Using an Avatar to Develop a System for the Predication of Human Body Pose from Moments

Song Hu and Bernard. F. Buxton

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Abstract. Tracking people using movie sequences is not straightforward because of the human body's articulation and the complexity of a person's movements. In this paper we show how a person's 3D pose can be reconstructed by using corresponding silhouettes of video sequences from a monocular view. Currently, a virtual avatar is used to train the model for inferring the pose and a different avatar is used to produce novel examples not in the training set in order to evaluate the approach. The approach was subsequently tested using the silhouettes of a walking person.

1 Introduction

In recent years, computer vision researchers have been interested in tracking people in video sequences since such a capability may enable a wide variety of applications in surveillance, entertainment, sports, computer games and even robotics. However, the task is not easy because of the human body's articulation and the complexity of a person's movements. Some work has been carried out to track a pedestrian's shape, for example in 2D [1]. However, instead of only tracking a person's 2D shape from the video frames, we aim to track a person's pose in 3D, which describes the movement more precisely than the 2D shapes do.

Bowden *et al* and Grauman *et al* have attempted to achieve 2D to 3D mapping by combining 2D and 3D data in single or mixture models [2,6] similar to local linear embeddings. In their approaches, 2D landmarks are labelled on the person's silhouette contour in each frame to represent the shape of the moving person through an image sequence (400 points are used in Bowden's work and 200 points are used in Grauman's work). This approach requires an accurate and reliable method to label each landmark at the same place of the silhouette contour otherwise the model cannot represent the changes of the moving person's shape reliably. The landmarks should also be located on important parts of the object, so that, for example, each anatomically important part (such as the hands and face) will be labelled with at least one landmark point to ensure that objects are modelled fully. Using more landmark points or even using the whole contour can reduce the possibility of missing important parts of the object and can also lower the requirements on landmark labelling accuracy. However, using

many landmark points is not ideal since it raises the dimension of the 2D data dramatically and does not provide, from contour data, a concomitant increase in the information available.

Thus, we have developed an approach based on global features of the silhouette contour such as its moments which, though they might be expected to be quite sensitive to noise and to details of the silhouette shape give us the benefit of a compact description and avoid the difficulties of building an accurate and reliable landmark labelling system. To achieve the goal of tracking 3D pose, we analyse the correlation between the silhouette's moments and the corresponding 3D pose of the body. To capture this correlation, we build a combined 2D and 3D statistical model, which can later be used to infer a moving person's 3D pose from a single video frame. In principle, the 3D pose may be inferred from 2D data by calculation of the posterior distribution from the combined 2D and 3D joint distribution. In practice, this is only straightforward when the distributions are Gaussian and the posterior may be obtained analytically. However, we also adjust the inferred 3D pose in order to optimise the reconstruction in case the distributions are significantly non-Gaussian or the moments unduly sensitive to noise and details of the silhouette shape, either of which could mean that the algebraic prediction is not accurate.

Training the model in order to capture the correlation between the 2D image and 3D pose requires access to both 2D image and 3D pose data. Such data could be provided by means of a specialised motion tracking system [4]. However, in the context of ordinary laboratory work and simple movements such as walking, it is more convenient to use data from an avatar to train the system. Use of an avatar gives full control over both its movement and of the virtual camera environment with the result that it is straightforward to obtain 2D image and corresponding 3D pose data. Our aim is thus to show: (i) that an avatar can be used to train such a system in this manner, (ii) that a few low-order nomalised central moments may be used to capture sufficient information from the silhouette shape for a first prediction of the 3D pose from the trained model, (iii) that the pose predicted in this manner may be refined and corrected by using the avatar model to match directly to the shape of the silhouette, thereby (iv) overcoming to a large extent both the potential sensitivity of the moment features and the limitations of the assumed Gaussian model. The accuracy of the pose reconstruction is evaluated from simulation experiments and that of the matching from both simulated and real data. We begin, however, in the next section, with a description of the data representation used.

2 Data representation and Gaussian model

We represent a person's 3D pose by the rotation of key joints, such as the root joint, which determines the balance and the orientation of the person, the knees, hips, elbows, shoulders *etc* as in Biovision's BVH format [5]. This approach, unlike others that represent 3D pose as a set of 3D joint locations (e.g. [6]), gives the potential of easily applying the estimated 3D pose to other objects,

which have different physique from the one 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 a person's 3D pose. To parameterise the silhouette, instead of using landmarks as in [2,6], we use normalised central moments η_{pq} [8] 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 on an ordinary, up to date, desktop workstation from the bounding contour of the silhouette, S_t, obtained, for example by thresholding of the image [7]. 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-order (e.g. 2^{nd}, 3^{rd}, 4^{th} and 5^{th}order) moments, can be introduced in a way that progressively introduces more detail of an object's shape. However, care must be taken in order to combat the known sensitivity of moments, even of comparatively low order, to noise and to details of the silhouette shape. This is one reason why, as described in section 3, after using the low-order moments to infer the 3D pose, we use the silhouette itself to correct and refine the pose estimates.

By definition, the zero order normalised moment is one and the first order central moments vanish. We therefore use moments of order $2 \leq p + q \leq l$ to represent the shape of the silhouette in the image by means of the $(l+4)(l-1)/2$ dimension column vector

$$m = [\eta_{20}, \eta_{11}, \eta_{02}, \cdots, \eta_{(l-1)}, \eta_{0l}]^T . \tag{2}$$

The vectors s and m are not in the same space and not of similar scale. Thus, principal component analysis (PCA) [9] is applied to both data sets:

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

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, $\bar{m}$ 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 are respectively chosen to contain the first t_s and t_m eigenvectors in each space, so as to explain a fraction f of the variation. Typically, $f = 90, 95, 98$ or 99%. Given the weights b_s and b_m for each training example, we can balance them by as suggested by Cootes [3] and use the scaled weights $b_s^{'}$ and $b_m^{'}$ after whitening to represent the 3D pose and the corresponding silhouette. Training data obtained from animation of the avatar implicitly enables us to construct the joint distribution $p(s, m)$. In practice, this is characterised by the means $\bar{s}$ and $\bar{m}$ and the covariance, which is calculated in the combined space of the vectors $b_s^{'}$ and $b_m^{'}$.

If we assume the joint distribution of the whitened weights $b_s^{'}$ and $b_m^{'}$ is Gaussian, then the conditional density $p(b_s^{'} \mid b_m^{'})$, which defines the distribution

of the $b_s^{'}$ given $b_m^{'}$ is also Gaussian. Moreover, the mean $\bar{b}_{s|m}^{'}$ and covariance $C_{s|m}$ of the conditional density are given by:

$$\bar{b}_{s|m}^{'} = C_{s,m}C_m^{-1}b_m^{'}\,, \qquad C_{s|m} = C_s - C_{s,m}C_m^{-1}C_{m,s}\,, \tag{4}$$

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. It is important to note that, according to 4, the conditional mean $\bar{b}_{s|m}^{'}$ is a function of the $b_m^{'}$. This means that, given a new example of whitened PCA weights $b_m^{'}$, the most likely corresponding 3D pose weights $b_s^{'}$ can be estimated as the mean of the conditional density $p(b_s^{'} \mid b_m^{'})$.

3 Adjusting the 3D pose

The system described in the preceding section uses the Gaussian approach to predict the 3D pose from image data. As we shall see from the results to be presented in section 4, such a system performs quite well, but owing to the assumption of Gaussian statistics, and the sensitivity of the moments to noise and details of the silhouette shape, is not always accurate. In this section, the pose estimates are refined and more accurate results obtained by using a search algorithm to adjust the initial pose estimates obtained from the Gaussian assumption to fit better to the observed image silhouette. Owing to the fact that the pose PCA parameter space $b_s^{'}$ usually has fewer dimensions than the original space of the joint angles, the search is carried out in the space $b_s^{'}$. Furthermore, in PCA, the eigenvectors of the matrix P_s are sorted with respect to the magnitudes of their variances, so successive weights contribute less to the manipulation of pose in equation 3. This encourages us to treat each pose PCA weight separately during the search process and starting with the most important weights will enable us to correct the largest errors first. Adjustment of the next weight then corrects for the next largest contribution to the remaining error and so on.

A one-dimensional golden section search [10] was carried out in this way to adjust the PCA pose weights. For each $b_s^{'}$ obtained from the conditional mean $\bar{b}_{s|m}^{'}$ as described in section 2, a pose is generated using equation 3 and applied to a virtual avatar that is similar to the moving person or target. The avatar's silhouette $S_a$is then obtained by projection onto a virtual plane. The accuracy of the match between S_a, regarded as a binary image with value one inside the silhouette and zero elsewhere, and the given target silhouette S_t, similarly encoded, was defined as:

$$D_S = \frac{2 \times (S_a \cup S_t - S_a \cap S_t)}{S_a \cup S_t + S_a \cap S_t}\,. \tag{5}$$

The search algorithm is iterative, so when we finish the golden section search on the last of the pose PCA weights, we return to the first and continue the pose adjustment until the system converges (i.e. the change in D_S is small enough). However, because at each point of the search the avatar has to be regenerated, the computation cost rises with the number of iterations.

4 Experiments and evaluation

Our approach was tested on a walking movement. An avatar was used to train the model as described in sections 2. Another avatar, which had a different physique was used to evaluate the system's performance. We also tested our method using the silhouettes of walking persons from the *Southampton Human ID at a Distance* database [11].

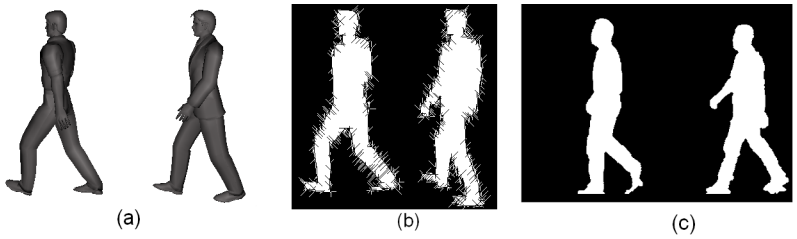


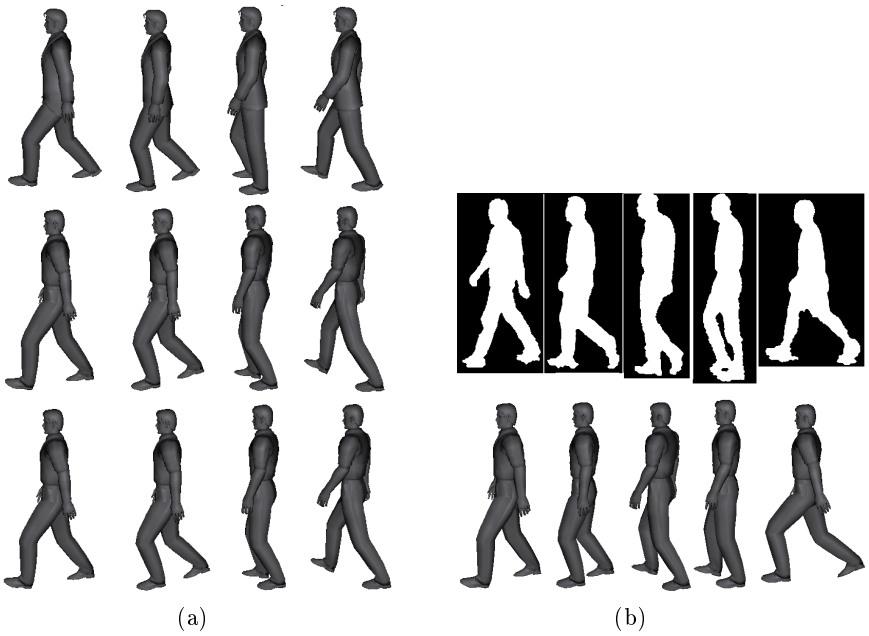
Fig. 1. (a) Examples of avatars used for training and evaluation. (b) Examples of avatar silhouettes with added random noise at a level of 12 pixels. (c) Examples silhouettes from the *Southampton Human ID at a Distance* database.

During the training process, we placed the virtual camera in front of the training avatar and made the avatar walk from right to left which was the same as for the real silhouette examples from the image database as shown in figure 1. Walking sequences were collected from different orientations at 60, 70, 80, 90, 100, 110 and 120 degrees from the front view (clockwise). The moments of the avatar's silhouettes, together with their corresponding 3D pose information, were collected as the training data set and were used to calculate the inverse of the covariance matrix C_m^{-1} and the covariance $C_{s,m}$ described in section 2. These matrices were then used in equation 4 for initial estimation of gait pose when given a new example silhouette. Adjustments to the 3D pose were then made as described in the section 3 to fine-tune the reconstructed pose by searching for a better match between the silhouette of the reconstructed avatar and the given new silhouette.

Because the data from the walking people lacks 3D pose information, evaluation of the accuracy of the 3D pose reconstruction was carried out using a virtual avatar. In order to test the robustness of our method, we used a different avatar from that used in training. The test avatar was constructed to perform movements similar to those of the avatar in the training set, while the orientation was set at different directions from those used in the training process (at $65, 75, 85, 95, 105$ and 115 degrees, respectively).

Noise was also introduced on the test silhouettes in order to evaluate the method's performance in a noisy situation. Figure 1 shows some examples of noisy silhouettes. The noise was added randomly to the avatar's silhouette contour along the normal of each contour pixel. A noise level of 12 pixels, as indicated

in figure 1 means we randomly generated a number between -12 and 12. 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.



(a) (b)

Fig. 2. (a) The first row at the top shows the ground truth test examples, the second row shows the reconstructed poses in a noise free situation, while the third row shows the reconstructed poses in a noisy situation. (b) Real silhouettes (first row) and corresponding reconstructed avatar poses (second row).

Figure 2 shows some examples of ground truth test poses and the reconstructions obtained as described in sections 2 and 3 in both noise free and noisy situations. There is little perceptible difference between the ground truth 3D poses and the reconstructions and it can be seen that adding noise on the test silhouette contour does not affect the reconstruction performance too adversely.

For each test example, the mean absolute difference (MAD) over 26 key joints between the angles of these key joints in the reconstructed pose and their ground truth values were used to assess the pose reconstruction performance. To do so, we represented the estimated and ground truth orientations of each key joint i as orthogronal 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 i may then be obtained from the fact that $tr(q_e(i)) = 1 + 2\cos(\theta(i))$.

As shown in figure 3, for both noisy and noise free situation, an accuracy of about 3 degrees mean absolute angular difference is achieved, which corresponds approximately to 2cm if we assume the average distance between key joints is 40 cm. It can also be seen that, although the 3D pose adjustment introduced extra errors for a few examples, the average performance is improved after the fine-tuning of the iterative search.

Our approach was also tested on real walking people's silhouettes provided by the *Southampton Human ID at a Distance* database. For testing on the real data, the virtual camera was set approximately the same as the real camera so that the real silhouette and the virtual avatar's silhouette were of the same scale. Some examples of the test on real silhouettes are shown in figure 2.

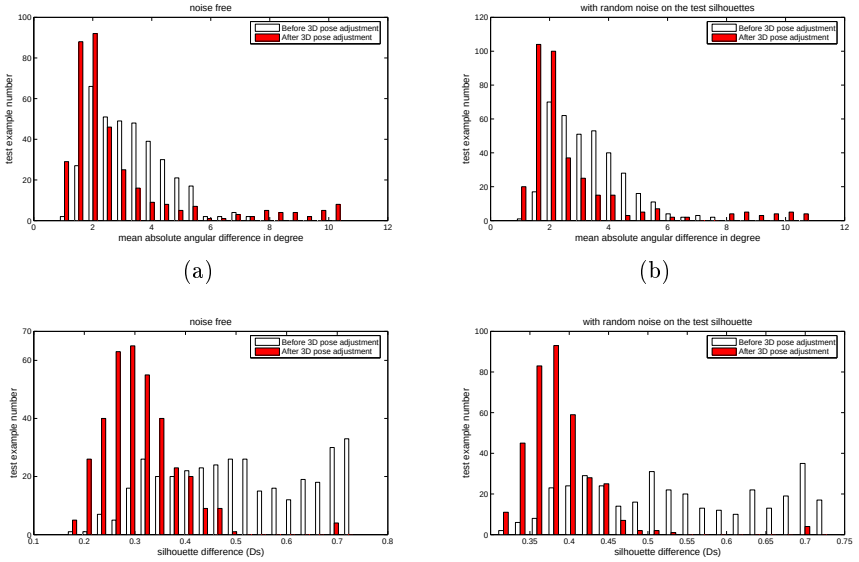


Fig. 3. (a) and (b) are the histograms of the angular MAD for noise free and noisy experiments respectively. Prior to the 3D pose adjustment the MAD is 3.23 degrees for noise free experiments and 3.28 for noisy ones, whilst after the 3D adjustment they are reduced to 2.87 degrees and 2.82 degrees respectively. (c) and (d) are the histograms of D_s for noise free and noisy experiments respectively. The mean of D_s is 0.5 in the noise free experiment and 0.53 in the noisy experiment prior to 3D pose adjustment, reduced to 0.31 and 0.39 respectively after 3D pose adjustment.

5 Conclusion and future work

As discussed in previous sections, by using the system described in section 2 we can estimate an avatar's 3D pose by using information from the avatar's

silhouette. The accuracy of the pose predicted from the conditional mean was acceptable (i.e to within a few degrees when a test avatar was used) and we were able to improve it by adjusting the initial estimated 3D pose to fit the silhouette. This reduces the average reconstruction errors and, as inspection of the sequence of reconstructed poses shows, the jitter. At present, there is no representation of temporal coherence in our model. However we intend to introduce temporal constraints in the future, for example, by use of a Kalman filter to eliminate the jitter and to make the model more accurate and specific.

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. A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. Technical Report 94.11, April 1994.
2. 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.
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. Ascension Technology Corporation. Real-time motion capture. [Online] Available at http://www.ascension-tech.com/products/motionstar.pdf, 2000. (accessed 03 March, 2005).
5. Inc. Curious Labs. Poser, 2000.
6. 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.
7. S. Hu and B. F. Buxton. A real-time tracking system developed for an interactive stage performance. In *WEC'05*, Istanbul, Turkey, April 2005.
8. 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.
9. I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1988.
10. William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing, 2nd edition.* Cambridge University Press, New York, NY, USA, 1992. page 397-402.
11. 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).