Tracking Articulated Motion using a Mixture of Autoregressive Models

Ankur Agarwal, Bill Triggs

Tracking Articulated Motion with Piecewise Learned Dynamical Models

Ankur Agarwal and Bill Triggs

GRAVIR-INRIA-CNRS, 655 Avenue de l'Europe, Montbonnot 38330, France
{Ankur.Agarwal,Bill.Triggs}@inrialpes.fr
http://www.inrialpes.fr/lear/{agarwal,triggs}

Abstract. We present a novel approach to modelling the non-linear and time-varying dynamics of human motion, using statistical methods to capture the characteristic motion patterns that exist in typical human activities. Our method is based on automatically clustering the body pose space into connected regions exhibiting similar dynamical characteristics, modelling the dynamics in each region as a Gaussian autoregressive process. Activities that would require large numbers of exemplars in example based methods are covered by comparatively few motion models. Different regions correspond roughly to different action-fragments and our class inference scheme allows for smooth transitions between these, thus making it useful for activity recognition tasks. The method is used to track activities including walking, running, *etc.*, using a planar 2D body model. Its effectiveness is demonstrated by its success in tracking complicated motions like turns, without any key frames or 3D information.

1. Introduction

Tracking and analyzing human motion in video sequences is a key requirement in several applications. There are two main levels of analysis: *(i)* detecting people and tracking their image locations; and *(ii)* estimating their detailed body pose, *e.g.* for motion capture, action recognition or human-machine-interaction. The two levels interact, as accurate detection and tracking requires prior knowledge of pose and appearance, and pose estimation requires reliable tracking. Using an explicit body model allows the state of the tracker to be represented as a vector of interpretable pose parameters, but the problem is non-trivial owing to the great flexibility of the human body, which requires the modelling of many degrees of freedom, and the frequent non-observability of many of these degrees of freedom in monocular sequences owing to self-occlusions and depth ambiguities. In fact, if full 3D pose is required from monocular images, there are potentially thousands of local minima owing to kinematic flipping ambiguities [18]. Even without this, pervasive image ambiguities, shadows and loose clothing add to the difficulties.

Previous work: Human body motion work divides roughly into *tracking based approaches*, which involve propagating the pose estimate from one time step to another,

To appear in the 2004 European Conference on Computer Vision.
This research was supported by European Union research projects VIBES and LAVA.

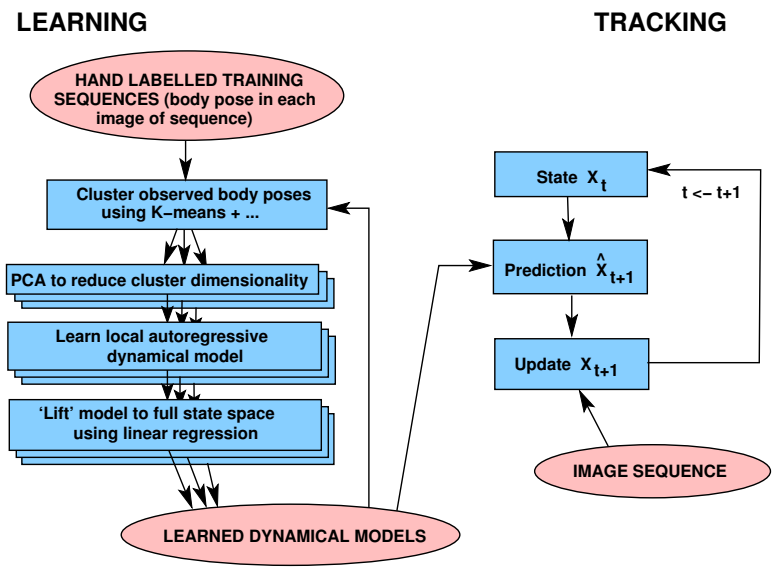


Fig. 1. Overview of the learning and tracking components of our algorithm (see text).

and *detection based approaches*, which estimate pose from the current image(s) alone. The latter have become popular recently in the form of 'exemplars' [21] and 'key frames' [19]. These methods allow the direct use of image data, which eliminates the need for predefined parametric models. But the interpretability of parametric models is lost, and large numbers of exemplars are needed to cover high dimensional example spaces such as those of human poses. (Tree-based structures have recently been explored for organizing these datasets [20], but they rely on the existence of accurate distance metrics in the appearance space).

Within the tracking framework, many methods are based on computing optical flow [9, 3, 2], while others optimize over static images (*e.g.* [18]). On the representation side, a variety of 2D and 3D parametric models have been used [9, 3, 16, 18], as well as non-parametric representations based on motion [4] or appearance [15, 11, 21]. A few learning based methods have modelled dynamics [8, 17, 14], motion patterns from motion capture data (*e.g.* [1]), and image features [16, 7, 6]. To track body pose, Howe *et al* [8] and Sidenbladh *et al* [17] propose plausible next states by recovering similar training examples, while Pavlovic *et al* [14] learn a weak dynamical model over a simplified 8-parameter body for fronto-parallel motions. We extend the learning based approach by modelling complex high dimensional motions within reduced manifolds in an unsupervised setting. In the past, nonlinear motion models have been created by combining Hidden Markov Models and Linear Dynamical Systems in the multi-class dynamics framework, *e.g.* in [13, 14]. However, this approach artificially decouples the switching dynamics from the continuous dynamics. We propose a simpler alternative that avoids this decoupling, discussing our philosophy in section 3.4.

Problem formulation: We use a tracking based approach, representing human motions in terms of a fixed parametric body model controlled by pose-related parameters,

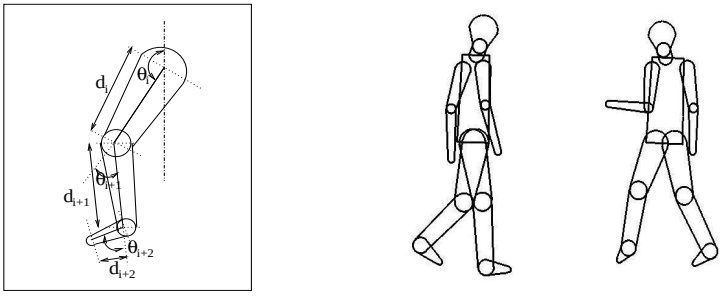


Fig. 2. *(a)* Human pose parametrization in the Scaled Prismatic Model. *(b)* Examples of different poses of the complete SPM. Each limb segment is overlayed with its corresponding template shape.

and focusing on flexible methods for learning the human dynamics. We specialize to monocular sequences using a 2D (image based) body model, but our methods extend immediately to the 3D and multicamera cases. Our main aim is to study how relationships and constraints in parameter space can be learned automatically from sample trajectories, and how this information can be exploited for tracking. Issues to be handled include the 'curse of dimensionality', complex nonlinear motions, and transitions between different parts of the space.

Overview of approach: Our approach is based on learning dynamical models from sample trajectories. We learn a collection of local motion models (Gaussian autoregressive processes) by automatically partitioning the parameter space into regions with similar dynamical characteristics. The piecewise dynamical model is built from a set of hand-labelled training sequences as follows: *(i)* the state vectors are clustered using K-means and projected to a lower dimensional space using PCA to stabilize the subsequent estimation process; *(ii)* a local linear autoregression for the state given the p previous reduced states is learned for each cluster (p = 1,2 in practice); *(iii)* the data is reclustered using a criterion that takes into account the accuracy of the local model for the given point, as well as the spatial contiguity of points in each model; *(iv)* the models are refitted to the new clusters, and the process is iterated to convergence.

We sidestep the difficult depth estimation problem by using a purely 2D approach, so our dynamical models are view dependent. Our tracking framework is similar to Covariance Scaled Sampling [18]: well-shaped random sampling followed by local optimization of image likelihood. Figure 1 illustrates the basic scheme of dividing the problem into learning and tracking stages.

2. Body representation

We choose a simple representation for the human body: a modified Scaled Prismatic Model [12] that encodes the body as a set of 2D chains of articulated limb segments. This avoids 3D ambiguities while still capturing the natural degrees of freedom. Body parts are represented by rounded trapezoidal image templates defined by their end

widths, and body poses are parametrized by their joint angles and apparent (projected) limb lengths. Including limb lengths, joint angles and hip and shoulder positions, our model contains 33 parameters, giving 33-D state vectors $\mathbf{x} = (\theta_1, d_1, \theta_2, d_2, \ldots \theta_n, d_n)$. Figure 2 illustrates the parametrization and shows some sample poses.

Three additional parameters are used during tracking, two for the image location of the body centre and one for overall scale. We learn scale and translation independently of limb movements, so these parameters are not part of the learned body model. The template for each body part contains texture information used for model-image matching. Its width parameters depend on the subject's clothing and physique. They are defined during initialization and afterwards remain fixed relative to the overall body scale, which is actively tracked.

3. Dynamical Model Formulation

Human motion is both complex and time-varying. It is not tractable to build an exact analytical model for it, but approximate models based on statistical methods are a potential substitute. Such models involve learning characteristic motions from example trajectories in parameter space. Our model learns the nonlinear dynamics by partitioning the parameter space into distinct regions or motion classes, and learning a linear autoregressive process covering each region.

3.1 Partitioning of State Space

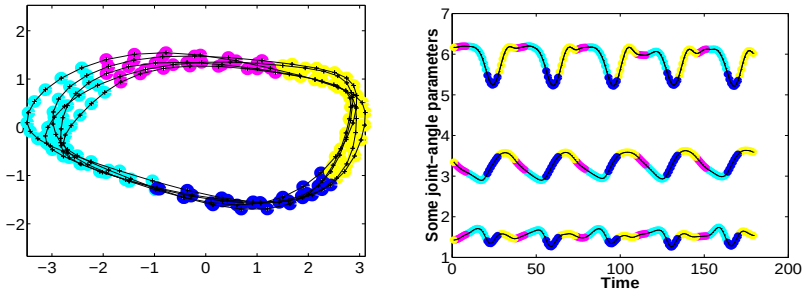
In cases where the dynamics of a time series changes with time, a single model is often inadequate to describe the evolution in state space. To get around this, we partition the state space into regions containing separate models that describe distinct motion patterns. The partitions must satisfy two main criteria: *(i)* different motion patterns must belong to different regions; and *(ii)* regions should be contiguous in state space. *I.e.*, we need to break the state space into *contiguous regions* with *coherent dynamics*. Coherency means that the chosen dynamical model is locally accurate, contiguity that it can be reliably deduced from the current state space position. Different walking or running styles, viewpoints, *etc.*, tend to use separate regions of state space and hence separate sets of partitions, allowing us to infer pose or action from class information.

We perform an initial partitioning on unstructured input points in state space by using K-means on Mahalanobis distances (see fig. 3). The clusters are found to cut the state trajectories into short sections, all sections in a given partition having similar dynamics. The partition is then refined to improve the accuracies of the nearby dynamical models. The local model estimation and dynamics based partition refinement are iterated in an EM-like loop, details of which are given in section 3.3.

3.2 Modelling the Local Dynamics

Despite the complexity of human dynamics and the use of unphysical image-based models, we find that the local dynamics within each region is usually well described by a linear Auto-Regressive Process (ARP):

$$\mathbf{x}_t = \sum_{i=1}^{p} \mathbf{A}_i \, \mathbf{x}_{t-i} + \mathbf{w}_t + \mathbf{v}_t \tag{1}$$



(a) (b)

Fig. 3. *(a)* The initial partition of the state space of a walking motion (5 cycles), projected to 2-D using PCA (see text). *(b)* The clusters correspond to different *phases* of the walking cycle, here illustrated using the variations of individual joint angles with time. (The cluster labels are coded by colour). These figures illustrate the optimal clustering obtained for a p=1 ARP. For p=2, a single class suffices for modelling unidirectional walking dynamics.

Here, $\mathbf{x}_t \in \mathbb{R}^m$ is the pose at time t (joint angles and link lengths), p is the model order (number of previous states used), $\mathbf{A}_i$ are $m \times m$ matrices giving the influence of $\mathbf{x}_{t-i}$ on $\mathbf{x}_t$, $\mathbf{w}_t \in \mathbb{R}^m$ is a drift/offset term, and $\mathbf{v}_t$ is a random noise vector (here assumed white and Gaussian, $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$).

The choice of ARP order is strongly dependent on the nature of the motions exhibited by the system. In practice, experiments on different kinds of motion showed that a second order ARP usually suffices for human tracking:

$$\mathbf{x}_t = \mathbf{A}_1 \, \mathbf{x}_{t-1} + \mathbf{A}_2 \, \mathbf{x}_{t-2} + \mathbf{v}_t \tag{2}$$

This models the local motion as a mass-spring system (set of coupled damped harmonic oscillators). It can also be written in differential form: $\ddot{\mathbf{x}}_t = \mathbf{B}_1 \, \dot{\mathbf{x}}_t + \mathbf{B}_2 \, \mathbf{x}_t + \mathbf{v}_t$.

3.3 Model Parameter Estimation

The parameters to be estimated are the state-space partitioning, here encoded by the class centers $\mathbf{c}^k$, and the ARP parameters $\{\mathbf{A}_1^k, \mathbf{A}_2^k, \ldots \mathbf{A}_p^k, \mathbf{Q}^k\}$ within each class ($k = 1 \ldots K$). There are standard ways of learning ARP models from training data [10]. We computed maximum likelihood parameter estimates. We also wanted to take advantage of the well-structured nature of human motion. People rarely move their limbs completely independently of one another, although the actual degree of correlation depends on the activity being performed. This can be exploited by learning the dynamics with respect to a *reduced set of degrees of freedom* within each class, *i.e.* locally projecting the system trajectories into a lower dimensional subspace. Thus, within each partition, we:

1. reduce the dimensionality using linear PCA (in practice to about 5);
2. learn an ARP model in the reduced space;
3. "lift" this model to the full state space using the PCA injection;
4. cross-validate the resulting model to choose the PCA dimension.

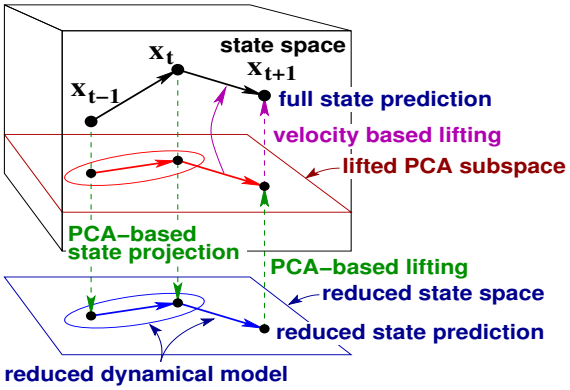


Fig. 4. Using a reduced dynamical model to predict states in a high-dimensional space. A given state is projected onto a low-dimensional space using PCA, within which a linear autoregressive progress is used to predict a current (reduced) state. This is then lifted back into full state space to estimate a noise model in the high-dimensional space. To prevent the state from being continually squashed into the PCA subspace, we lift the velocity prediction and not the state prediction.

The basic scheme is illustrated in figure 4, and the complete algorithm is given below. Before applying PCA, the state-space dimensions need to be statistically normalized. This is done by dividing each dimension by its observed variance over the complete set of training data.

Algorithm for estimation of maximum-likelihood parameters:

1. Initialize the state-space partitions by K-means clustering based on scaled (diagonal Mahalanobis) distance.
2. Learn an autoregressive model within each partition.
3. Re-partition the input points to minimize the dynamical model prediction error. If the class assignments have converged, stop. Otherwise go to step 2.

Step 2 above is performed as follows:

1. Reduce the vectors in the class to a lower dimensional space by:
   (a) Centering them and assembling them into a matrix (by columns): $\mathbf{X} = [\,(\mathbf{x}_{p_1} - \mathbf{c})\ (\mathbf{x}_{p_2} - \mathbf{c})\ \cdots\ (\mathbf{x}_{p_m} - \mathbf{c})\,]$, where $p_1 \ldots p_m$ are the indices of the points in the class and $\mathbf{c}$ is the class mean.
   (b) Performing a Singular Value Decomposition of the matrix to project out the dominant directions: $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$.
   (c) Projecting each vector into the dominant subspace: each $\mathbf{x}_i \in \mathbb{R}^m$ is represented as a reduced vector $\mathbf{q}_i = \tilde{\mathbf{U}}^T(\mathbf{x}_i - \mathbf{c})$ in $\mathbb{R}^{m'}$ ($m' < m$), where $\tilde{\mathbf{U}}$ is the matrix consisting of the first m' columns of $\mathbf{U}$.
2. Build an autoregressive model, $\hat{\mathbf{q}} = \sum_{i=1}^{p} \mathbf{A}_i\,\mathbf{q}_{t-i}$, and estimate $\mathbf{A}_i$ by writing this in the form of a linear regression:

$$\mathbf{q}_t = \tilde{\mathbf{A}}\,\tilde{\mathbf{q}}_{t-1}, \quad t = t_{p_1}, t_{p_2}, \ldots t_{p_n} \tag{3}$$

where

$$\tilde{\mathbf{A}} = (\,\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_p\,), \quad \tilde{\mathbf{q}}_{t-1} = \begin{pmatrix} \mathbf{q}_{t-1} \\ \mathbf{q}_{t-2} \\ \vdots \\ \mathbf{q}_{t-p} \end{pmatrix}$$

3. Estimate the error covariance $\mathbf{Q}$ from the residual between $\{\hat{\mathbf{x}}_i\}$ and $\{\mathbf{x}_i\}$ by "lifting" $\hat{\mathbf{q}}_t$ back into m dimensions:

$$\hat{\mathbf{x}}_t = \mathbf{c} + \tilde{\mathbf{U}}\hat{\mathbf{q}}_t \tag{4}$$

Step 3 above is performed as follows: The K-means based partitions are revised by assigning training points to the dynamical model that predicts their true motion best, and the dynamical models are then re-learned over their new training points. This EM / relaxation procedure is iterated to convergence. In practice, using dynamical prediction error as the sole fitting criterion gives erratic results, as models sometimes "capture" quite distant points. So we include a spatial smoothing term by minimizing:

$$\sum_{\text{training points}} (\text{prediction error}) + \lambda \cdot (\text{number of inter-class neighbors})$$

where λ is a relative weighting term, and the number of inter-class neighbors is the number of edges in a neighborhood graph that have their two vertices in different classes (*i.e.*, a measure of the lack of contiguity of a partition).

3.4 Inter-Class Transitions

Many example-based trackers use discrete state HMMs (transition probability matrices) to model inter-cluster transitions [21, 20]. This is unavoidable when there is no state space model at all (*e.g.* in exemplars [21]), and it is also effective when modelling time series that are known to be well approximated by a set of piecewise linear regimes [5]. Its use has been extended to multi-class linear dynamical systems exhibiting continuous behavior [14], but we believe that this is unwise, as the discrete transitions ignore the location-within-partition information encoded by the continuous state, which strongly influences inter-class transition probabilities. To work around this, quite small regions have to be used, which breaks up the natural structure of the dynamics and greatly inflates the number of parameters to be learned. In fact, in modelling human motion, the current continuous state already contains a great deal of information about the likely future evolution, and we have found that this alone is rich enough to characterize human motion classes, without the need for the separate hidden discrete state labels of HMM based models.

We thus prefer the simpler approach of using a piecewise linear dynamical model over an explicit spatial partition, where the 'class' label is just the current partition cell. More precisely, we use soft partition assignments obtained from a Gaussian mixture model based at the class centres, so the dynamics for each point is a weighted random mixture over the models of nearby partitions. Our classes cover relatively large regions of state space, but transitions typically only occur at certain (boundary) areas

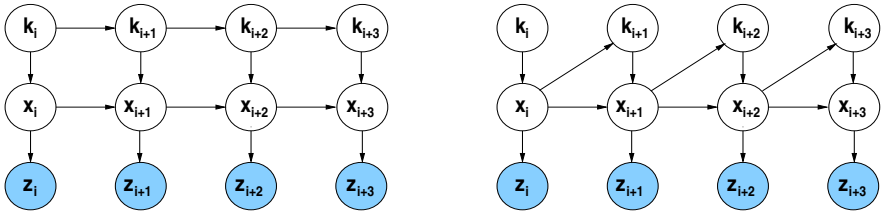


Fig. 5. Graphical models for inter-class transitions of a system. *(a)* An HMM-like mixed-state model, and *(b)* our inter-class transition model ($\mathbf{z}_i$: observation, $\mathbf{x}_i$: continuous state, k_i: discrete class). Transitions in an HMM are learned as a fixed transition probability matrix, while our model allows location-sensitive estimation of the class label by exploiting continuous state information.

within them. Constant transition probabilities given the current class label would thus be inappropriate in our case.

Figure 5 compares the two schemes in graphical form. By modelling the class-label to be conditional on continuous state, we ensure a smooth flow from one model to the next, avoiding erratic jumps between classes, and we obviate the need for complex inference over a hidden class-label variable.

4. Image Matching Likelihood

At present, for the model-image matching likelihood we simply use the weighted sum-of-squares error of the backwards-warped image against body-part reference templates fixed during initialization. Occlusions are handled using support maps. Each body part P has an associated support map whose j^{th} entry gives the probability that image pixel j currently 'sees' this part. Currently, we use hard assignments, $p(j \text{ sees } P) \in \{0, 1\}$. To resolve the visibility ambiguity when two limbs overlap spatially, each pose has an associated *limb-ordering*, which is known a priori for different regions in the pose space from the training data. This information is used to identify occluded pixels that do not contribute to the image matching likelihood for the pose. We charge a fixed penalty for each such pixel, equal to the mean per-pixel error of the visible points in that segment. Some sample support maps are shown in figure 8(b).

5. Tracking Framework

Our tracking framework is similar to Covariance Scaled Sampling [18]. For each mode of $\mathbf{x}_{t-1}$, the distribution $\mathcal{N}(\hat{\mathbf{x}}_t, \mathbf{Q})$ estimated by the dynamical model (1,5) is sampled, and the image likelihood is locally optimized at each mode. State probabilities are propagated over time using Bayes' rule. The probability of the tracker being in state (pose) $\mathbf{x}_t$ at time t given the sequence of observations $\mathcal{Z}_t = \{\mathbf{z}_t, \mathbf{z}_{t-1} \ldots \mathbf{z}_0\}$ is:

$$p(\mathbf{x}_t \,|\, \mathcal{Z}_t) \;=\; p(\mathbf{x}_t \,|\, \mathbf{z}_t, \mathcal{Z}_{t-1}) \;\propto\; p(\mathbf{z}_t \,|\, \mathbf{x}_t)\, p(\mathbf{x}_t \,|\, \mathcal{Z}_{t-1})$$

where $\mathcal{X}_t$ is the sequence of poses $\{\mathbf{x}_i\}$ up to time t and

$$p(\mathbf{x}_t \,|\, \mathcal{Z}_{t-1}) \;=\; \int p(\mathbf{x}_t \,|\, \mathcal{X}_{t-1})\, p(\mathcal{X}_{t-1} \,|\, \mathcal{Z}_{t-1})\, d\mathcal{X}_{t-1} \qquad (5)$$

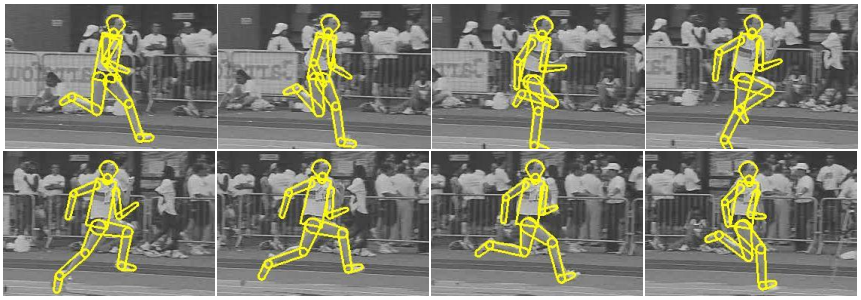


Fig. 6. Results from tracking athletic motion (frames 0,4,8,12,16,20,24). The tracker was trained on a different athlete performing a similar motion. Strong priors from the dynamical model allow individual limbs to be tracked in the presence of a confusing background. Note that the left arm is not tracked accurately. This is due to the fact that it was occluded in the initial image and hence no information about its appearance was captured in the template. However, the dynamics continue to give a good estimate of its position.

The likelihood $p(\mathbf{z}_t \mid \mathbf{x}_t)$ of observing image $\mathbf{z}_t$ given model pose $\mathbf{x}_t$ is computed based on the image-model matching error. The temporal prior $P(\mathbf{x}_t \mid \mathcal{X}_{t-1})$ is computed from the learned dynamics. In our piecewise model, the choice of discrete class label k_t is determined by the current region in state space, which in our current implementation depends only on the previous pose $\mathbf{x}_{t-1}$, enabling us to express the probability as

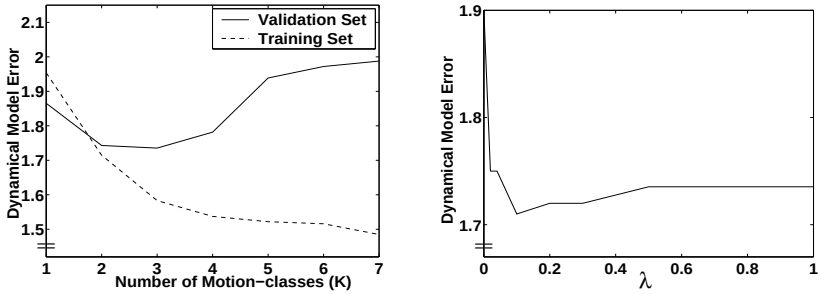
$$p(\mathbf{x}_t \mid \mathcal{X}_{t-1}) \;=\; p(\mathbf{x}_t \mid \mathcal{X}_{t-1}, k_t)\, p(k_t \mid \mathbf{x}_{t-1}) \tag{6}$$

The size and contiguity of our dynamical regions implies that $p(k_t \mid \mathbf{x}_{t-1})$ is usually highly unimodal. The number of modes increases when the state lies close to the boundary between two or more regions, but in this case, the spatial coherence inherited from the training dynamics usually ensures that any of the corresponding models can be used successfully, so the number of distinct modes being tracked does not tend to increase exponentially with time. For each model $k = 1 \ldots K$, we use a Gaussian posterior for $p(k|\mathbf{x}_t)$: $p(k \mid \mathbf{x}_t) \propto e^{-((\mathbf{x}_t - \mathbf{c}_k)\Sigma^{-1}(\mathbf{x}_t - \mathbf{c}_k))/2}$ where $\mathbf{c}_k$ is the center of the k^{th} class. Note that with a second order ARP model, $p(\mathbf{x}_t \mid \mathcal{X}_{t-1}) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_{t-2})$.

6. Results

We demonstrate our technique by learning models for different classes of human motion and using them to track complete body movements in unseen video sequences. Here, we present results from two challenging sequences.

1. Fast athletic motion: This is a case where traditional methods typically fail due to high motion blur. A hand-labelled sequence covering a few running cycles is used to train a model and this is used to track a different person performing a similar motion. For a given viewing direction, we find that a single 2nd order autoregressive process in 5 dimensions suffices to capture the dynamics of such running motions. A tracking example is shown in figure 6.



(a) (b)

Fig. 7. *(a)* Dynamical model prediction error w.r.t. number of motion-classes in the turning experiment. Minimizing the validation error selected 3 classes, corresponding to the two walking directions and turning between them. *(b)* The influence of spatial regularization when re-partitioning the state space. A weak regularization $\lambda \sim 0.1$ gives the optimal dynamical estimates. A larger λ causes the partition to remain too close to the suboptimal initial K-means estimate.

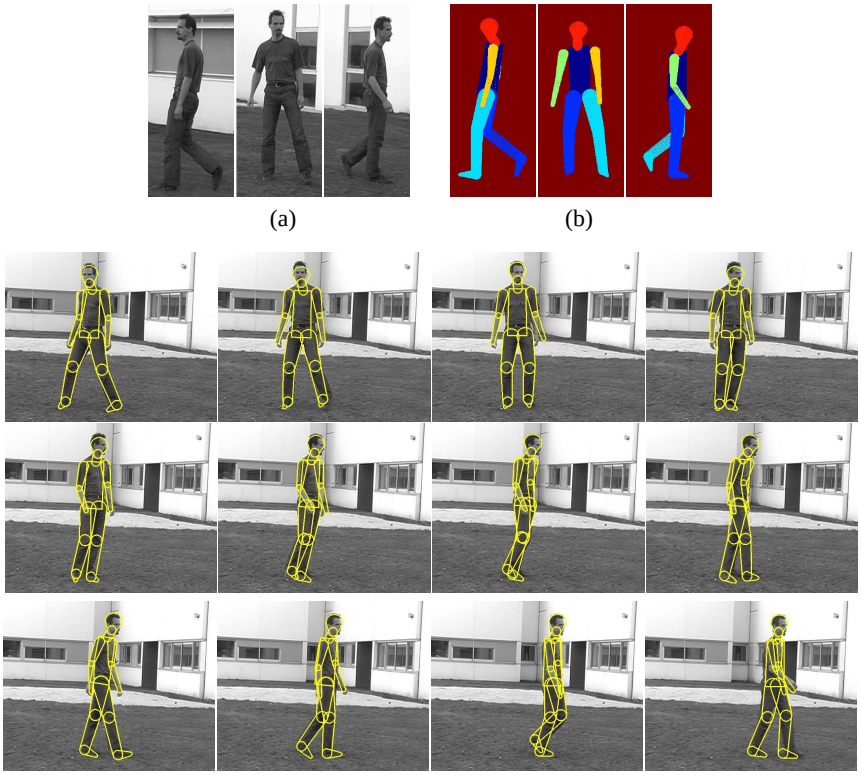
2. Switching between turning and walking: This experiment illustrates the effectiveness of our inter-class transition model. A 300-frame sequence consisting of walking in different directions and turning motion is used as training data. Our learning algorithm correctly identifies 3 motion patterns (see figure 7(a)), corresponding to two different walking directions and turning between them. The frames corresponding to the centers of these 3 classes are shown in figure 8(a). While tracking a new sequence, the model correctly shifts between different classes enabling smooth switching between activities. Figure 8(c) shows complete tracking results on an unseen test sequence.

In both cases, the models were initialized manually (we are currently working on automatic initialization), after which only the learned dynamics and appearance information were used for tracking. Position and scale changes were modelled respectively as first and zeroth order random walks and learned online during tracking. This allows us to track sequences without assuming either static or fixating cameras, as is done in several other works. The dynamical model alone gives fairly accurate pose predictions for at least a few frames, but the absence of clear observations for any longer than this may cause mistracking.

Figure 7(b) shows how repartitioning (step 3 of our parameter estimation algorithm) improves on the initial K-means based model, provided that a weak smoothing term is included.

7. Conclusion

We have discussed a novel approach to modelling dynamics of high degree-of-freedom systems such as the human body. Our approach is a step towards describing dynamical behavior of high-dimensional parametric model spaces without having to store extremely large amounts of training data. It takes advantage of local correlations between motion parameters by partitioning the space into contiguous regions and learning individual local dynamical behavior within reduced dimensional manifolds. The approach was tested on several different human motion sequences with good results,



(a) (b)

(c)

Fig. 8. Examples from our turning experiment. *(a)* Poses characterizing the 3 motion classes learned. *(b)* Support maps illustrating occlusion information for the 3 classes (color coded by body part). *(c)* Tracking results (every 6th frame from 0–66). The corresponding state vectors show a smooth transition between the turning and walking models.

and allows the tracking of complex unseen motions in the presence of image ambiguities. The piecewise learning scheme developed here is practically effective, and scalable in the sense that it allows models for different actions to be built independently and then stitched together to cover the complete 'activity space'. The learning process can also be made interactive to allow annotation of different classes for activity recognition purposes.

In terms of future work, the appearance model needs to be improved. Adding detectors for characteristic human features and allowing the appearance to evolve with time would help to make the tracker more robust and more general. Including a wider range of training data would allow the tracker to cover more types of human motions.

An open question is whether non-parametric models could usefully be incorporated to aid tracking. Joint angles are a useful output, and are probably also the most appropriate representation for dynamical modelling. But it might be more robust to use

comparison with real images, rather than comparison with an idealized model, to compute likelihoods for joint-based pose tracking.

Acknowledgements

This work was supported by the European Union FET-Open Project VIBES.

References

1. Matthew Brand and Aaron Hertzmann. Style Machines. In *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192, 2000.
2. C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *International Conference on Computer Vision and Pattern Recognition*, pages 8–15, 1998.
3. T. Cham and J. Rehg. A Multiple Hypothesis Approach to Figure Tracking. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 239–245, 1999.
4. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *International Conference on Computer Vision*, 2003. To appear.
5. Z. Ghahramani and G. Hinton. Switching State-Space Models. Technical report, Department of Computer Science, University of Toronto, Canada, 1998.
6. Tony Heap and David Hogg. Nonlinear Manifold Learning for Visual Speech Recognition. In *International Conference on Computer Vision*, pages 494–499, 1995.
7. Tony Heap and David Hogg. Wormholes in Shape Space: Tracking Through Discontinuous Changes in Shape. In *International Conference on Computer Vision*, pages 344–349, 1998.
8. N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. In *Neural Information Processing Systems*, 1999.
9. S. Ju, M. Black, and Y. Yacoob. Cardboard People: A Parameterized Model of Articulated Motion. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 38–44, 1996.
10. H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, Germany, second edition, 1993.
11. G. Mori and J. Malik. Estimating Human Body Configurations Using Shape Context Matching. In *European Conference on Computer Vision*, volume 3, pages 666–680, 2002.
12. D. Morris and J. Rehg. Singularity Analysis for Articulated Object Tracking. In *International Conference on Computer Vision and Pattern Recognition*, pages 289–296, 1998.
13. B. North, A. Blake, M. Isard, and J. Rittscher. Learning and Classification of Complex Dynamics. *Pattern Analysis and Machine Intelligence*, 22(9):1016–1034, 2000.
14. V. Pavlovic, J. Rehg, and J. MacCormick. Learning Switching Linear Models of Human Motion. In *Neural Information Processing Systems*, pages 981–987, 2000.
15. D. Ramanan and D. Forsyth. Finding and Tracking People from the Bottom Up. In *International Conference on Computer Vision and Pattern Recognition*, 2003.
16. H. Sidenbladh and M. Black. Learning Image Statistics for Bayesian Tracking. In *International Conference on Computer Vision*, volume 2, pages 709–716, 2001.
17. H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, volume 1, 2002.
18. C. Sminchisescu and B. Triggs. Covariance Scaled Sampling for Monocular 3D Body Tracking. In *International Conference on Computer Vision and Pattern Recognition*, 2001.
19. J. Sullivan and S. Carlsson. Recognizing and Tracking Human Action. In *European Conference on Computer Vision*, 2002.
20. A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Learning a kinematic prior for tree-based filtering. In *Proc. British Machine Vision Conference*, volume 2, pages 589–598, 2003.
21. K. Toyama and A. Blake. Probabilistic Tracking in a Metric Space. In *International Conference on Computer Vision*, pages 50–59, 2001.