

# Statistical Personal Tracker

Song Hu and Bernard Buxton  
University College London

Abstract—Tracking people using movie sequences is not straightforward because the human body is articulated (and therefore far from rigid), parts of the body are frequently occluded by other parts, and video data is inevitably degraded by noise. In this paper we show how a person’s 3D pose can be tracked by using corresponding silhouette moments from video sequences. The moment computation has been implemented in real-time for real data and shown to work satisfactorily for data obtained in the laboratory or a dance studio or theatre. 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 this approach.

## 1. Introduction

In recent years, computer vision researchers have been interested in tracking people using movie sequences since such a capability brings a wide variety of applications in surveillance, entertainment, sports, computer games and even robotics.

Some work has already been done to track a pedestrian’s shape, for example in 2D[1]. However, instead of only tracking a person’s 2D shape from the image frames, our aim is to track a person’s pose in 3D, which describes the movement more precisely than the 2D shapes do. To achieve the goal of tracking 3D pose, we will analyse the correlation between the silhouettes obtained from the video frames and the corresponding 3D pose of the body, by building a combined 2D and 3D statistical model, which can be used to track moving people’s pose in 3D from only a sequence of monocular images from a single, static conventional camera.

Bowden and Kristen have attempted to achieve 2D to 3D mapping by combining 2D and 3D data in single or mixture models [2,4] 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). This approach required 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 people’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. Also, 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 amount of information. Using our new approach based on global features of the silhouette contour such as the moments we benefit from the compactness of the moment description and avoid the difficulty of building an accurate and reliable landmark labelling system.

It is convenient to use an avatar moving around in a virtual environment to train the statistical model because we have full control over both the avatar and the virtual camera environment, which means we can obtain 2D and 3D data easily.

## 2. Methods

Suppose we have two random vectors  $x$ ,  $y$  of dimensions  $m$  and  $n$  respectively. Assume the  $(m+n)$  dimensional vector  $z^T = (x^T, y^T)$  belong to a joint Gaussian distribution. The conditional density  $p_{y|x}$  is also Gaussian. The mean of the conditional density  $p_{y|x}$  is  $m_{y|x} = m_y + C_{yx}C_x^{-1}(x - m_x)$ , where  $C_{yx}$  is cross-covariance matrix of the vectors  $x$ ,  $y$  and  $C_x^{-1}$  is the inverse covariance matrix of  $x$  [3]. Thus if we concatenate the 2D data set  $X$  and 3D data set

Y into one vector  $Z^T = (X^T, Y^T)$  and assume Z has a Gaussian distribution, given a new example of 2D vector X', the corresponding 3D vector Y' can be approximately estimated as the mean of the conditional density  $\text{pylx}'$ . This superior to solving a set of linear equations for the PCA weights of the Y given the X, as one might as first sight do.

### 3. Implementation and Results

To test the method above, a virtual avatar that waves its arm and leg is used to train the model as shown in Fig.1.a. The viewpoint is set in front of the avatar and 5 different orientations are set during the training process (0, 10, 20, 30, 40 and 50 degree from the frontal view). The avatar's joint rotation angles (around x, y, z axis) represent the pose (3D) parameter set and the avatar's silhouette moments (up to 4<sup>th</sup> order) represent the 2D data set. Then a different avatar (Fig.1.b) performing similar movements is used to produce a monocular video sequence in different orientations (at 5, 15 and 45 degrees, respectively). 3D postures are estimated using the method discussed in section 2 according to the video sequence. The results are compared with the ground truth pose (male avatar) in Fig1.(c-k).

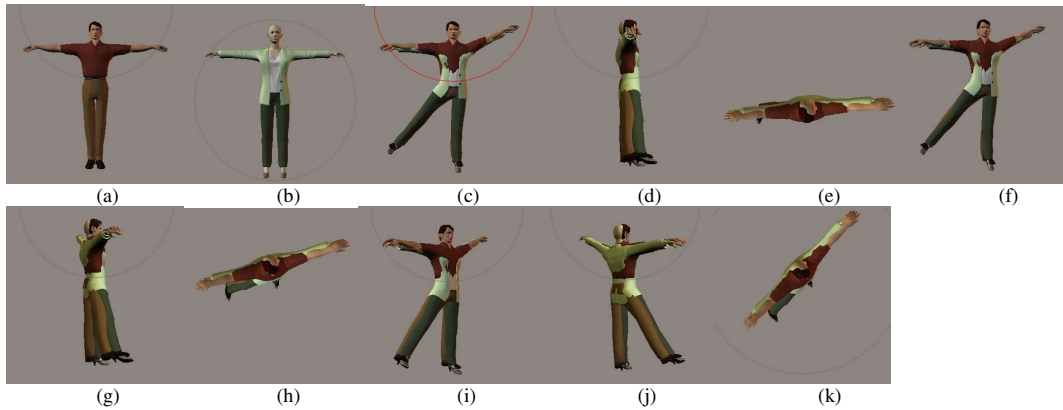


Fig 1

### 4. Conclusion and Future work

We have discussed a method to infer 3D pose from 2D silhouette moments. As shown in Fig1.(c-k), good pose estimation can be obtained by using only up to 4<sup>th</sup> order moments. The method has been shown quantitatively to be superior to the naïve solution indicated at the end of section 2. In future work, we will investigate the effect of increasing the maximum order of moments that used to represent the 2D information and combine this system with our real-time implementation of the moment computation in order to infer the pose of real people, such as modern dance performers.

#### References:

- [1] A M Baumberg & D C Hogg, *An Efficient Method for Contour Tracking using Active Shape Models*, Workshop on Motion of Non-Rigid and Articulated Objects, Austin, Texas, IEEE. 1994.
- [2] R Bowden, T A Mitchell and M Sarhadi, *Reconstructing 3D Pose and Motion from a Singel Camera View*, In British Machine Vision Conference 1998, volume 2, pages 904-913, 1998.
- [3] A Hyvarinen, J Karhunen and E Oja, *Independent Component Analysis*, pages, 32-33, 2001, ISBN 0-471-40540-X
- [4] K Grauman, G Shakhnarovich and T Darrell, *Inferring 3D Structure with a Statistical Image-Based Shape Model*. In Proceedings IEEE International Conference on Computer Vision, Nice, France, October 2003