Behaviour Based Particle Filtering for Human Articulated Motion Tracking

J. Darby, B. Li and N. Costen
Department of Computing and Mathematics, Manchester Metropolitan University, U.K.
j.darby@mmu.ac.uk, b.li@mmu.ac.uk, n.costen@mmu.ac.uk

Abstract

This paper presents an approach to human motion tracking using multiple pre-trained activity models for propagation of particles in Annealed Particle Filtering. Hidden Markov models are trained on dimensionally reduced joint angle data to produce models of activity. Particles are divided between models for propagation by HMM synthesis, before converging on a solution during the annealing process. The approach facilitates multi-view tracking of unknown subjects performing multiple known activities with low particle numbers.

1. Background

Techniques based on particle filtering have been widely used in human motion tracking [4, 5]. Given enough particles it is possible to approximate a posterior distribution for the configuration of a human body given a series of observations [1]. However, the typically high number of degrees of freedom in full body tracking feature spaces results in a large particle requirement. The evaluation of a weighting measure for each hypothesis makes human motion tracking a computationally demanding task. Annealed Particle Filtering (APF) [4] is a variation of Sampling Importance Resampling (SIR) [1] which reduces computational load by attempting to recover only the global maximum of the posterior distribution at each time step. APF has been shown to achieve better tracking accuracy than SIR given the same number of particles [2].

To avoid searching in high dimensional feature spaces, approaches to tracking often make assumptions about the class of movement and look for solutions in low dimensional pose spaces recovered from training data [3, 7]. Inspired by earlier work [5], we extend a previous approach using a single activity model [3] to give simultaneous consideration to multiple models. Low dimensional activity feature spaces are produced by the application of PCA to training data. The resulting data distributions and associated dynamics are modelled by training hidden Markov models (HMMs). Synthesis is used for efficient particle dispersion in APF to allow tracking with low particle numbers. PCA is computationally cheap compared with other, nonlinear, dimensionality reduction techniques [7], as is the transfer of particles from one activity subspace to another. Particles are free to migrate between models both at each frame and at each annealing layer and we consider the recovered distribution as an activity classifier.

2. Method

The HumanEva-II dataset [6] contains multi-camera synchronised video sequences of subjects performing various activities. Associated ground truth MoCap data allows for the quantitative evaluation of tracking accuracy (§ 3) and separate training MoCap data for the estimation of a body model (§ 2.2) and learning of activity models (§ 2.3). We start by briefly reviewing particle filtering and its annealing extension.

2.1. Particle Filters and Annealing

Human motion can be modelled as the evolution of a system state $\mathbf{x}_t$ over time $t = 1, 2, ..., T$, described by a Markov process and observed by a sensor providing independent observations, $\mathbf{Z}_t = (\mathbf{z}_1, ..., \mathbf{z}_t)$. The posterior distribution can be obtained according to the recursion

$$p(\mathbf{x}_t|\mathbf{Z}_t) \propto p(\mathbf{z}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{Z}_{t-1}). \tag{1}$$

In SIR a multimodal system state is represented via a finite set of $b = 1, ..., B$ normalised, weighted particles. *Dispersion* by a model of temporal dynamics, *evaluation* by a weighting function and *resampling* with probability proportional to weighting score, propagates the probability distribution over time [1]. An estimate of the system's current state can be obtained by calculating the sample mean of the distribution $\mathcal{E}[\mathbf{x}_t]$.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

Annealed particle filtering [4] attempts to recover a global maximum by cooling the weighting distribution and then gradually introducing sharp peaks over $r = R, R-1, ..., 1$ resampling layers at each time step t,

$$w_t^r(\mathbf{z}_t, \mathbf{x}_t) = w(\mathbf{z}_t, \mathbf{x}_t)^{\beta_t^r}, \quad \beta_t^1 > ... > \beta_t^R. \tag{2}$$

The value of β_t^r is chosen to control the particle survival rate α_r, the proportion of particles that will be resampled; here $\alpha_R = ... = \alpha_1 = 0.5$. The particle survival rate is also used to scale the addition of Gaussian noise for dispersion, $\mathbf{P}_r = (\alpha_r)^{R-r} \times \mathbf{P}$, where $\mathbf{P}$ is a noise covariance matrix estimated from training data. The posterior distribution is not fully represented, giving a reduction in computation at the expense of a departure from the formal Bayesian framework.

2.2. Body Model: Training and Evaluation

Human body configuration is approximated by a simple geometric body model. The model consists of a kinematic tree containing 10 truncated cones and is specified by the location and orientation of the torso and relative joint angles between limbs, $\boldsymbol{\omega}$. In offline training, body model configuration vectors were calculated from HumanEva-II MoCap training data and used to learn models of activity. During tracking, particles were evaluated by projection of the corresponding cone configuration into the image planes for the weighting function calculation.

Training: For all $m = 1, ..., M$ training data vectors available for a given activity, the 6 global translation and rotation elements and the 3 head orientation elements were removed and each subject's mean vector subtracted from each of their activity vectors. A single global PCA was used and the $\eta = 4$ highest-variance eigenvectors were retained. Each training pose was specified by the 13D feature vector

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

The head and torso parameters were used to estimate a covariance matrix $\mathbf{P}$ for dispersion as in standard APF, while the remaining pose approximations given by $\mathbf{f}_m$ were used to train an HMM for dispersion by synthesis.

Evaluation: The dimensions of the body model cones were estimated from the MoCap data of the tracking subject. For a given particle b, the model as specified by $\mathbf{x}_t^{(b)}$ was projected into each image plane and an evaluation of its correlation with image data $\mathbf{z}_t$ made, $w(\mathbf{z}_t, \mathbf{x}_t^{(b)}) \approx p(\mathbf{z}_t | \mathbf{x}_t^{(b)})$. The coordinates of points on the surface and edges of each component cylinder were used to sample from silhouette and smoothed edge maps calculated from the current image evidence (the reader is referred to [4] for a more detailed discussion). Particles describing a pose with intersecting cones were given a zero weighting.

2.3. Behavioural Models

In the acquisition of training data (§ 2.2), both a human's performance of an intended activity and the observation of that performance are stochastic processes. HMMs allow us to describe such a doubly stochastic system. An HMM was learned from each batch of activity training data $F = \{\mathbf{f}_1, ..., \mathbf{f}_M\}$ using the Baum-Welch algorithm. An HMM λ is specified by a set of states $S = \{s_1, ..., s_N\}$; a prior A_i giving the probability of a sequence starting in state i; an $N \times N$ matrix A_{ij} giving the probability of a transition from state i to state j; and a set of observation densities $p_i(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ giving the probability of observing features $\mathbf{f}$ while in state i. The prior was kept flat and the observation densities modelled by single multivariate Gaussians, initial estimates of which were found by k-means clustering.

The approach to particle dispersion at each time step can be reposed as two separate predictive tasks where for each particle $\mathbf{x}_{t-1}^{(b)}$; the first 9 parameters are re-estimated by adding Gaussian noise to give $\boldsymbol{\omega}_t^{'(b)} | \boldsymbol{\omega}_{t-1}^{'(b)}$, while the last η parameters are re-estimated by querying activity HMMs to get $\mathbf{f}_t^{(b)} | \mathbf{f}_{t-1}^{(b)}$.

For each particle in the first annealing layer, R, the parameters $\mathbf{f}_{t-1}^{(b)}$ are randomly assigned to one of the activity models. The system state s_i most likely to have been active after the model λ has emitted the sequence $\{\mathcal{E}[\mathbf{f}_{t-2}], \mathbf{f}_{t-1}^{(b)}\}$ is found. The model is initialised in state s_i and allowed to make one state transition via A_{ij} and one emission via $p_j(\mathbf{f})$, this is the new estimate $\mathbf{f}_t^{(b)} | \mathbf{f}_{t-1}^{(b)}$. The new estimate $\boldsymbol{\omega}_t^{'(b)}$ is found by sampling from the Gaussian distribution with mean $\boldsymbol{\omega}_{t-1}^{'(b)}$ and covariance matrix $\mathbf{P}_R$. The new particle location is given by the feature vector $\mathbf{x}_t^{(b)} = (\boldsymbol{\omega}_t^{'(b)}, \mathbf{f}_t^{(b)})$.

If the particle is resampled, the most likely active state in both models is calculated and the allocation made with likelihood proportional to their associated probabilities. The chosen state does not transition before emitting a further observable. In line with the re-scaling of $\mathbf{P}_r$ for the re-estimation of $\boldsymbol{\omega}_t^{'(b)}$ (§ 2.1), all observation densities are re-scaled at each annealing layer, $\boldsymbol{\Sigma}_i^r = (\alpha_r)^{R-r} \times \boldsymbol{\Sigma}_i$. The mean $\boldsymbol{\mu}_i$ is replaced by the current estimate of $\mathbf{f}_t^{(b)}$ in order that weighting function scores rather than training data guides final convergence.

3. Results

Tracking was performed on the HumanEva-II dataset [6]. In each experiment the tracking subject's activity training data were excluded from the HMM training. The particle set was initialised using ground truth and the absolute error – the average distance between 15 3D joint-centre locations [2] – between the recovered sample mean body model configuration $\mathcal{E}[\mathbf{x}_t]$ and the ground truth MoCap was calculated at each frame. All sequences are 60fps.

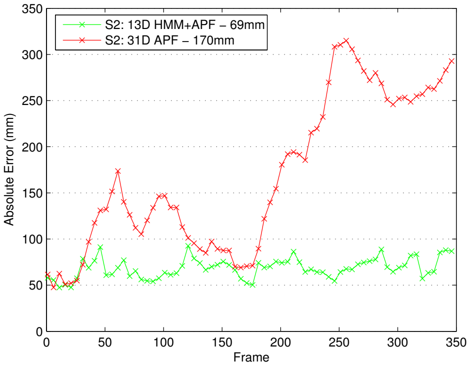


Figure 1. Absolute error for HumanEva-II Subject 2 (S2) walking portion of the Combo sequence. Note that S2 is excluded from the training data.

3.1. Tracking Walking

A 30-state HMM and noise covariance matrix $\mathbf{P}$ were estimated using body model configuration data taken from 3 subjects (700 consecutive frames each) as described in § 2.3. The resulting dynamical models were then used to guide APF tracking of a 350 frame sequence featuring Subject 2 (S2) walking. The absolute error at every 5th frame is shown in Figure 1 with standard APF in the full 31D space included for comparison. Both experiments used 20 particles and 5 annealing layers. Tracking is lost early using standard APF, it is briefly recovered due to the cyclic nature of the activity before permanently failing. For APF, 100 particles per frame is too few to sufficiently explore the feature space. Using the HMM for propagation facilitated robust tracking of the walking sequence.

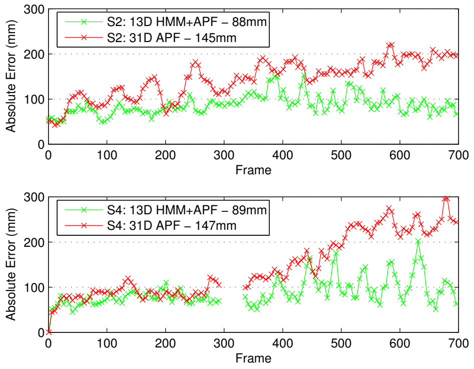


Figure 2. Error for HumanEva-II Walking/Jogging portion of the Combo sequences for S2 and S4. No error is plotted where ground truth MoCap data were unavailable.

3.2. Tracking Walking and Jogging

A second 30-state HMM was trained using jogging data and used in combination with the walking model from § 3.1 to attempt tracking on two longer sequences, half the frames of walking followed by half of jogging. Issues of training data quality meant the jogging model was based on a single subject. The number of particles was doubled to 40 over 5 annealing layers. The absolute error at every 5th frame is shown in Figure 2.

With twice as many particles APF performs better on walking, but performance is worse for jogging. Inclusion of jogging training data in the estimation of the covariance matrix caused tracking to fail during walking. Tracking guided by the HMMs produced lower absolute error scores. The use of more training subjects should help improve performance further, reducing the variations in error for jogging which we attribute at least in part to stylistic differences between the tracking subjects and the single training subject.

Generally, all or most particles in the 'wrong' model are gone within 2-3 annealing layers due to their low weighting scores. The percentage of particles remaining in each model after the annealing process has completed, averaged over the previous 12 frames (0.2sec), is shown in Figure 3. Complete migration of the particle set is seen in a small number of frames, e.g. around frame 600 for S4, and produces an incorrect albeit highly weighted mean pose. In each case the correct pose is recovered within a few frames and a smoothed version of the particle distribution between multiple models could be used as an activity classifier.

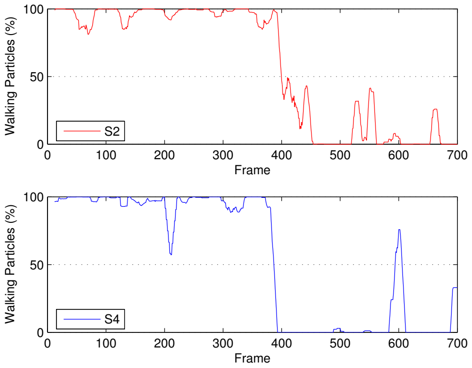


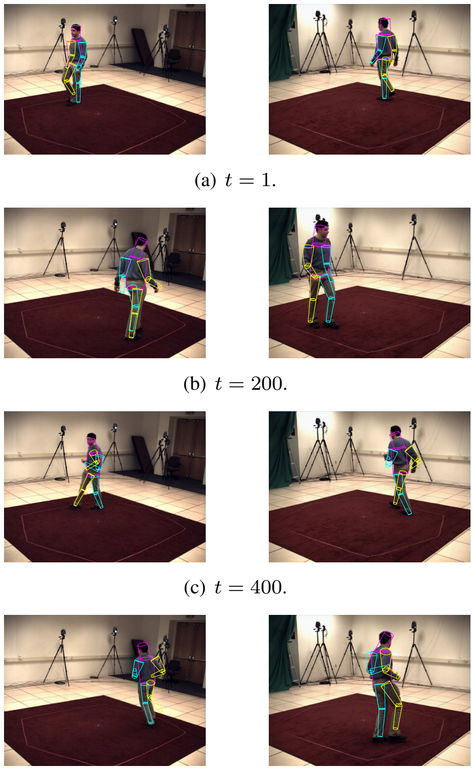
Figure 3. Average over the previous 0.2sec of the distribution of particles between walking and jogging models at final annealing layer.

4. Discussion and Conclusions

The tracking results for the Combo sequences represent good accuracy, with errors in tracking quickly recovered. Image data from 2 of the 4 cameras with the tracking model superimposed is shown in Figure 4. We have used models of behaviour to reduce the computation required for a given level of tracking accuracy. They could also be used to help guarantee reliable tracking given degraded test data. For example, it should be possible to combine the set of predictive models with standard APF, handing over a proportion of particles to an activity model depending on its proximity in terms of the original feature space. Such a scheme could improve tracking robustness where image evidence is weak due to poor silhouettes, fewer cameras or subject occlusion. In ongoing work we are investigating the learning of low dimensional activity spaces from hierarchical subtrees of the body model, rather than solely at the scale of full body configurations.

We have shown how the assumption of a known activity e.g. [3, 7] may be relaxed to one of a set of known activities. PCA is used to create multiple activity spaces from training data and HMMs are trained to guide their exploration. A particle-based approach allows us to consider multiple hypotheses from multiple activity models, with annealing providing a method to distill out the best candidate at each frame.

Acknowledgments: We thank the authors of [2] for making their tools and data available; the HMM implementation used Kevin Murphy's HMM Toolbox for Matlab.



(a) $t = 1$.

(b) $t = 200$.

(c) $t = 400$.

(d) $t = 600$.

Figure 4. Tracking results for S2, every 200th frame from 2 of 4 cameras used.

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, 2002.

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

[3] J. Darby, B. Li, and N. Costen. Tracking a walking person using activity-guided annealed particle filtering. In *F&G*, 2008. (Accepted to appear).

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

[5] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, pages 702–718, June 2000.

[6] 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, Providence, RI, 2006.

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