Using Temporal Coherence for Gait Pose Estimation From a Monocular Camera View

Song Hu and Bernard F. Buxton
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK
{s.hu, b.buxton}@cs.ucl.ac.uk

Abstract

Predication of a pedestrian's 3D pose using movie sequences is not straightforward because of the human body's articulation and the complexity of biped locomotion. In this paper we show how temporal coherence can be employed to help the reconstruction of gait pose using corresponding silhouette moments of a video sequence from a monocular view. Virtual avatars were used to train the model for inferring the pose. The system was tested on a walking real person's silhouettes and evaluated using different avatars that were not in the training set.

1 Introduction

In recent years, computer vision researchers have been interested in tracking people's 3D pose from monocular camera views [1]. Such a monocular tracking system is less expensive and easier to set-up than multi-camera 3D tracking systems [9]. However, inferring the pose of the subject is not easy because of the human body's articulation and the complexity of the human body's movement.

Bowden *et al* have attempted to achieve 2D to 3D mapping by combining 2D and 3D data in a single model [1] . In their approach, 2D landmarks (400 points) were labelled on the person's silhouette contour in each frame to represent the shape of the moving person through an image sequence. They then concatenated the 2D shape with corresponding 3D body structure and positions of the hands and face to build a non-linear model similar to a local linear embedding [11]. By applying such a model to a given image the corresponding 3D structure could be reconstructed. This work showed that 2D and 3D information can be combined to provide a mapping between them. However, in such an approach, the 3D structure reconstructions were performed on each example separately without introducing the constraint of temporal coherence.

Grauman *et al* [5] have introduced a similar approach to infer 3D structure using an image-based shape model. They concatenated the 2D shapes in images obtained from calibrated cameras and the corresponding 3D locations of key joints by means of a mixture of probabilistic principal components analyzers (PPCA) [13]. A mean joint location error of 3 cm was achieved while using four cameras. They have also tested their method by training and testing the model with single views. However, when only one input view was given the performance dropped to an accuracy of 5 cm on average, which corresponds

BMVC 2005 doi:10.5244/C.19.48

approximately to 7 degrees of rotation error if we assume the average distance between key joints is 40 cm. As in [1], Grauman *et al* treated the examples at each time separately.

In practice there are temporal relationships between adjacent examples of a pedestrian's pose and patterns for certain movements such as walking, running and jumping can be learned during a training process. Thus, we can and should consider introducing temporal constraints to make the 2D to 3D mapping more accurate and specific for these kind of movements. We have used this approach to develop a method that employs an extended Kalman filter (EKF) [2, 14] to estimate a person's 3D pose using both the information from the silhouette obtained from a single camera and the characteristics of the movement pattern (e.g. walking).

2 Data representation and preparation

We represented 3D pose by the rotation of the key joints of an avatar, such as the knees, hips, elbows, shoulders etc as in Biovision's BVH format [12]. Within this approach, unlike others that use 3D joint locations (e.g. [5]), the 3D pose may be applied to other objects which have different physiques from that of the object being tracked. If we suppose there are L key joints, then the column vector

$$s = [x_1, y_1, z_1, \cdots x_i, y_i, z_i, \cdots x_L, y_L, z_L]^T, \tag{1}$$

in which x_i, y_i, z_i stand for the rotations of the i^{th} joint around the X, Y, Z axes, will represent the 3D pose. To parameterise the silhouette, instead of using contour points and landmarks as in [1, 5], we used normalised central moments η_{pq} [7], which are invariant to image rotation and approximately invariant to changes in viewing distance. This is an attractive option as the moments are easily computable in real-time from the silhouette contour and the silhouette itself may similarly be obtained from the image on an ordinary, up to date, desktop workstation [6]. Moreover, the moments are not dependent on the presence of particular landmark points which may sometimes be obscured and, by focusing first on the low-orders, moments can be used in a way that progressively introduces more detail of an object's shape. By definition, the zero order normalised moment is one and the first order central moments vanish. We therefore used moments of order $2 \leq p+q \leq l$ with $l = 5$ (i.e. up to fifth order) to represent the shape of the silhouette in the image by means of the $(l+4)(l-1)/2$ dimension column vector

$$m = [\eta_{20}, \eta_{11}, \eta_{02}, \cdots, \eta_{0l}]^T. \tag{2}$$

The vectors s and m are not defined in the same space and are not of similar scale. Principal component analysis (PCA) [8] was therefore applied to both data sets:

$$s = \bar{s} + P_s b_s, \quad m = \overline{m} + P_m b_m, \tag{3}$$

where b_s, $\bar{s}$ and P_s are respectively the weight parameters, mean vector and matrix of principal components of the 3D pose data set, and b_m, $\overline{m}$ and P_m are the weight parameters, mean vector and matrix of principal components of the silhouette moments data set. The matrices P_s and P_m were respectively chosen to contain the first t_s and t_m eigenvectors in each space, so as to explain a fraction f of the respective total variations. Typically, $f = 0.90, 0.95, 0.98$ or 0.99. Given the weights b_s and b_m for each training example, we

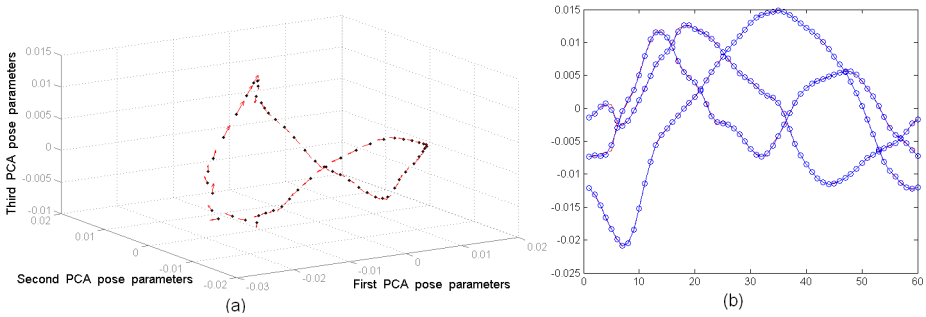


Figure 1: (a) A single cycle of the trajectory of the whitened pose PCA parameters of walking. The figure shows the first three principal components which describe, on average, 83.5% of the total variation. (b) Trajectories of the first three whitened pose PCA parameters of a walking avatar compared with their approximations using equation 4 when $\Gamma = 15$. From top to bottom at zero phase, $\acute{b_s}(2)$, $\acute{b_s}(3)$ and $\acute{b_s}(1)$, the open circles show values of the PCA components, the lines show the truncated Fourier approximations.

balanced them by whitening as suggested by Cootes [3] and used the scaled, whitened weights $\acute{b_s}$ and $\acute{b_m}$ to represent the 3D pose and the corresponding silhouette. Training data obtained from animation of the avatar implicitly enabled us to construct the joint distribution $p(s,m)$. In practice, this was characterised by the means $\bar{s}$ and $\bar{m}$ and the covariance, which was calculated in the combined space of the vectors $\acute{b_s}$ and $\acute{b_m}$.

3 Learning the movement pattern

Owing to the fact that people's movements such as walking, running and jumping are mainly composed of regular, repetitive motions, these movements present obvious periodic features. Such a periodic movement pattern can be seen from the trajectories of the pose PCA parameters of one walking cycle shown in figure 1. Thus, we can learn such a movement pattern from one selected period.

Practically, in our approach, we used Fourier expansions on one interval $[0, 2T]$, which corresponds to one walking cycle, of the pose PCA parameters to describe the movement pattern trajectory by means of the functions:

$$\acute{b_s}(k,\varphi) = \frac{A_0(k)}{2} + \sum_{n=1}^{\Gamma} A_n(k)\cos(n\pi\varphi/T) + \sum_{n=1}^{\Gamma} B_n(k)\sin(n\pi\varphi/T). \qquad (4)$$

Here $A_0(k)$, $A_n(k)$ and $B_n(k)$ are the Fourier coefficients of the k^{th} PCA pose parameter, T is the number of training samples taken for one stride, φ is the phase of the movement cycle, and $n = 1, 2, 3, \cdots, \Gamma$. Figure 1 shows the trajectories of the first three pose PCA parameters of a walking avatar and also the approximations to them obtained from equation 4 when $\Gamma = 15$. These truncated Fourier series approximations to the 3D pose PCA parameters had a RMSE of 2.55×10^{-8} in PCA space and a mean joint orientation difference of 1.02 degrees when projected back into the original pose space. The avatar's walking

movement was created using POSER [4] and the orientation differences were measured over 19 key joints.

4 Initial estimation of gait pose given a silhouette

If we assume the joint distribution of the whitened weights b_s' and b_m' are Gaussian, then the conditional density $p(b_s' | b_m')$, which defines the distribution of the b_s' given b_m' is also Gaussian and can be denoted as $p(b_s' | b_m') = N_{t_s}(\overline{b_{s|m}'}, C_{s|m})$. Moreover, the mean $\overline{b_{s|m}'}$ and covariance $C_{s|m}$ of the conditional density are given by:

$$\overline{b_{s|m}'} = C_{s,m} C_m^{-1} b_m' \quad and \quad C_{s|m} = C_s - C_{s,m} C_m^{-1} C_{m,s} \,, \tag{5}$$

where C_m^{-1} is the inverse covariance matrix of the b_m' and $C_{s,m}$ is the $t_s \times t_m$ cross-covariance matrix of the b_s' and the b_m'. In practice, the matrices C_m^{-1} and $C_{s,m}$ are constructed using vectors b_s' and b_m', which are computed from training examples. It is important to note that, according to equations 5 and 3, the conditional mean $\overline{b_{s|m}'}$ is a function of the silhouette vector m. Given a new example silhouette, the most likely corresponding 3D pose weights b_s' may therefore be estimated as the mean of the conditional density $p(b_s' | b_m')$.

5 Construction of the EKF

We now use equation 4 as the description of the trajectories of the PCA pose parameters of a person's movement and the prediction of the pose obtained from the silhouette by use of the posterior Gaussian density model as described in section 4 to construct an extended Kalman filter (EKF).

This EKF enables us to estimate the person's 3D pose using both the silhouette obtained from a single camera and the characteristics of the movement pattern, such as walking. Used together in this way, the EKF and the pose trajectory model enable us to exploit the temporal coherence of the movement.

In the EKF, the state variables are the phase φ and the velocity of the phase $\dot{\varphi}$. The φ indicates the object's position in a walking cycle and the $\dot{\varphi}$ indicates how fast the object walks. The state evolution equations were assumed to be:

$$\begin{bmatrix} \varphi_r \\ \dot{\varphi}_r \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varphi_{r-1} \\ \dot{\varphi}_{r-1} \end{bmatrix} + \begin{bmatrix} 0 \\ w \end{bmatrix}, \tag{6}$$

where w is zero mean random noise. The subscripts r and $r-1$ represent the current and previous sampling times respectively. The measurements were taken to be the person's PCA pose parameters $b_s'(k, \varphi_r)$ as obtained from the observed silhouette at time r corresponding to phase φ_r as discussed in section 4. The measurement equations of the EKF are thus represented by equations 4 (to each of which noise $v(k)$ say, should be added) and need to be linearized for example, as described in [14]. This requires the Jacobian matrices H and V which are the partial derivatives of the measurement equations with respect to $[\varphi_r, \dot{\varphi}_r]^T$ and v respectively. In our case V is a diagonal $t_s \times t_s$ matrix whilst H

Figure 2: (a) Examples of avatars used for training and evaluation. (b)Examples of walking people's silhouettes from the *Southampton Human ID at a Distance* database

is the $t_s \times 2$ matrix:

$$H = \frac{n\pi}{T} \begin{bmatrix} \sum_{n=1}^{\Gamma} B_n(1)\cos(n\pi\varphi/T) - \sum_{n=1}^{\Gamma} A_n(1)\sin(n\pi\varphi/T) & 0 \\ \sum_{n=1}^{\Gamma} B_n(2)\cos(n\pi\varphi/T) - \sum_{n=1}^{\Gamma} A_n(2)\sin(n\pi\varphi/T) & 0 \\ \vdots & \vdots \\ \sum_{n=1}^{\Gamma} B_n(t_s)\cos(n\pi\varphi/T) - \sum_{n=1}^{\Gamma} A_n(t_s)\sin(n\pi\varphi/T) & 0 \end{bmatrix} \tag{7}$$

evaluated at $\varphi = \varphi_r$.

When new silhouettes of a walking person are given, initial estimates of the gait pose parameters (in PCA space) are calculated as discussed in section 4. These initial parameters were then used as measurements in of our EKF and the walking person's current state variables φ and $\dot{\varphi}$ were estimated. Given the phase φ we could then reconstruct the gait pose by using equations 4 and 3.

6 Experiments and evaluation

Our approach was tested on a walking movement. We used a number of avatars with the same walking pattern but different physiques, hair styles and clothes to train the model as described in sections 2 and 3. Avatars, which had a slightly different gait and which were not used in the training were used to evaluate the system's performance. We also tested our method on a real walking person's silhouette from the *Southampton Human ID at a Distance* database [10].

During the training process, we placed the virtual camera in front of the training avatars and made the avatars walk from right to left, which was the same as for the real silhouette examples from the image database as shown in figure 2. Each training avatar performed one cycle of walking and 60 frames, together with their corresponding 3D pose information, were collected as the training data set for that avatar. All the training data sets from different training avatars were then concatenated and used to calculate the inverse of the covariance matrix C_m^{-1} and the covariance $C_{s,m}$ described in section 4. These were then used in equation 5 for initial estimation of gait pose when given a new example silhouette.

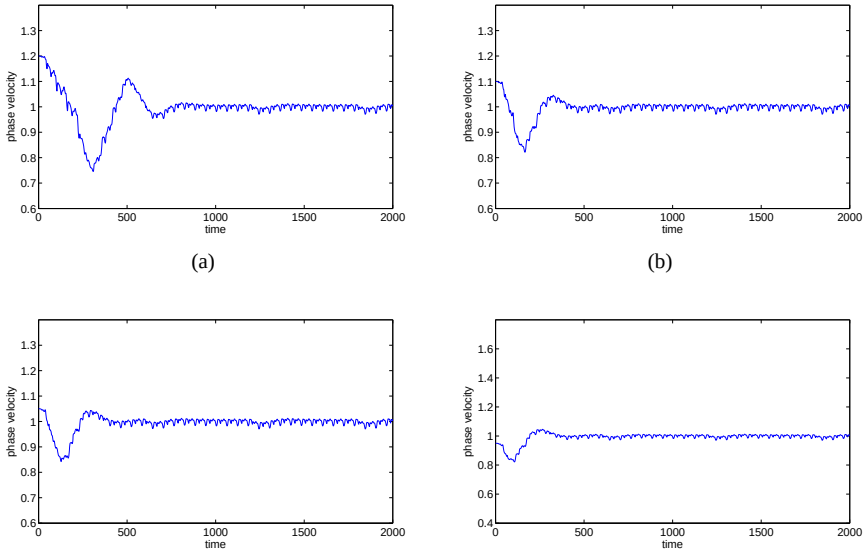


Figure 3: The estimated phase velocity $\dot{\varphi}$ when it was initialized to 1.2, 1.1, 1.05 and 0.95 for (a), (b), (c) and (d) respectively.

6.1 Evaluation using testing avatars

Because the real walking data lacks 3D pose information, the evaluation of our approach's accuracy of gait pose reconstruction was carried out using virtual avatars. In order to test the robustness of our method, we used a different walking pattern from that used in training by introducing a different stride length. Since we sampled 60 examples within each walking cycle, the true period $2T$, velocity of the phase $\dot{\varphi}$ and initial phase φ were 60, 1, and 0 respectively. In figure 3, we show for one testing avatar the result obtained when we set the initial phase to 12 and the initial phase velocity to 0.95, 1.05, 1.1 and 1.2 respectively. It can be seen that the phase velocity converges to the correct value of 1 for all four cases. The approach to the correct velocity was more rapid when the initial estimate of $\dot{\varphi}$ was set closer to 1 although in all four cases there was an initial undershoot/overshoot and subsequent oscillation.

As a measure of accuracy, we took one walking cycle from a testing avatar and compared the PCA pose parameters obtained from the EKF, computed using equation 6, with corresponding measurement (i.e. the initial conditional mean estimates), and ground truth parameters. The initial state was set as $\varphi = 12$, $\dot{\varphi} = 1.2$ and the PCA parameters were compared after the state variable $\dot{\varphi}$ became stable after approximately $t = 800$. From figure 4 we can see that pose parameters estimated from the EKF are much closer to the ground truth than the frame-by-frame initial measurement values obtained from the conditional mean estimates. It is particularly notable that the system uses the relationship between the PCA pose parameters to correct the estimates of the second component (figure 4(b)).

The difference between the angles of key joints in the reconstructed pose and their

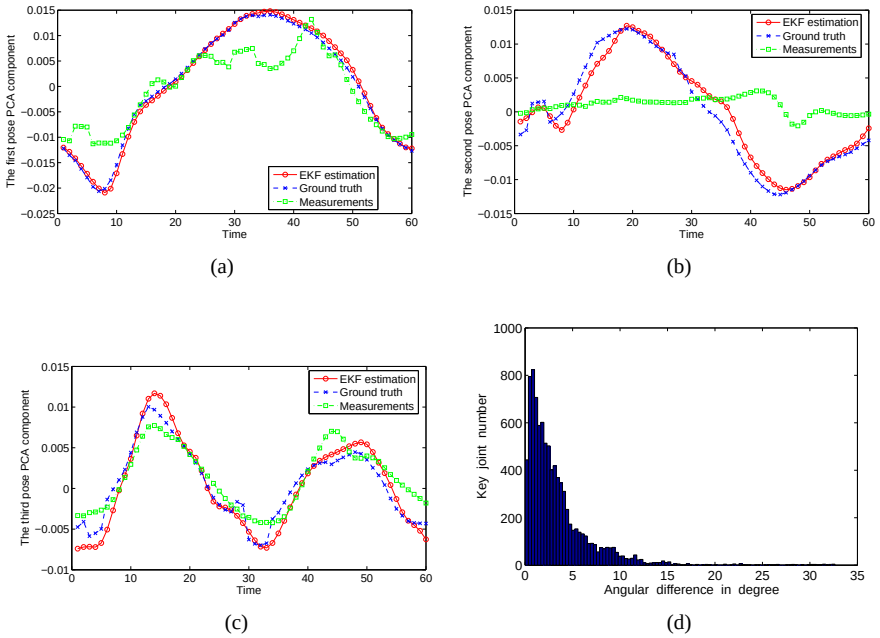
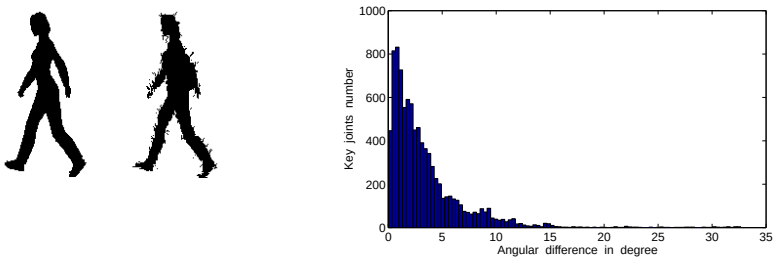


Figure 4: The pose PCA parameters estimated from the EKF compared with ground truth parameters and with the parameters predicted directly by the measurement equation. (a), (b) and (c) are for the first, second and third PCA components respectively. (d) is the histogram of angular differences $\theta(i)$ and the mean is 3.38 degrees.

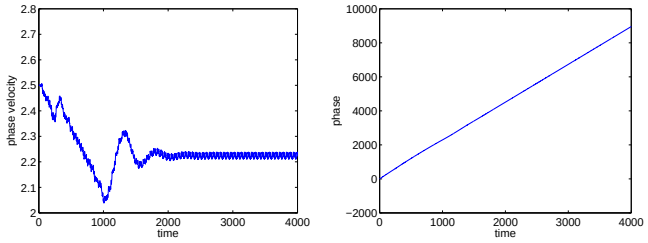
ground truth values were also used to assess the system's performance. To do so, we represented the estimated and ground truth orientations of each key joint i as orthogonal matrices $q_r(i)$ and $q_t(i)$ respectively and, by solving $q_t(i) = q_r(i) \cdot q_e(i)$ we obtained the matrix $q_e(i)$ that would rotate $q_r(i)$ into $q_t(i)$. The angular difference $\theta(i)$ between the estimated and ground truth rotations for each key joint may then be obtained from the fact that $tr(q_e(i)) = 1 + 2cos(\theta(i))$. For each test avatar, once the EKF had settled down, we took one walking cycle (60 examples) and measured the $\theta(i)$ over the 19 key joints. The result are shown as a histogram in figure 4(d). The mean value of the difference is 3.38 degrees.

The experiments discussed above were carried out under noise free conditions. In order to evaluate the system's performance in a noisy situation, we introduced noise on the test avatar's silhouettes. The noise was added randomly (even though, in practice, noise on the contour at neighbouring pixels would be correlated) to the avatar's silhouette contour along the normal of each contour pixel. A noise level of 10 pixels means we randomly generated a number between -10 and 10. If the random number obtained was positive, we added this number of noise pixels outside the silhouette contour along the direction of the contour's normal, while if the random number was negative, we similarly added noise pixels inside the silhouette. The mean value of the $\theta(i)$ when the above experiment was repeated on noisy silhouettes is 3.37 degrees, which is close to the accuracy obtained for noise-free silhouettes. This indicates that noise on the silhouette contour will not affect



(a) (b)

Figure 5: (a) The image on the left shows a noise-free silhouette, the image on the right shows the silhouette with a noise level of 10 pixels. (b) The histogram of angular differences $\theta(i)$. The mean is 3.37 degrees.



(a) (b)

Figure 6: Estimated phase velocity (a) and phase (b) for real silhouettes. The initial phase was set as 0 and the initial phase velocity was 2.5.

the system's performance too adversely.

6.2 Testing on real images

The Southampton database provides walking people silhouettes which we sampled at approximately 30 frames per walking cycle. Inspection of the data showed that the actual phase velocity $\dot{\varphi}$ for the walking people was about 2.

Figure 6 shows the estimated phase velocity for one set of real images from which it can be seen that $\dot{\varphi}$ converged to about 2.23 and that the estimated phase φ is approximately linear as it should be with little drift. Figure 7 shows some examples of the silhouettes from the database and the corresponding estimated pose of the avatar used. Although we cannot assess the pose estimation errors in this case, we measure the accuracy of the match quality between the real silhouettes, represented by the binary masks S_s, and the silhouettes S_e, similarly represented, obtained by projection of the avatar. The match quality was defined as

$$\frac{2 \times (S_s \cup S_e - S_s \cap S_e)}{(S \cup S_e + S_s \cap S_e)}, \tag{8}$$

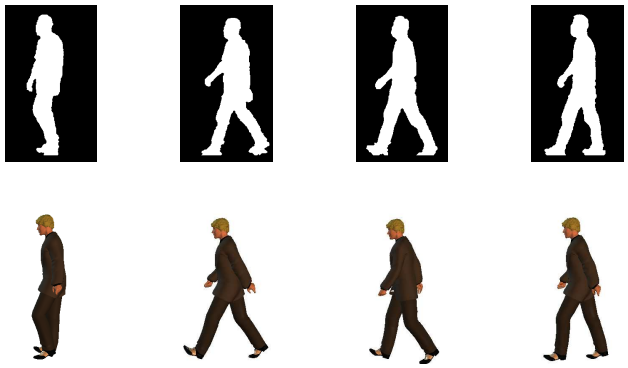


Figure 7: Real silhouettes (first row) and corresponding avatar poses (second row)

for each frame in the sequence. An average match quality of 0.34 was obtained for 57 real silhouettes. For comparison, we also computed the silhouette match quality from the experiments in which the test avatars were used. An average match quality of 0.15 was obtained.

7 Conclusion and future work

In this paper we introduced an avatar modelling and Kalman filtering approach to the estimation of gait from monocular imagery. Initial estimates of a subject's pose were obtained by constructing a posterior density of the pose given a silhouette. Temporal coherence was introduced to improve the initial estimation by learning the gait characteristics and employing an EKF which took the initial pose estimates from the posterior conditional mean noisy measurements. The posterior density and gait characteristics were obtained by using avatars and a walk animation script [4] to provide training data. The system's performance was evaluated by using a test avatar different from those used in the training and experiments carried out on real data obtained from the silhouettes of walking people in the *Southampton Human ID at a Distance* database. Satisfactory results were obtained in the experiments on both the test avatars and on real data. Once the EKF had settled down, the former was accurate, for the key joints, to a mean rotation angular error of approximately 3.5 degrees with either clean test silhouettes or silhouettes to which noise had been added. Analysis of the silhouettes indicates that the fit to the real data in the experiments on the Southampton database is of a comparable accuracy to that obtained in the test experiments when virtual data was used.

8 ACKNOWLEDGMENT

The authors would like to thank the ISIS research group at the University of Southampton for providing access to the Automatic Gait Recognition for Human ID at a Distance database.

References

[1] R. Bowden, T.A. Mitchell, and M. Sarhadi. Reconstructing 3d pose and motion from a single camera view. In John N. Carter and Mark S. Nixon, editors, *In Proceedings of the British Machine Vision Conference*, volume 2, pages 904–913, University of Southampton, September 1998.

[2] Robert Grover Brown and Patrick Y.C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, New York, NY, USA, 2 edition, 1992.

[3] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, Manchester M13 9PT, U.K., March 2004.

[4] Inc. Curious Labs. Poser, 2000.

[5] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 641–648, Washington, DC, USA, 2003. IEEE Computer Society.

[6] S. Hu and B. F. Buxton. A real-time tracking system developed for an interactive stage performance. In *WEC'05*, Istanbul, Turkey, April 2005.

[7] Anil K. Jain. *Fundamentals of digital image processing*. Number 0-13-336165-9. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989. page 377-380.

[8] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1988.

[9] Simon G Penny, Jeffrey Smith, and Andre Bernhardt. Traces: Wireless full body tracking in the cave. In *Ninth International Conference on Artificial Reality and Telexistence (ICAT'99)*, December 1999.

[10] ISIS research group. Southampton human id at a distance database. [Online] Available at http://www.gait.ecs.soton.ac.uk/database/index.ph3, February 2004. (accessed 03 March, 2005).

[11] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. In *Science*, volume 290, pages 2323–2326, December 2000.

[12] J. Thingvold. Biovision bvh format. [Online] Available at http://www.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html, (accessed 03 March, 2005), 1999.

[13] Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[14] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.