Tracking a Walking Person using Activity-Guided Annealed Particle Filtering

John Darby, Baihua Li and Nicholas Costen
Department of Computing and Mathematics, Manchester Metropolitan University
John Dalton Building, Chester Street, Manchester, M1 5GD, UK.
{j.darby,b.li,n.costen}@mmu.ac.uk

Abstract

Tracking human pose using observations from less than three cameras is a challenging task due to ambiguity in the available image evidence. This work presents a method for tracking using a pre-trained model of activity to guide sampling within an Annealed Particle Filtering framework. The approach is an example of model-based analysis-by-synthesis and is capable of robust tracking from less than 3 cameras with reduced numbers of samples. We test the scheme on a common dataset containing ground truth motion capture data and compare against quantitative results for standard Annealed Particle Filtering. We find lower absolute and relative error scores for both monocular and 2-camera sequences using 80% fewer particles.

1. Introduction

Research into the markerless tracking of human motion has recently benefitted from the introduction of common data sets that include ground truth motion capture (MoCap) data [3, 10]. These have allowed for the quantitative evaluation and cross-comparison of tracking approaches. Annealed Particle Filtering (APF) [4] and Sampling Importance Resampling (SIR) [1] have been shown to recover pose from multiple cameras using silhouette and edge cues [3]. However, both approaches fail when limited to observations from only 1 or 2 cameras. Many distinct pose hypotheses may agree well with available image evidence. Despite large particle numbers, ambiguous evidence causes tracking to fail.

Even simple models of human pose possess a relatively high number of degrees of freedom (DOF). The resulting feature space is computationally prohibitive to explore, requiring large numbers of particles. We therefore appeal to the idea that human motion is well described by a low-dimensional subspace of the original feature space. We use Principal Components Analysis (PCA) to reduce the dimensionality of joint angle vectors recovered from MoCap training data. However, PCA recovers a space that is both linear and continuous, containing many illegal configurations and making it unsuitable for direct sampling.

The dimensionally reduced data constitutes a set of noisy observations of a stochastic process. We learn a temporal model and set of continuous observation density functions by training a hidden Markov model (HMM) from the distribution of data in the PCA space. Sampling guided by such an activity model produces poses close to the training data, with the recovered observation densities precluding the sampling of illegal regions. Sample propagation also benefits from a first order Markov model of dynamics rather than the addition of noise [2, 4]. Only small but pertinent portions of the original feature space are explored, which can be achieved with low particle numbers. In summary we make the following contributions:

- Use of an HMM to model a walking activity and simulate a nonlinear activity axis within a linear subspace recovered using PCA.
- Integration of the HMM into an Annealed Particle Filtering framework for particle propagation.
- Introduction of a temperature parameter into the HMM synthesis process, allowing the tracker to escape incorrect interpretations during ambiguous image evidence.
- Modification of the APF weighting function to allow for better estimation of global position in monocular sequences, thus aiding accurate pose recovery.

We test the HMM-guided APF (HMM-APF) approach on the walking dataset presented in [3] and evaluate against ground truth MoCap data. Known and unknown subject tracking is demonstrated from 2 cameras using 200 particles. Known subject tracking is also shown for a monocular sequence with 200 particles by using a modified weighting function. Both results represent a considerable improvement in accuracy over standard APF using 80% fewer particles.

978-1-4244-2154-1/08/$25.00 ©2008 IE

2. Related Work

Balan et al. [3] made a quantitative study of APF performance, finding an average lower bound on absolute joint location error of 41mm for a walking subject using 1000 particles and 3 cameras. However, tracking was found to fail for the 2-camera and monocular cases after approximately 1 second, despite equivalent particle numbers. Model-based analysis-by-synthesis approaches have proven successful in constraining the tracking problem given limited image evidence [6, 9, 11–13].

Sidenbladh et al. [9] used PCA to build a database of low dimensional walking, running, dancing, skipping and lifting models in order to propagate particles in a SIR tracking scheme. The authors showed how the database may be efficiently searched using the recently estimated pose history in order to obtain relevant future predictions over the range of activities. Successful tracking was demonstrated on a 30 frame monocular walking sequence.

Urtasun et al. [13] built separate models of walking and running activities by performing PCA on joint angles obtained from MoCap data. Tracking was achieved by minimising a differential objective function with respect to the first 5 coefficients of the principal components. They demonstrated tracking on a sequence of stereo data containing a transition from walking to running.

PCA is only capable of learning linear subspaces, limiting its effectiveness in modelling nonlinear human motion. More recent approaches to monocular tracking have adopted invertible nonlinear dimensionality reduction techniques such as the Gaussian Process Latent Variable Model (GPLVM) [5] to improve performance [6,12]. The Gaussian Process Dynamical Model (GPDM) due to Wang et al. [14] allows for the simultaneous learning of activity dynamics in addition to a nonlinear low-dimensional latent space [11].

We hope that the use of HMMs will provide a natural framework for the consideration of multiple activity models in multiple spaces during a single sequence, which is an aim of future work. Although an invertible nonlinear dimensionality reduction method could potentially be adopted within our framework, PCA is inexpensive versus techniques such as the GPLVM, as is the projection of new feature space data into subspaces previously recovered by PCA.

3. Training Data and Ground Truth

In this work, we use the synchronised video and MoCap dataset described in [3] in order to draw direct comparisons with standard APF performance. The dataset contains multi-camera synchronised video sequences of a walking subject taken at 60fps. Associated ground truth MoCap data allow the measurement of absolute error at the joint locations of pose hypotheses. Training MoCap data featuring the same subject walking is also available. We use MoCap data for 3 other walking subjects (S1, S2 and S3 from the HumanEva-I dataset [10]) to train the activity HMM for the tracking of an unknown subject.

4. Dimension Reduction by PCA

All M MoCap data frames available for the activity are converted into a set of joint angle vectors describing the corresponding configurations of a 31-DOF body model, $\Omega = \{\boldsymbol{\omega}_m | m = 1, ..., M\}$ where $\boldsymbol{\omega}_m = (\omega_m^1, ..., \omega_m^{31})^T$. The body model consists of a kinematic tree containing 10 truncated cones and the model pose is fully described by the 31 parameters in $\boldsymbol{\omega}_m$, comprising the position and orientation of the torso and the relative joint angles between limbs.

The 6 global translation and rotation elements of the body pose vector and the 3 head orientation elements, $\boldsymbol{\omega}_m^{'} = (\omega_m^1, ..., \omega_m^9)^T$, are removed from the training data as we do not want a subject's route or line of sight to form part of the generic activity class. Each subject's mean vector is then subtracted from each of their activity vectors, leaving a set of 22-D vectors $E = \{\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_M\}$. PCA is used to analyse each subject's departure from their mean pose vector (excluding pelvis and head parameters) during a particular activity sequence. The mean $\bar{\boldsymbol{\epsilon}}$ and covariance matrix $\mathbf{S}$ are calculated for the data and Single Value Decomposition used to find the eigenvectors, ϕ_i and eigenvalues, χ_i of $\mathbf{S}$. This allows for an estimate of any datapoint in the training set, $\boldsymbol{\epsilon}_m$, using

$$\boldsymbol{\epsilon}_m \approx \bar{\boldsymbol{\epsilon}} + \boldsymbol{\Phi}\mathbf{f}_m, \tag{1}$$

where $\boldsymbol{\Phi} = (\phi_1 | \phi_2 | ... | \phi_\eta)$ contains the first η eigenvectors corresponding to the largest eigenvalues, and the weighting vector is given by

$$\mathbf{f}_m = (f_m^1, ..., f_m^\eta)^T = \boldsymbol{\Phi}^T(\boldsymbol{\epsilon}_m - \bar{\boldsymbol{\epsilon}}). \tag{2}$$

In this way the training data sequence E is approximated by the set of feature vectors $F = \{\mathbf{f}_1, ..., \mathbf{f}_M\}$, which are plotted in Figure 1(a) for the 3 HumanEva-I subjects.

4.1. Feature Vector

The full feature vector used in this work can be viewed as a concatenation of free and learned body parameters. It is $(9 + \eta)$-dimensional and allows, using the approximation in Eq. (1), for the complete specification of the body model $\boldsymbol{\omega}_m$ at any time m,

$$\mathbf{x}_m = (\omega_m^1, ..., \omega_m^9, f_m^1, ..., f_m^\eta)^T = \boldsymbol{\omega}_m^{'} || \mathbf{f}_m. \tag{3}$$

5. Hidden Markov Models

A hidden Markov model can be used to model a time series of observations such as $F = \{\mathbf{f}_1, ..., \mathbf{f}_M\}$, derived in the

previous section. The approach is based upon the assumption that the underlying system is a Markov process, where the system's state at any timestep m is assumed to depend only on its state at $m-1$. A standard Markov model is described by a set of states and a set of transition probabilities between these states. The state of the system is allowed to evolve stochastically and is directly observable.

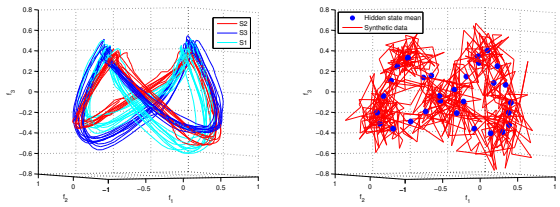


Figure 1. Training data and corresponding HMM, projected onto the 3 highest variance eigenvectors of the body configuration data E. 1(a) Several walking cycles for each of 3 subjects. 1(b) 30-state HMM trained from walking data and used to emit a synthetic sequence 500 feature vectors in length.

This approach may be extended with the introduction of a hidden layer between state and observer. Each state emits an observable symbol from an alphabet common to all states, according to some probability distribution over that alphabet. In our own application, both the human's performance of an intended activity and the measurement of that performance are stochastic processes. HMMs allow us to describe such a doubly stochastic system.

A HMM λ is specified by the parameters S,A_{ij},A_i,$p_i(\mathbf{f})$, where $S = \{s_1, ..., s_N\}$ is the set of states; the $N \times N$ matrix A_{ij} gives the probability of a transition from state i to state j; A_i gives the probability of a sequence starting in state i and $p_i(\mathbf{f})$ is the probability of observing feature vector $\mathbf{f}$ while in state i. In this work the emission probability is modelled by a single multivariate Gaussian $p_i(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. HMM training is performed using the Baum-Welch algorithm, for further detail the reader is referred to [8].

6. Annealed Particle Filtering

Human motion can be viewed as the evolution of a system state $\mathbf{x}_m$ over time, $m = 1, 2, ..., M$, described by a Markov process and observed by some sensor providing independent observations given $\mathbf{x}_m$. The state density $p_m(\mathbf{x}_m)$, given by $p(\mathbf{x}_m|\mathbf{Z}_m)$, where $\mathbf{Z}_m = (\mathbf{z}_1, ..., \mathbf{z}_m)$ is the set of all observations up until time m, may be propagated over time with the following rule:

$$p(\mathbf{x}_m|\mathbf{Z}_m) \propto p(\mathbf{z}_m|\mathbf{x}_m) \int_{\mathbf{x}_{m-1}} p(\mathbf{x}_m|\mathbf{x}_{m-1}) p(\mathbf{x}_{m-1}|\mathbf{Z}_{m-1}). \tag{4}$$

SIR [1] allows for the representation of a multimodal posterior, $p(\mathbf{x}_m|\mathbf{Z}_m)$ via a finite set of normalised, weighted particles,

$$\left\{ (\mathbf{x}_m^{(0)}, \pi_m^{(0)}), ..., (\mathbf{x}_m^{(B)}, \pi_m^{(B)}) \right\}. \tag{5}$$

Having initialised the particle set with ground truth $\mathbf{x}_0$, each with equal weight, particles are randomly sampled and dispersed using some model of temporal dynamics, $p(\mathbf{x}_m|\mathbf{x}_{m-1})$. Each new point in the feature space $\mathbf{x}_m^{(b)}$ is evaluated using a weighting function $w(\mathbf{z}_m, \mathbf{x}_m^{(b)})$ and assigned a weighting $\pi_m^{(b)}$ approximating $p(\mathbf{z}_m|\mathbf{x}_m^{(b)})$. Resampling then takes place, with B new particles randomly sampled from the existing distribution with likelihood proportional to their weighting (and with replacement) and then dispersed. In this way, the particle set may be propagated over time to maintain a representation of $p(\mathbf{x}_m|\mathbf{Z}_m)$.

Annealed particle filtering [4], a variation on SIR, cools the weighting distribution and then gradually introduces sharp peaks over $r = R, R-1, ..., 1$ separate resampling layers at each timestep m, where

$$w_m^r(\mathbf{z}_m, \mathbf{x}_m) = w(\mathbf{z}_m, \mathbf{x}_m)^{\beta_m^r}, \tag{6}$$

for $\beta_m^0 > \beta_m^1 > ... > \beta_m^R$. The value of β_m^r is chosen to control the particle survival rate τ_m^r, that is, the proportion of particles that will be resampled. A high survival rate results in an evenly spread probability distribution, while a low survival rate concentrates the probability distribution into just a few particles. We use the crossover operator and survival rate values proposed in [4]:

$$\tau_m^R = ... = \tau_m^1 = 0.5. \tag{7}$$

The effect of APF is to recover the pose that maximises the weighting function, leading to a gradual concentration of particles into a particular mode of the distribution. Thus the posterior distribution is not fully represented, constituting a departure from the formal Bayesian framework.

6.1. Weighting Function

The weighting function $w(\mathbf{z}_m, \mathbf{x}_m)$ provides an approximation of $p(\mathbf{z}_m|\mathbf{x}_m)$ using the geometric body model as specified by $\mathbf{x}_m$ [4].

The model is projected into the image and a set of points $\{\xi\}$ taken from each component cone, used to sample from image evidence. We calculate $w(\mathbf{z}_m, \mathbf{x}_m)$ for the case where image evidence, V, is: (i) a silhouette map and (ii) an edge map, and sum the results,

$$-\log p(\mathbf{z}_m|\mathbf{x}_m) \propto \frac{1}{W|\{\xi\}|} \sum_{\xi} (1 - V(\xi))^2. \tag{8}$$

In either case, $\{\xi\}$ is a set of sample points on (i) the cone surfaces and (ii) the cone edges, respectively. The edge data are extracted from each frame by convolution with a gradient based edge detection mask. Results are thresholded and smoothed with a Gaussian mask before being rescaled to the interval $[0, 1]$, giving each pixel a measure of proximity to an edge. Silhouette data are found using a foreground classifier trained on background only image sequences. A Gaussian mixture model of the background distribution is learned for each pixel over 1000 background images [3]. Particles describing a pose requiring the intersection of limbs are given a zero weighting as, in the case of known subject tracking, are those that exceed joint angle limits learned from the training data.

For monocular tracking, we make a modification to the weighting calculation. In the treatment of weightings above, there is no consideration of foreground in image evidence which is left unaccounted for by a pose hypothesis. An XOR-type comparison of foreground regions is desirable. In order to address this problem, we calculate weightings from silhouette evidence alone and take account of the disparity in the ratio of foreground to background pixels in image evidence, $g_{\mathbf{z}}$ and particle pose, $g_{\mathbf{x}}$ by calculating W as follows:

$$W = \left(1 - \text{abs}\left(\frac{g_{\mathbf{z}} - g_{\mathbf{x}}}{\max(g_{\mathbf{z}}, g_{\mathbf{x}})}\right)\right)^{\gamma}. \tag{9}$$

The requirement for silhouette sizes to match may be enforced by varying the exponent γ, with $\gamma = 0$ corresponding to standard APF weightings ($W = 1$ in Eq. (8)).

6.2. Temporal Dynamics

A drift-and-spread model of temporal dynamics $p(\mathbf{x}_m|\mathbf{x}_{m-1})$ is generally employed for particle propagation in SIR tracking schemes e.g. [2]. Quantitative evaluation in [3] found results using a presumption of constant angle velocity are worse than with a presumption of constant joint angles. Dispensing with the drift term leaves only a spreading function: the addition of noise. In standard APF, Gaussian noise is applied to each feature vector element x_m^i with covariance equal to half the maximum inter frame change in x_m^i, calculated from training data. This leads to the diagonal covariance matrix $\mathbf{P}$. The covariance matrix is multiplied by the particle survival rate τ_m^r at each annealing layer, the application of noise decreasing at the same rate as the resolution of the particle set increases [4]. In our approach the use of $\mathbf{P}_r$ is retained for the 9 free parameters (see Eq. (3)). Temporal dynamics for the remaining η learned body parameters are provided by the pre-trained HMM. We retain the idea of scaling covariance, multiplying the covariance matrices describing the Gaussian observation density at each state, $p_i(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, by τ_m^r at each annealing layer.

$$\mathbf{P}_r = (\tau_m^r)^{R-r} \times \mathbf{P}; \quad \boldsymbol{\Sigma}_i^r = (\tau_m^r)^{R-r} \times \boldsymbol{\Sigma}_i. \tag{10}$$

We also use a temperature parameter in the synthesis process, requiring a non-self state transition from state i with some minimum probability ρ_T. In general, hypotheses that extend too far along the activity cycle die out during annealing, discounted by comparison with image evidence in the evaluation of $w(\mathbf{z}_m, \mathbf{x}_m)$. However, where image evidence is weak, the particle set is able to escape becoming 'stuck' in an incorrect mode of the posterior distribution during the annealing process by sampling 'hot' future poses. The temperature ρ_T is also multiplied by the particle survival rate τ_m^r at each annealing layer,

$$\rho_T^r = (\tau_m^r)^{R-r} \times \rho_T. \tag{11}$$

We require K_i^r transitions from state i at layer r before the emission of an observable, where

$$K_i^r = \left\lceil \frac{\log(1 - \rho_T^r)}{\log(A_{ii})} \right\rceil. \tag{12}$$

For learned parameter reestimation in later annealing layers, self state transitions are more common as K_i^r becomes small. In the case where $s_j = s_i$ we draw noise using (a weighted version of) the parent state's covariance matrix $\boldsymbol{\Sigma}_j^r$ but replace $\boldsymbol{\mu}_j$ with the particle's current estimate of $\mathbf{f}_m^{(b)}$. This stops the training data from dominating the choice of new pose hypotheses, allowing the weighting function scores to guide refinement. In summary, each particle undergoes the following steps at each time step:

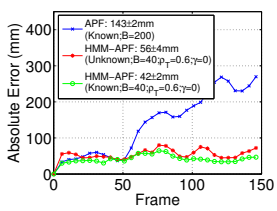
1. For a sampled particle $\mathbf{x}_{m-1}^{(b)}$ in the annealing layer R, the state s_i most likely to have been active after the HMM has emitted the observable $\mathbf{f}_{m-1}^{(b)}$, is found.
2. The activity model is initialised in state s_i and allowed to make K_i^R state transitions via A_{ij}. The final emission via $p_j(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j^R)$ forms the new estimate $\mathbf{f}_m^{(b)}$.
3. The remaining particle location parameters given by $\boldsymbol{\omega}_{m-1}^{'(b)}$ are reestimated by drawing from the multivariate Gaussian distribution with covariance matrix $\mathbf{P}_R$ and mean $\boldsymbol{\omega}_{m-1}^{'(b)}$.
4. The new particle location is now given by the feature vector $\mathbf{x}_m^{(b)} = (\mathbf{f}_m^{(b)} || \boldsymbol{\omega}_m^{'(b)})$ and the weighting $\pi_m^{(b)}$ may be calculated. If resampled, the prediction is subsequently refined over annealing layers with all Gaussians rescaled as in Eq. (10). In addition, where $s_j = s_i$ after K_i^r transitions, we add noise to the learned parameters from the distribution $p_i'(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \boldsymbol{\mu} = \mathbf{f}_m^{(b)}, \boldsymbol{\Sigma}_i^r)$.

7. Experiments

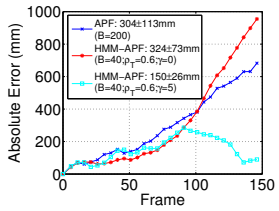
We added the steps detailed in the last section to the Matlab implementation of APF made available by Balan et al. [3] and attempted tracking on the first 150 frames of the walking test sequence. In the case of known subject tracking, PCA was applied to walking joint angle data $E_k = \{\epsilon_1, ..., \epsilon_{1900}\}$ taken from a MoCap training sequence of the subject to be tracked (data featuring the subject standing still was removed for the case of monocular tracking). In the case of unknown subject tracking, PCA was applied to walking joint angle data $E_{\bar{k}} = \{\epsilon_1, ..., \epsilon_{2100}\}$ taken from subjects S1, S2 and S3 [10] (700 consecutive frames each). The first $\eta = 4$ eigenvectors were retained (preserving 93% and 92% of the walking training data in E_k and $E_{\bar{k}}$, respectively) and an $N = 30$ state HMM trained with the resulting 4-D timeseries. In order to recover the known subject's body pose vector at time m from an observable emission, the relationship in Eq. (1) was used and their corresponding mean body pose added. In the case of an unknown subject, the average of all the training subject's mean body poses was added.

For HMM training, initial estimates of the state means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ were found by K-means clustering. The transition matrix A_{ij} was initialised randomly (and each row normalised) and the prior A_i fixed with every value being $1/N$. The transition probabilities, state means and state covariances were reestimated using no more than 50 iterations of the Baum-Welch update equations, performed using the HMM Toolbox for Matlab [7]. For tracking $R = 5$ annealing layers and 40 particles were used (effectively 200 particles per frame).

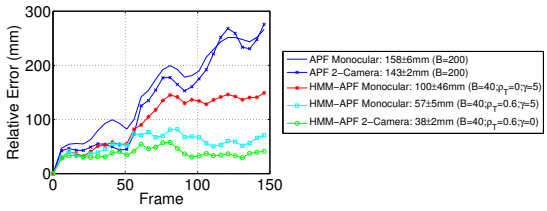
Results for the HMM-APF approach are shown in Figure 2 with standard APF (using $\mathbf{P}_r$ for the propagation of particles and edge-plus-silhouette weighting function, with $W = 1$ in Eq. (8)) included for comparison. For each setting, tracking of the sequence was attempted 10 times with the HMM reestimated from training data each time. The absolute error is calculated as the average distance in millimetres between a set of virtual markers located at the joint positions of the body model and 15 corresponding MoCap markers on the joints of the subject. The optimistic absolute error [3], is given by the lowest error of any particle in the set and provides a lower bound on error suitable for the cross-comparison of particle based tracking methods. The average optimistic error across the 10 runs is plotted at every 5th frame. The average optimistic error was also calculated across each run and the mean and standard deviation of error across the 10 runs is shown in the legends. For the relative error calculation in Figure 2(c), the global coordinates of the virtual and MoCap pelvis markers are set equal before taking the average marker error. A weighted average over all particle errors is then calculated, with the error of each particle weighted by $\pi_m^{(b)}$.



(a) 2-camera tracking: average optimistic absolute error for known and unknown subject.



(b) Monocular tracking: average optimistic absolute error for known subject.



(c) Weighted relative error: known subject.

Figure 2. Tracking results for HMM-APF versus standard APF.

8. Discussion and Future Work

For the 2-camera case HMM-APF outperforms standard APF in the case of both known and unknown subject tracking with a greater than 50% reduction in optimistic absolute error. HMM-APF also appears robust, maintaining a good estimate of current pose throughout the sequence in each of the 10 runs.

In the case of monocular tracking, maintaining an accurate estimate of the subject's global coordinates is very challenging. The body model tends to 'sit back', ensuring it is enveloped by image evidence and scoring well in terms of the APF weighting function. This can be seen in Figure 2(b) for APF and for HMM-APF with $\gamma = 0$ where the high absolute errors are due, overwhelmingly, to error in estimating the subject's global position. Enforcing agreement between the silhouette sizes by setting $\gamma = 5$ causes the model to move with the subject as they start to walk towards the camera (around frame 90 in Figures 2(b) and 3). Absolute error arising from inaccuracy in the global coordinates is difficult to eliminate entirely, still reaching almost 300mm for $\gamma = 5$, but (with $\rho_T = 0.6$) correct pose recovery is now observed across all 10 runs, giving a mean weighted relative error of $57 \pm 5mm$ (Figure 2(c)).

We hope that modelling activity within the HMM framework will allow for the simultaneous consideration of multiple activity models. Here we require an approach capable of quantifying the probability that a pose vector has been emitted by a set of distinct activity models, allowing for the appropriate distribution of particles between activity sub-

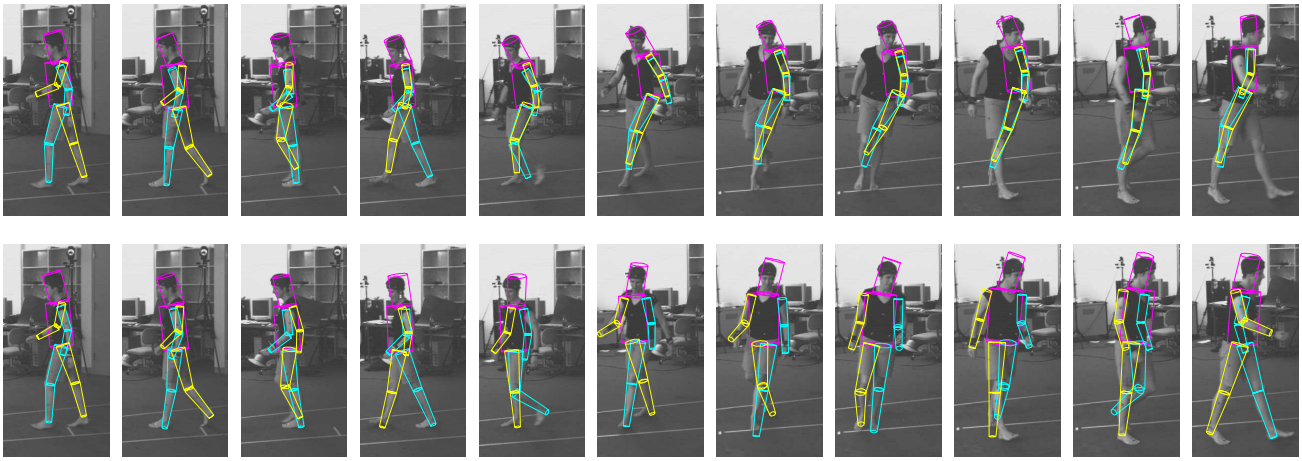


Figure 3. Monocular tracking for $B = 200$ APF (top) and $B = 40$ HMM-APF (bottom), every 15th frame (video is 60fps).

spaces. We aim to track sequences featuring multiple activities in future work.

9. Conclusion

We have demonstrated a method for human motion tracking by using a HMM for the propagation of particles in a modified APF scheme. The PCA-based dimensionality reduction of the feature space reduces the difficulty of both the particle filtering task and the HMM training task. The activity HMM simulates a nonlinear activity axis within the reduced space. APF guided by traversal of the HMM is able to recover pose for the walking sequence using fewer than 3 cameras and 200 particles per frame. Standard APF experiences rapid failure using 5 times as many particles e.g. see Figure 3.

10. Acknowledgments

This research was supported by an MMU Dalton Research Institute research studentship and EPSRC grant EP/D054818/1. We would also like to thank the authors of [3] for making their data and Matlab code available and for their helpful comments.

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 Trans. on Sig. Proc.*, 50(2):174–188, 2002.

[2] P. Azad, A. Ude, R. Dillmann, and G. Cheng. A full body human motion capture system using particle filtering and on-the-fly edge detection. In *ICHR*, pages 941–959, 2004.

[3] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3D person tracking. In *VS-PETS*, pages 349–356, 2005.

[4] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2):185–205, 2005.

[5] N. D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[6] R. Li, M.-H. Yang, S. Sclaroff, and T.-P. Tian. Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers. In *ECCV*, pages 137–150, 2006.

[7] K. Murphy. Hidden Markov model toolbox for Matlab. www.cs.ubc.ca/˜murphyk/Software/HMM/hmm.html.

[8] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

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

[10] L. Sigal and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI, 2006.

[11] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with gaussian process dynamical models. In *CVPR*, pages 238–245, 2006.

[12] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *ICCV*, pages 403–410, 2005.

[13] R. Urtasun and P. Fua. 3D human body tracking using deterministic temporal motion models. In *ECCV*, pages 92–106, 2004.

[14] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *NIPS*, pages 1441–1448, 2005.