

In Fig 2.(a) and 2.(b) we show the true posture on the right and a set of predicted postures on the left for a particular frame of the same *aBend* and *aSit* performances used in Fig 1. The set shown is randomly selected from the 100 predicted postures. The results obtained point out that this dynamic model is focused on generating the most suitable postures while performing an action, and naturally reduces the searching space avoiding the evaluation of improbable and impossible body configurations.

6 Conclusions and Future Work

This paper presents a temporal prior distribution suitable to be used as a dynamic human body model for tracking. The drawn particles from this distribution correspond to predicted feasible poses of the body given the history of estimated poses over time. The method learns a human motion model from a database of 3D actions acquired with a commercial Motion Capture System.

The results point out that this procedure, if used in a particle filtering framework, will drastically reduce the number of particles needed to track a human body while performing an action. Even though the proposed example-based dynamic model is less flexible than generic models for articulated objects motion, it is generic and accurate enough for making the tracking of human motion an achievable task.

Future research relies on integrating this approach into a particle filtering framework and developing appropriate likelihood measures for human bodies in 2D images. To reduce the problems of extrapolating from the *p-action* model, a more refined action model could be developed by probabilistically modeling each action using Mixtures of Gaussians, for example. Furthermore, transitions between actions could be naturally modelled by interpolating between the keyframes of several *p-actions*. Another open issue is the high computational cost of the probabilistic search, which could be addressed by efficient indexing the motion database.

Acknowledgments

This work has been supported by the Spanish CICYT TIC2003-08865.

References

1. M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, February 2002.
2. Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and data association*. Academic Press, 1988.
3. Jordi Gonzàlez. *Human Sequence Evaluation: the Key-frame Approach*. PhD thesis, Universitat Autònoma de Barcelona, May 2004.
4. Jordi Gonzàlez, Javier Varona, F. Xavier Roca, and J. José Villanueva. Analysis of human walking based on aSpaces. *3rd International Workshop on Articulated Motion and Deformable Objects (AMDO'2004)*, September 2004.
5. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
6. O. King and David A. Forsyth. How does CONDENSATION behave with a finite number of samples? In *ECCV (1)*, pages 695–709, 2000.
7. Hedvig Sidenbladh, Michael J. Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV (1)*, pages 784–800, 2002.
8. L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.