Generative Model for Human Motion Recognition

David Excell A. Taylan Cemgil William J. Fitzgerald
Cambridge University Engineering Department
Signal Processing and Communications Laboratory
Trumpington Street, Cambridge, CB2 1PZ, UK
{dae30, atc27, wjf1000}@cam.ac.uk

Abstract

This paper describes a generative Bayesian model designed to track an articulated 3D human skeleton in an image sequence. The model infers the subjects appearance, pose, and movement. This technique provides a novel method for implicity modelling depth and self occlusion, two issues that have been identified as drawbacks of existing models. We also employ a switching linear dynamical system to efficiently propose skeleton configurations. The model is verified using synthetic data. A video clip from the Caviar data set is used to demonstrate the potential of the methodology for tracking on real data.

1 Introduction

The task of tracking humans in an image sequence has seen a significant research investment. To date most algorithms consider tracking the body as a solid moving object that stands out from the background. These models allow for the general tracking of people as they move through a scene. These models are unable to assist with understanding detailed behaviour of humans. Articulated models provide this detail but introduce added complexity to the inference. One particular complexity occurs when body parts self-occlude each other, for example, when walking is observed in the sagittal plane (side on), it is often difficult to identify the left and right legs. The ambiguity is reduced when the legs cross as the layering of legs becomes evident. Utilising a 3D skeleton model this layering of joints can be inferred when it is supported by the data. Our inference model associates a depth with each body part in the image plane thus allowing occlusion ambiguities to be resolved efficiently.

The positions of all model objects are tracked in 3D space allowing the depths of multiple people to be calculated in the image plane. When scaling this model to a crowded scene these depths will enable occlusion amongst crowd members to be appropriately represented. Maintaining 3D locations also allows multiple observations from different locations to be fused in a general framework.

There are several applications for a robust method of tracking the articulate motion of a human. Surveillance systems used in public spaces can be improved to interpret the actions of the individuals. Human Computer Interfaces can be significantly improved to bypass the traditional inputs of a keyboard and mouse. Expanding methods of interaction will enable technology use to become more efficient in areas such as manufacturing and repair. Visual and motion inputs can also be used to enhance the game play in computer games as is currently being explored by Sony's EyeToy and the Nintendo Wii. Enhancing the range and accuracy during this interaction will increase the possibilities of future game development.

1.1 Related Work

There is a significant library of literature on tracking humans in video. Some of the more recent papers include [2, 4, 13]. In [13] each human object is modelled by a head position, hight, thickness and 2D inclination. A colour histogram is used for the appearance model and three ellipsoids are used for the shape. Kalman filters are used for the temporal estimation of these parameters. In [2] image features, calculated on a frame-by-frame basis are use to learn feature trajectories and then infer independent motion of clustered features.

Star structures have been used to classify different types of motion in [6]. The star shape was used as it is an efficient representation of a person walking when viewed form side-on. The homogeneous structure used to detect the person as a whole suffers under occlusion as the detection methods do not degrade gracefully. Pavolvić uses a Switching Linear Dynamic System (SLDS) model to track humans walking from a side-on view [10]. The model described is restricted to 2D and uses templates learnt from the first frame to match the object in future frames.

Individual human part detectors have become a standard method to detect the pose of humans in images [4, 8, 9, 13]. In [8] for static images and [9] for an image sequence a data driven MCMC algorithm to propose possible body configurations is described. The observation likelihood function is calculated by synthesising the human form and comparing it to the input image. The comparison considers region coherency, colour dissimilarity with the background, skin

colour likelihood and foreground matching. The results of the part detectors are integrated to generate proposal maps of joint configurations. For the image series, a dynamical model is used to propose a sequence of skeleton configurations. Batch processing is used to enable forward and backward propagation of state information.

2 Human Skeleton Model

Given a person with location and orientation described by a 6D vector $\mathbf{p}_t = \{x, y, z, \alpha, \beta, \gamma\}$ and their current pose defined by the skeleton configuration κ_t at time instant t, let the location, orientation and joint positions be described by Gaussian random variables. The dynamics of the global position and orientation of the person can be described by the state space model $\{\mathbf{A}_p, \mathbf{B}_p, \mathbf{Q}_p\}$

$$\begin{aligned} \mathbf{p}_t &\sim \mathcal{N}\left(\mathbf{p}_t; \mathbf{A}_p \mathbf{p}_{t-1} + \mathbf{B}_p, \mathbf{Q}_p\right) \\ \mathbf{p}_0 &\sim \mathcal{N}\left(\mathbf{p}_0; 0, \mathbf{\Sigma}_p\right) \end{aligned}$$

In this paper the global motion is restricted to forward movement. Therefore $\mathbf{A_p}$ is the identity matrix and $\mathbf{B}_p = [v_x, 0, 0, 0, 0, 0]^T$. The forward velocity v_x is a random variable assumed to be normally distributed,

$$v_x \sim \mathcal{N}\left(v_x, \mu_{v,x}, \sigma_{v,x}\right).$$

To reduce the computational requirement of tracking the global motion, the mean shift algorithm [3] could be used to generate efficient proposal locations in a new frame. The rotation variables need still to be inferred but the admissible range of rotational changes in human motion is generally small and well predictable and thus efficient inference is possible.

The dynamics of the joint positions are defined within a local coordinate framework with the origin at the center of the body, the x-axis though the sagittal plane, the y-axis through the coronal plane and the z-axis through the transverse plane. The coordinate system is illustrated in Figure 1. The dynamics of the local joints are defined by a SLDS model, defined by $\{\mathbf{A}_\kappa^{m_t}, \mathbf{C}_\kappa^{m_t}, \mathbf{Q}_\kappa^{m_t}, \mathbf{R}_\kappa^{m_t}, \mu_\kappa^{m_t}, \mathbf{\Sigma}_\kappa^{m_t}\}$, where m_t describes the index of the model at time t. The number of known linear models is denoted by M. The SLDS is learnt from motion capture data acquired at a rate of 120 Hz, to time-align the dynamic system $\mathbf{A}^m$ is raised to the power 120/30 for video capture at 30 Hz. Details of the learning process for these models is described in Section 2.1. The dynamics of the local joints is given by

$$\begin{aligned} \mathbf{x}_{\kappa,t} &\sim \mathcal{N}\left(\mathbf{x}_{\kappa,t}; \mathbf{A}_\kappa^{m_t} \mathbf{x}_{\kappa,t-1}, \mathbf{Q}_\kappa^{m_t}\right) \\ \kappa_t &\sim \mathcal{N}\left(\kappa_t; \mathbf{C}_\kappa^{m_t} \mathbf{x}_{\kappa_t}, \mathbf{R}_\kappa^{m_t}\right) \end{aligned}$$

The initial state of the dynamic system for a newly observed body is estimated from the library of possible states defined by the SLDS. For repetitive behaviours, such as walking, selecting the initial phase is equivalent to identifying the phase of the system. The initial state and rate of evolution define all future states. The likelihood that any states be selected as the initial state is assumed to be uniform across the library,

$$\begin{aligned} \mathbf{x}_{\kappa,0} &= \mathbf{x}_{\kappa,\delta} \\ \delta &\sim \mathcal{U}\left(\delta; 0, T^m\right). \end{aligned}$$

The time instant of switching between linear models is parameterised by μ_m and σ_m, where μ_m denotes the mean period that model m is active and Σ_m describes the variance of this measure. The next switching instant τ_j is given by

$$\tau_j \sim \mathcal{N}\left(\tau_j; \tau_{j-1} + \mu_m, \sigma_m\right) \quad (1)$$

The transition from model $m_{\tau_{j-1}}$ to model m_{τ_j} is given by the $M \times M$ transition matrix $\mathbf{A}_m$. For the walking behaviour used in this paper $\mathbf{A}_m$ is the identity matrix.

The SLDS describes the joint positions relative to the humans local coordinate system. To generalise the model for arbitrary camera locations the local joint positions (κ_t) are projected to positions in a global coordinate system via the projection matrix parameterised by the individual's location vector $\mathbf{p}_t$

$$\kappa_{g,t} = H\left(\mathbf{p}_t\right) \kappa_t$$

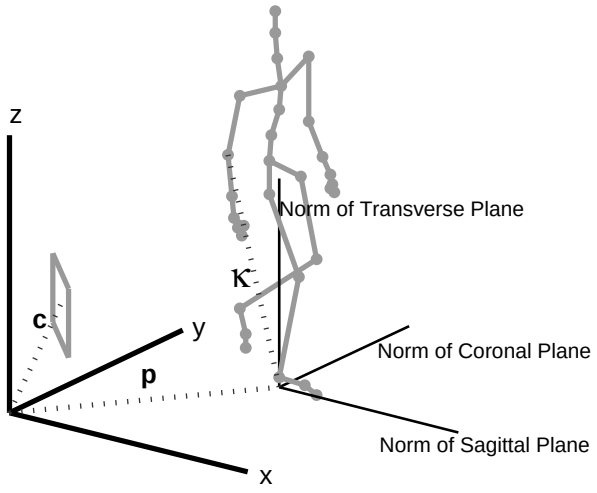


Figure 1. Global and Local coordinate system with camera and skeleton positions.

$H\left(\mathbf{p}_t\right)$ is defined as,

$$H\left(\mathbf{p}_t\right) = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \mathbf{T}\left(x, y, z\right) = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\mathbf{R} = \mathbf{R}_x\left(\alpha\right) \mathbf{R}_y\left(\beta\right) \mathbf{R}_z\left(\gamma\right)$$

The rotations around the axis are defined with $c_\alpha = \cos\left(\alpha\right)$ and $s_\alpha = \sin\left(\alpha\right)$ as

$$\mathbf{R}_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_\alpha & s_\alpha \\ 0 & -s_\alpha & c_\alpha \end{bmatrix} \quad \mathbf{R}_y(\beta) = \begin{bmatrix} c_\beta & 0 & -s_\beta \\ 0 & 1 & 0 \\ s_\beta & 0 & c_\beta \end{bmatrix}$$

$$\mathbf{R}_z(\gamma) = \begin{bmatrix} c_\gamma & s_\gamma & 0 \\ -s_\gamma & c_\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Given a known, fixed, camera location and orientation $\mathbf{p}_c = \{x, y, z, \alpha, \beta, \gamma\}$, the global joint positions are projected into the the observed image space, denoted by $\mathbf{J} = [u, v, 1]^T$ by

$$J_t = \mathbf{K}[\mathbf{R}_c|\mathbf{t}_c]\kappa_{g,t} \qquad K = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix}$$

The image space is defined by $N = W \times H$ pixels, where $u = [-(W-1)/2, (W-1)/2]$ and $v = [-(H-1)/2, (H-1)/2]$. The camera properties are encapsulated by $\mathbf{K}$ which includes the focal plane, f, and principle point, p_x, p_y. $\mathbf{R}_c$ defines the rotation of the camera, and $\mathbf{t}_c = -\mathbf{R}_c[x_c, y_c, z_c]^T$ denotes translation. For a more detailed description of the camera transformation matrix see [7, pp. 153–158].

Given two connected joints denoted by $\mathbf{J}_{a,t}$ and $\mathbf{J}_{b,t}$, a body part is defined by the rectangle $\mathbf{s}_k = [w_k, h_k]$, $k = \{a, b\}$ where, a and b are index's of the joints and k is an index over the body parts. There are a total of K body parts. The region enclosed by s_k in the image plane corresponds to the estimated position where the body part would be observed given no occlusion. The rectangle is defined by its width w_k and height h_k, the rectangle is oriented such that its major axis coincides with the line connecting the joints $\mathbf{J}_{a,t}$ and $\mathbf{J}_{b,t}$. To simplify the model the height is equal to the distance between joints, $h_k = \|\mathbf{J}_{a,t} - \mathbf{J}_{b,t}\|$ and w_k is a random variable distributed according to

$$w_k \sim \mathcal{P}(w_k)$$

Occlusion is the biggest source of errors in previous human tracking papers cited in the introduction. For each body joint we define a depth variable z_a. The depth value is measured perpendicular to the image plane and can be calculated directly from the 3D skeleton model

$$z_a = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \kappa_{g,t}$$

Background pixels are defined to have a depth of $z = 0$. The depth of a body is calculated from the mean of the two connecting joints. Body parts have depths greater than 0, the greater the value the more likely the body part will be visible. To maintain this property the depth of all body parts are normalised.

Given model parameters, $\mathbf{p}_c$, $\mathbf{p}_t$, κ_t and w_k we define a variable r_i to indicate how likely pixel i will corresponds to body part k.

$$p(r_{1:N}|z_k, w_k) = \mathcal{C}(r; \pi_{i,0}, \ldots, \pi_{i,k}, \ldots, \pi_{i,K})$$

$$\pi_{i,j} = \frac{\exp(g_{i,k})}{\sum_{k'=0}^{K} \exp(g_{i,k'})}$$

$$g_{i,0} = 0$$
$$g_{i,k} = z_k \phi(s_k, J_{t,k}, x_i)$$

where $\mathcal{C}$ is the categorical distribution with cell probabilities $\pi_{i,k}$. The indicator function $\phi(s_k, J_{t,k}, x_i) = 1$ if x_i is located within the region of the body part and 0 otherwise.

This model has an explicit method for representing self occlusion of body parts through the indicator variable r_i. If two body parts, k and k' are co-located at pixel i, the depth variables z_k and $z_{k'}$ enable the model to explicitly describe the probability of observing the appearance of either object relative to their distance from the camera. Therefore if $z_k < z_{k'}$ then the appearance model φ_k is more likely than $\varphi_{k'}$. As the difference in depth approaches zero the appearance model of either body part become equally likely.

2.1 Dynamical Model

To model the evolution of the pose throughout the image sequence we use a switching linear dynamic system. SLDS's are shown to have superior performance modelling skeleton dynamics than linear systems [1]. Subspace techniques are used to segment data captured from a motion capture system to train the individual linear systems [11]. For walking behaviour demonstrated in this paper two linear systems are learnt, a left and right leg swing model. For each model we learn the dynamics in the form

$$\mathbf{x}_t = \mathbf{A}_\kappa^m \mathbf{x}_{t-1} + \mathbf{B}_\kappa^m + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \mathbf{Q}_\kappa^m)$$
$$\mathbf{y}_t = \mathbf{C}_\kappa^m \mathbf{x}_t + e \qquad e \sim \mathcal{N}(0, \mathbf{R}_\kappa^m)$$

A standard EM algorithm is used to learn the parameters $\mathbf{A}_\kappa^m$, $\mathbf{B}_\kappa^m$, $\mathbf{C}_\kappa^m$ and the initial conditions $\mathbf{x}_0$. The EM algorithm is initialised from a closed from subspace solution. To reduce the complexity of the optimisation, Principle Component Analysis was applied to the skeleton configuration reducing it from 56 angles to 8 dimensions. Figure 2 demonstrates the accuracy of the learnt model to describe the walking motion. Expanding the framework to infer a richer set of behaviours simply requires expanding the number of linear systems contained within the model. Data for training the models was obtained from the CMU motion capture database (`mocap.cs.cmu.edu`).

2.2 Appearance Model

Given two connected joints denoted by positions $\mathbf{J}_a$ and $\mathbf{J}_b$, we know that there is a rigid connecting bone, denoted by index value k. We will assume that each bone has a fixed width w_{ij}. The appearance of the bone will be modelled as a Gaussian with mean φ_k and variance Σ_φ, where φ_k denotes the mean colour in the colour space and Σ_φ denotes the

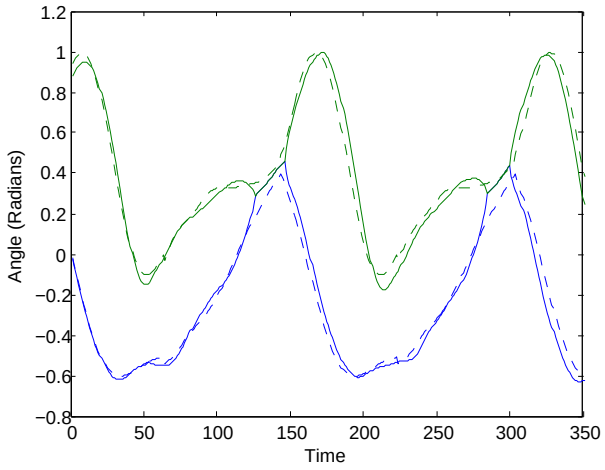


Figure 2. Comparison between motion capture (solid line) and simulated (dashed line) human walking motion. The data shown is the angle of the right knee (lower line pair) and hip (upper line pair) joints.

noise. Background pixels are similarly modelled as Gaussian random variables with mean b_i and variance Σ_b. The 3D red-green-blue colour space is used throughout this paper. The noise models of the background and foreground pixels are parameterised separately to allow the noise of the foreground pixels to be reduced during inference. The likelihood of a pixel colour matching the appearance of the k bone is given by

$$p\left(y_i|\varphi_{1:K}, b_i, r_i = k\right) = \begin{cases} \mathcal{N}\left(y_i; b_i, \Sigma_b\right) & \text{if } r_i = 0 \\ \mathcal{N}\left(y_i; \varphi_k, \Sigma_\varphi\right) & \text{if } r_i = j \end{cases}$$

More advanced appearance models, exploiting patterns within clothing could be incorporated within this framework to improve tracking at the expense of computation.

3 Estimation Framework

Given an image sequence $\mathbf{Y}_{1:T}$ the goal is to estimate the behaviour (position and pose sequence) of the subject. The Bayesian framework is utilised to allow the uncertainty in parameters to be propagated through the model. The behaviour is specified by the parameters $\mathbf{B}_t = \{\mathbf{p}_t, \kappa_t, m_t\}$. The behaviour is observed in the image via an observation model described by the parameters $\mathbf{O}_t = \{\mathbf{w}_{1:K}, \varphi_{1:K}\}$. The coupling of the two parameter sets is shown in Figure 3. The posterior probability is given by

$$p\left(B_{1:T}, O_{1:T}|Y_{1:T}\right) \propto p\left(Y_{1:T}|B_{1:T}, O_{1:T}\right) p\left(B_{1:T}, O_{1:T}\right) \quad (2)$$

The behavioural model evolves with time as shown by the graphical model in Figure 4(a). These parameters then become inputs to the observation model. The relationships between observation parameters is shown in Figure 4(b). Through the relationships described in the observation model we are able to determine the likelihood of

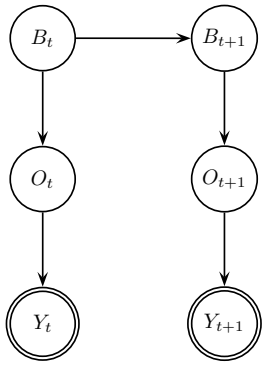


Figure 3. Coupling of behavioural and observation model parameters. The behavioural parameters evolve with time and the observation parameters are derived directly from these values.

the behaviour parameters given the image sequence. The observation model contains parameters for the appearance (φ_k) and width (w_k) that are assumed to remain constant throughout the sequence.

All model parameters are randomly initialised. To improve the rate of convergence more intelligent initialisations could be made from measuring the difference between the current frame and background. Similar approaches are used in existing techniques [13]. The input image where the skeleton is thought to be can be sampled to initialise the appearance model.

4 Inference

Given the model defined in the previous sections to infer the most likely model parameters we need to compute the MAP estimate

$$\{B^*_{1:T}, O^*_{1:T}\} = \underset{B_{1:T}, O_{1:T}}{\arg\max}\, P\left(B_{1:T}, O_{1:T}|Y_{1:T}\right) \quad (3)$$

On real data, we are only interested in the behavioural parameters so our MAP estimate is redefined as

$$\begin{aligned} B^*_{1:T} &= \underset{B_{1:T}}{\arg\max}\, p\left(B_{1:T}|Y_{1:T}\right) \\ &= \underset{B_{1:T}}{\arg\max} \int dO_{1:T} p\left(B_{1:T}, O_{1:T}|Y_{1:T}\right). \end{aligned} \quad (4)$$

Therefore if both the behaviour and observation parameters are of interested the MAP estimate is defined by (3), otherwise if only the behaviour is of interest the MAP estimate becomes (4).

A Metropolis-Hastings Markov Chain Monte Carlo algorithm has been implemented to obtain the MAP estimates. A new sample generated from the Markov Chain is accepted by the acceptance criteria defined by the Metropolis Hastings algorithm. Let the entire parameter space be denoted by $\theta = \{B_{1:T}, O_{1:T}\}$. The acceptance criteria is defined as

$$\alpha\left(\theta^{(n)} \rightarrow \theta'\right) = \min\left(1, \frac{P\left(\theta'\right) Q\left(\theta' \rightarrow \theta^{(n)}\right)}{P\left(\theta^{(n)}\right) Q\left(\theta^{(n)} \rightarrow \theta'\right)}\right)$$

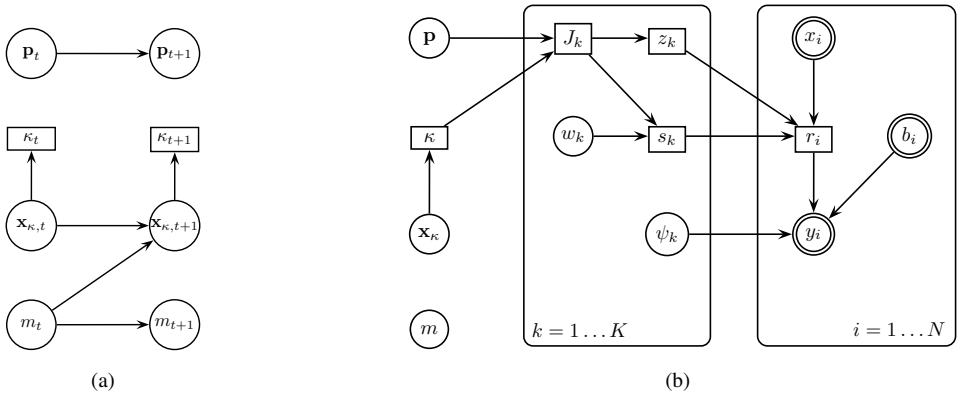


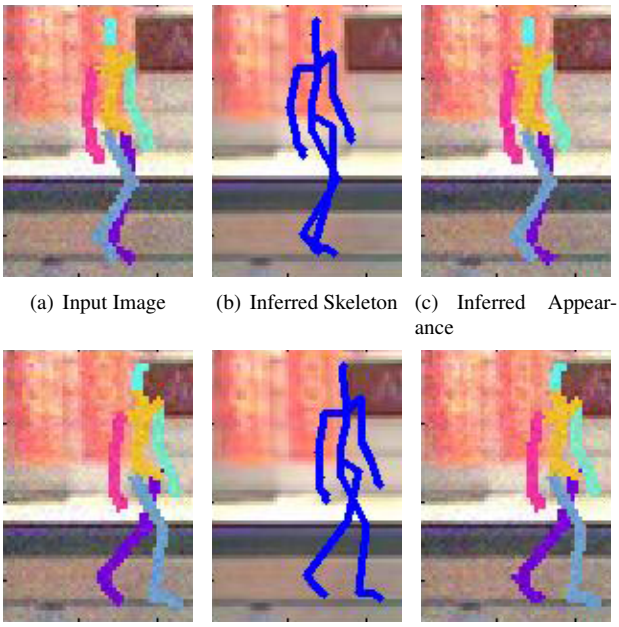
Figure 4. (a) shows a dynamic model of parameters describing behavioural properties of the subject. (b) shows the graphical model relating the behavioural parameters to the data observed in the image sequence. As the behaviour varies with time there is an implicit time subscript on each observation parameter.

where Q denotes the transition probability and P denotes the likelihood of the parameters. If the new sample θ' is accepted it forms the next sample in the Markov chain $\theta^{(n+1)} = \theta'$. To ensure the space is explored efficiently the acceptance of new proposals is controlled by annealing. To obtain the MAP estimate the parameters that achieve the maximum likelihood are stored.

5 Results

To demonstrate the correctness of the model we have generated a series of 10 frames of data obtained by by sampling the model. Figure 5 show the input image, inferred skeleton pose and the observation parameters respectively. The log likelihood at each sampling step is shown in Figure 6. There is a jump in the likelihood value at iteration $10,000$ as the variance in the appearance model is decreased. For the first half of the estimation process the position and skeleton pose is given a higher priority through the elevated appearance variance.

The front view of the image sequence 'threepastshop2' available in the Caviar data [5] is used to perform initial analysis on real data. The advantage of using this data set is the pre-labeled ground truth values. The algorithm was initialised with the background model, skeleton appearance model and the initial skeleton location. From the 5 frame image sequence we estimate the initial pose and the evolution of the skeleton position and pose. The evolution of poses is obtained from our learnt SLDS. Note that the motion used to train the SLDS is independent to the motion contained in the images. A restricted set of parameters is considered to reduce computation time. The results of the inference are shown in Figure 7. The image sequence contains a foreground railing introducing an error in our inference model. This unmodelled effect appears to have a negligible impact on tracking performance.



(a) Input Image (b) Inferred Skeleton (c) Inferred Appearance

(d) Input Image (e) Inferred Skeleton (f) Inferred Appearance

Figure 5. Figures (a), (b) and (c) are obtained from frame 5 of the synthetic data set and Figures (d), (e) and (f) are from frame 10. (a) and (d) are the input frames. The inferred skeleton position and pose are shown in (b) and (e). A sample of all inferred model parameters is shown in (c) and (f).

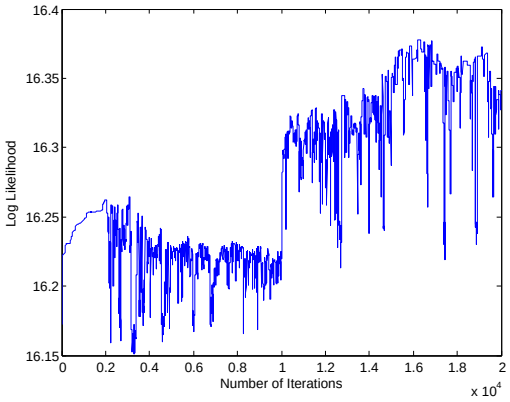


Figure 6. Log likelihood of MCMC samples for estimating parameters of simulated data.

(a) (b) (c) (d)

Figure 7. Figures (a) and (c) show input frames 1 and 5. Figures (b) and (d) show the inferred skeleton pose and position at frames 1 and 5 respectively. It is observed that the position of the back leg in (d) has lost track while the remaining body joint positions are accurate. It is anticipated that if subsequent frames are added to the inference algorithm this error would be corrected.

The execution time of the Metropolis-Hastings algorithm to infer the behavioural parameters of the 5 real data frames is 40 minutes for $15,000$ iterations. The bottleneck for computation is the estimation of the observation parameters. To reduce computational load, structural relationships between behavioural parameters to the observed images (see Figure 4(b)) can be exploited. The required integration over the observation parameters O_t can be achieved approximately with fast deterministic methods (such as variational techniques [12]), thus reducing the structure of the model effectively to a hidden Markov model where the behavioural parameters correspond to the latent states (see Figure 3).

6 Conclusions and Future Work

In this paper we have presented a novel graphical model to infer human behaviour from a sequence of images. Behaviours are encapsulated by a switching linear dynamic system. The observation model has an implicit method for describing occlusion. The model has been demonstrated on synthetic and real data. To extent this model we intend to investigate more efficient inference schemes and expand the range of behaviours that the system interprets.

7 Acknowledgements

David Excell would like to thank the support given by his funding bodies, the Cambridge Commonwealth Trust, the Trinity Hall Brookhouse Scholarship and the John Crampton Travelling Scholarship.

References

[1] A. Bissacco, "Modeling and learning contact dynamics in human motion", In *CVPR*, San Diego, June 2005.

[2] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds", In *IEEE Computer Vision and Pattern Recognition*, 2006.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid bbjects using mean shift", In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR'00)*, volume 2, pp. 142–149, South Carolina, 2000.

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition", *International Journal of Computer Vision*, 61(1), January 2005, pp. 55–79.

[5] R. Fisher, Caviar test case scenarios, Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/, January 2004.

[6] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization", *IEICE Trans. Inf & Syst.*, E87-D(1), January 2004, pp. 113–120.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[8] M. W. Lee and I. Cohen, "A model-based approach for estimating human 3d poses in static images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6), June 2006, pp. 905–916.

[9] M. W. Lee and R. Nevatia, "Dynamic human pose estimation using markov chain monte carlo approach", In *Proceedings of the IEEE Workshop on Montion and Video Computing (WACV/Motion'05)*, volume 2, pp. 168–175, 2005.

[10] V. Pavolvic, J. M. Rehg, T. Cham, and K. P. Murphy, "A dynamic bayesian network approach to figure tracking using learned dynamic models", In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pp. 94–101, Kerkyra, Greece, September 1999.

[11] P. van Overschee and B. D. Moor, *Subspace Identification for Linear Systems*, Kluwer Academic Publishers, 1996.

[12] M. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference", Technical Report 649, Department of Statistics, UC Berkeley, September 2003.

[13] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), September 2004, pp. 1208–1221.