MARKERLESS MOTION CAPTURE WITH SINGLE AND MULTIPLE CAMERAS

Paris Kaimakis and Joan Lasenby

University of Cambridge
Department of Engineering
Trumpington Street, CB2 1PZ
United Kingdom
{pk228, jl}@eng.cam.ac.uk

ABSTRACT

The aim of Optical Motion Capture is to sequentially estimate the true state $\mathbf{X}^*$ of the subject (generally an articulated body) at any time instant t_k from a set of data D_k, captured by N calibrated cameras each of resolution $U \times V$ pixels. Our aim is to achieve this without the need for markers. This is to enhance both utility and portability for the motion capture system and to render it suitable for surveillance issues. Even though several stochastic techniques to address this issue exist, we aim to solve this problem in a deterministic way, for future real-time performance. We adopt a Bayesian framework under which we employ a 3D articulated model $\mathcal{M}$ and a rendering function Λ to describe the data. Unlike other existing approaches [1, 2], we solely use silhouette matching to obtain a measure of how well the model describes the data.

1. INTRODUCTION

In the past two decades many systems have been developed for achieving motion capture, with main applications in the movie industry and computer game design, but also in sports coaching etc. These systems typically make use of marker technology, which imposes heavy restrictions on the clothing of the subject. In most cases lycra clothing has to be worn, and IR emitters/reflectors have to be attached all over the subject's body, usually close to the joints. Even worse, this has the additional disadvantage of limiting the motion of the subject, as the subject becomes aware that the markers might fall off the clothing if special care is not taken during data acquisition. Furthermore, these methods usually have limited portability, with filming taking place in dedicated studios.

There is an increasing need for systems that impose as few restrictions on the subject as possible. Markerless Motion Capture circumvents the use of markers with all their disadvantages, and concentrates on model-based as well as image-based techniques to solve the problem of articulated body motion recovery. Applications of real-time markerless systems could cover those of the marker-based systems, but its additional portability and transparency also renders Markerless Motion Capture appropriate for automated surveillance, man-machine interaction, security and medical purposes.

Invariably, in every model-based motion capture system a fitness function is derived and subsequently optimised to estimate the true body state. Problems arise from (i) the multi-modality and (ii) the size of the state-space of the fitness function. Although several stochastic optimisation methods exist that can handle multimodal functions by cleverly sampling from the entire state space [2, 3, 4], any such method would be incapable of running in real time given the processing capabilities of today's computers. For the development of real-time motion capture systems recent research interest has been concentrating on *deterministic* algorithms that can track articulated bodies with fewer computations [1, 5] at the expense of adequate initialisation [6].

In this paper we present a system that uses a gradient-based search to optimise a differentiable fitness function to track a human subject. We begin by first stating the problem of motion capture in a statistical framework in Section 2. We then introduce our model and explain how we use it to estimate solutions to motion capture in Section 3. Our experiments are discussed in Section 4 and then proposals for future work as well as final conclusions are outlined in Section 5.

2. FORMULATION

The state vector $\mathbf{X}^{(k)} = \begin{bmatrix} x_1^{(k)} & x_2^{(k)} & \dots & x_s^{(k)} & \dots & x_S^{(k)} \end{bmatrix}^T$ denotes any possible configuration that the model may attain at any given time t_k. Let $\mathcal{M}(\mathbf{X}^{(k)})$ be the 3D model under some arbitrary state. A projection of the model's silhouette onto the image planes of the cameras may be expressed as $\Lambda\left(\mathcal{M}\left(\mathbf{X}^{(k)}\right)\right)$. For brevity we will be using $\mathbf{X}$ for the state and since only one model is considered, the

0-7803-8554-3/04/$20.00 ©2004 IEEE.

notation $\Lambda(\mathbf{X})$ will be used for the model's silhouette as viewed from the cameras.

Given the correct state $\mathbf{X}^*$ and assuming a suitable model the data may be described by,

$$D_k = \Lambda(\mathbf{X}^*) + N_k \tag{1}$$

where dataset D_k is the silhouette extracted from the k^{th} frameset captured by the cameras, $\Lambda(\mathbf{X})$ the silhouette produced by rendering $\mathcal{M}(\mathbf{X})$, and N_k is a random process which describes all the discrepancies between the model's silhouette and the data. $\Lambda(\mathbf{X})$, D_k and N_k are all vectors of size NUV, e.g.,

$$\Lambda(\mathbf{X}) = \big[\lambda_1(\mathbf{X})\ \ \lambda_2(\mathbf{X})\ \ \ldots\ \ \lambda_j(\mathbf{X})\ \ \ldots\ \ \lambda_{NUV}(\mathbf{X})\big]^T \tag{2}$$

where λ_j is the intensity of the j^{th} pixel. Our task is to find the best estimate for the true state $\mathbf{X}^*$. Adopting a Bayesian framework, we choose $\hat{\mathbf{X}}_{MAP}$, the maximum *a posteriori* estimate, i.e. the state that maximises the posterior probability distribution:

$$\hat{\mathbf{X}}_{MAP} = \arg\max_{\mathbf{X}}\ P\big(\Lambda(\mathbf{X})|D_k\big) \tag{3}$$

where, using Bayes' theorem,

$$P\big(\Lambda(\mathbf{X})|D_k\big) = \frac{P\big(D_k|\Lambda(\mathbf{X})\big)P\big(\Lambda(\mathbf{X})\big)}{P(D_k)} \tag{4}$$

with $P\big(D_k|\Lambda(\mathbf{X})\big)$ being the likelihood, $P\big(\Lambda(\mathbf{X})\big)$ the prior distribution, and $P(D_k)$ the evidence.

Regarding the evidence as a normalising constant and assuming no prior knowledge about the position of the subject's body in the k^{th} dataset, the prior distribution becomes flat, and we therefore need to maximise the likelihood, this way using the *maximum likelihood* estimator:

$$\hat{\mathbf{X}}_{ML} = \arg\max_{\mathbf{X}}\ P\big(D_k|\Lambda(\mathbf{X})\big) \tag{5}$$

Furthermore we assume that N_k is normally distributed with zero mean and that the intensity of the noise on each pixel in the datasets is independent of the noise on neighbouring pixels, giving a diagonal covariance matrix for N_k:

$$N_k \sim\ \text{i.i.d.}\ \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I}) \tag{6}$$

where $\mathbf{I}$ is the identity matrix of size NUV and σ_n^2 is a constant. This gives

$$\begin{aligned} P\big(D_k|\Lambda(\mathbf{X})\big) &= \big(2\pi\sigma_n^2\big)^{-\frac{NUV}{2}} \\ &\times\ \exp\Big\{-\frac{1}{2\sigma_n^2}\big(D_k-\Lambda(\mathbf{X})\big)^T\big(D_k-\Lambda(\mathbf{X})\big)\Big\} \end{aligned}$$

and we therefore need to *minimise* the negative log-likelihood function, i.e.,

$$\hat{\mathbf{X}}_{ML} = \arg\min_{\mathbf{X}}\big(D_k-\Lambda(\mathbf{X})\big)^T\big(D_k-\Lambda(\mathbf{X})\big) \tag{7}$$

To converge to the maximum likelihood estimator $\hat{\mathbf{X}}_{ML}$ that minimises (7) we use the Gauss-Newton method for Non-Linear Least Squares. For each iteration i a new approximation for the maximum likelihood estimator is given by,

$$\hat{\mathbf{X}}_i = \hat{\mathbf{X}}_{i-1} + M(\hat{\mathbf{X}}_{i-1})^{-1}J(\hat{\mathbf{X}}_{i-1})^T\big(D_k-\Lambda(\hat{\mathbf{X}}_{i-1})\big) \tag{8}$$

where $M(\mathbf{X}) = J(\mathbf{X})^TJ(\mathbf{X})$ and $J(\mathbf{X}) = \frac{\partial}{\partial\mathbf{X}}\Lambda(\mathbf{X})$ is the Jacobian of the model's silhouette.

3. THE MODEL

3.1. The State

Our model $\mathcal{M}$ is a 3D articulated structure that simulates the anatomy of a human skeleton (Fig. 1, left). The state vector $\mathbf{X}$ encodes the position and orientation of every limb of the model. The first three coefficients of $\mathbf{X}$ form the spatial offset vector $\mathbf{x}_{\text{sp}}$ of the body's centre of mass (in our case the base of the torso). The rest of the coefficients in $\mathbf{X}$ group up in fours and each of these groups describes the orientation at each joint. In other words,

$$\mathbf{X} = \big[\mathbf{x}_{\text{sp}}^T\ \ \mathbf{x}_{\text{r},1}^T\ \ \mathbf{x}_{\text{r},2}^T\ \ \ldots\ \ \mathbf{x}_{\text{r},n}^T\big]^T \tag{9}$$

where n is the number of limbs on the body that are attached on a rotating joint, and

$$\mathbf{x}_{\text{sp}} = [x\ \ y\ \ z]^T \quad \mathbf{x}_{\text{r}} = \Big[\cos\frac{\theta}{2}\ \ u_x\sin\frac{\theta}{2}\ \ u_y\sin\frac{\theta}{2}\ \ u_z\sin\frac{\theta}{2}\Big]^T\ .$$

Each quartet of state rotation coefficients $\mathbf{x}_{\text{r}}$ constitutes a *quaternion* [7] and parameterises the rotation of a single limb about a *unit* axis $\hat{\mathbf{u}} = [u_x\ \ u_y\ \ u_z]^T$ and through an angle θ. Even though each rotation is now described with 4 parameters, these are constrained in that $|\mathbf{x}_{\text{r}}| = 1$. The new estimate for the state is calculated using (8) and subsequently all the quaternions are normalised, effectively resulting in 3 degrees of freedom per limb rotation.

Euler angles ($\mathbf{x}_{\text{r}} = [\phi_x\ \phi_y\ \phi_z]^T$) could have been used to parameterise the limbs' rotations, resulting in a smaller size for the state vector $\mathbf{X}$. Such a parameterisation, however, suffers from discontinuities in $\mathbf{x}_{\text{r}}$ even during small variations of the limb's orientation [8]. Using quaternions on the other hand, smooths the state-space for $\mathbf{X}$ thus eliminating local minima and singularities.

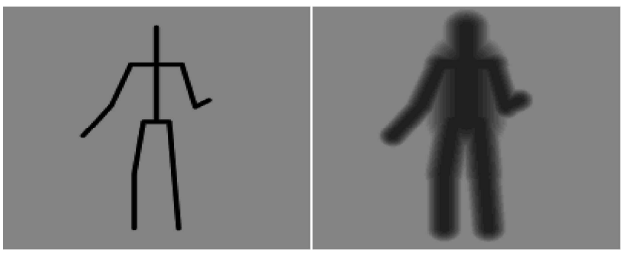

Fig. 1: Left: front view of $\mathcal{M}(\mathbf{X})$. Right: model silhouette $\Lambda(\mathbf{X})$ as seen from the same view.

3.2. The Rendering Function

The rendering function $\Lambda(\mathbf{X}) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is a perspective projection process that uses $\mathcal{M}$ to produce 2D framesets that are compared with the (2D) datasets D_k at any time. Equation (8) shows that the algorithm depends on gradient information provided by the rendering function $\Lambda(.)$. Hence, the *pixel-wise* rendering function $\lambda(.)$ is chosen to be such that linearly smooth silhouettes are produced under *all* possible states $\mathbf{X}$; one can achieve this by modelling each limb as a cylinder with hemispherical ends.

The collective relationship between λ and Λ is shown in equation (2). For the j^{th} pixel of the frameset we have,

$$\lambda_j(\mathbf{X}) = \begin{cases} 0 & \text{if } d_j(\mathbf{X}) < \frac{1}{2}r \\ \frac{255}{r} d_j(\mathbf{X}) - \frac{255}{2} & \text{if } \frac{1}{2}r \leq d_j(\mathbf{X}) \leq \frac{3}{2}r \\ 255 & \text{if } d_j(\mathbf{X}) > \frac{3}{2}r \end{cases} \quad (10)$$

This gives a linear intensity gradient w.r.t. the distance measure d_j, as shown in Fig. 2 on the left. Furthermore, d_j is defined to be the smallest of 3 candidate distance measures between the model's limb and the ray passing through the current pixel:

$$d_j(\mathbf{X}) = \min \left\{ \hat{d}_j(\mathbf{X}), \check{d}_j(\mathbf{X}), \bar{d}_j(\mathbf{X}) \right\} \ .$$

$\hat{d}_j$ and $\check{d}_j$ are the distances between the ray and either of the two ends of the limb, whereas $\bar{d}_j$ denotes the smallest distance between the axis of the limb and the current ray, as illustrated in Fig. 2 on the right.

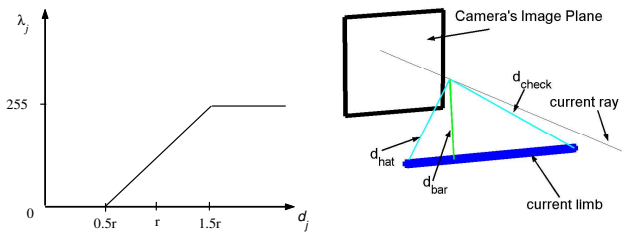

Fig. 2: Left: plot of $\lambda_j(\mathbf{X})$ vs $d_j(\mathbf{X})$. Right: illustration of the three candidate distance measures $\hat{d}_j, \check{d}_j, \bar{d}_j$.

4. RESULTS

4.1. Simulated Data

In the following examples a *single* data frameset D_1 is produced by rendering the model itself under a state $\mathbf{X}^*$, and then 'forgetting' about the state. We then initialise our model's state at $\hat{\mathbf{X}}_0$ and the algorithm is employed to converge to $\hat{\mathbf{X}}_{ML}$, (which in this case is identical to $\mathbf{X}^*$). Under this scenario the additive noise process $N_1 = \mathbf{0}$ and equation (1) implies that the model is *perfect*, as it can describe the data with no error.

4.1.1. Single Camera

In Fig. 3 we display three snapshots from the silhouette matching process for the case of a single-camera system. For reasons of clarity the model's silhouette is rendered white and is superposed on the data silhouette which is painted black. Despite the ambiguity problem when working with pure silhouettes, we see that the algorithm can converge to the global minimum provided that $\hat{\mathbf{X}}_0$ lies in the vicinity of $\hat{\mathbf{X}}_{ML}$.

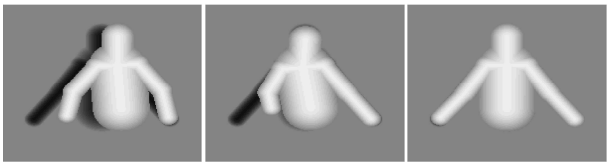

Fig. 3: Silhouette matching on a simulated dataset using 1 camera. $\Lambda(\hat{\mathbf{X}}_i)$ superposed on D_1 for iterations $i = 0$ (left), $i = 7$ (middle), $i = 120$ (right).

4.1.2. Stereo Vision

Double camera configurations helped the algorithm to converge towards the global minimum with a smaller number of iterations due to the additional constraints of the extra data, as shown in Fig. 4: The stereo vision configuration (right) has converged correctly with only 30 iterations while the single camera configuration (left) needed 120, even though identical initial state approximations ($\hat{\mathbf{X}}_0$) and true states ($\mathbf{X}^*$) have been used. Put in a different way, we are now able to choose the initial state $\hat{\mathbf{X}}_0$ for silhouette matching with greater freedom, and still converge to the correct state $\mathbf{X}^*$. The maximum allowable distance that guarantees correct state recovery $\epsilon = |\hat{\mathbf{X}}_0 - \mathbf{X}^*|$ for the double camera configuration is now greater than the one for the single camera, implying that the camera frame capture rate may now be reduced. We can generalise this by noting that data captured by additional cameras made the state-space more *convex*, and the optimisation method consequently more robust.

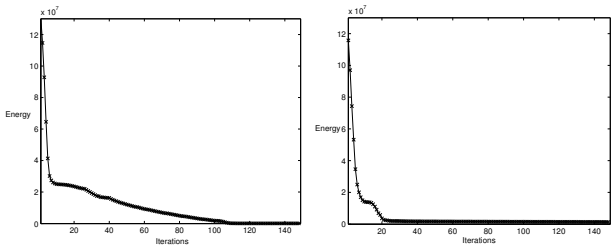

Fig. 4: Variation of the negative log-likelihood w.r.t. number of iterations i during silhouette matching with simulated data on dataset D_1 for the single camera (left) and double camera (right) experiments.

4.2. Real Data

Datasets D_k were now produced by removing the background from a real person's raw footage and subsequently

smoothing. This process produces data that are far from ideal, as there are now two types of distortion: (i) the model and its rendering function are not complex enough to describe the natural curvature of the human body and the clothing around it, and (ii) D_k has been further corrupted during the process of background removal. All these contribute towards N_k, the random process described in (6). The performance of our algorithm under these conditions for the case of the 1^{st} dataset D_1 is shown in Fig 5.

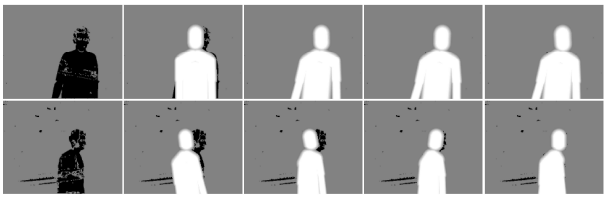

Fig. 5: Silhouette matching on a *real* dataset using 2 cameras. 1^{st} row: view from camera 1; 2^{nd} row: view from camera 2. Column 1 shows real data D_1; columns 2 to 5 show $\Lambda(\hat{\mathbf{X}}_i)$ superposed on D_1 for iterations $i = 0$, $i = 6$, $i = 10$ and $i = 25$ respectively.

When silhouette matching is achieved repeatedly over the continuous flow of datasets we achieve *motion recovery*. For each new dataset D_k a new optimisation process of 25 iterations takes place and by default the initial approximation for the current state $\hat{\mathbf{X}}_0^{(k)}$ is taken to be the converged state of the previous frame $\hat{\mathbf{X}}_{25}^{(k-1)}$.

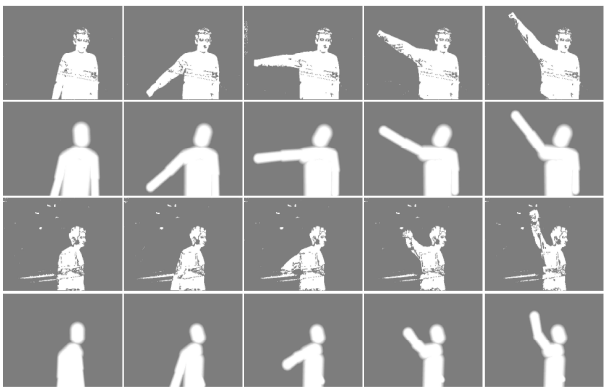

Fig. 6: Motion recovery after successful silhouette matching on a video sequence of real datasets $D_{1:50}$ using 2 cameras. 1^{st} row: D_k as viewed from camera 1; from left to right: D_1, D_9, D_{17}, D_{25} and D_{33}. 2^{nd} row: $\Lambda(\hat{\mathbf{X}}_{25})$ for each corresponding D_k of row 1 from the same view. 3^{rd} row: same D_k as viewed from camera 2. 4^{th} row: $\Lambda(\hat{\mathbf{X}}_{25})$ for each corresponding D_k of row 3 from the same view.

Fig. 6 shows the real datasets D_k as well as the converged model silhouettes $\Lambda(\hat{\mathbf{X}}_{25})$ as observed by the double-camera system for different time instants t_k. We see that despite the low quality of the data the system recovers the subject's state at each frame to a considerable degree of accuracy.

Our method is still in a developing stage and details on computational complexity are not available as yet. Processing time for 50 frames of real data (including the time taken for the extraction of the silhouettes) took about 30 minutes using Matlab. The system will later be upgraded into C++ code and its real-time performance will be tested.

5. CONCLUSIONS AND FUTURE WORK

We have presented a simple system that uses Bayesian inference to track human subjects using silhouette matching, Likelihood Maximisation and a smooth parameterisation of the model state space using quaternions. Provisional results on both simulated and real data are very promising. Future work on the system includes incorporation of both prior distributions $P\big(\Lambda(\mathbf{X})\big)$, and priors for the reduction of the effect of local minima during the optimisation part of the algorithm. Priors could be derived by methods similar to the ones in [1] where a rough classification of the motion in an early stage yields a cue of the type of motion tracked. The anatomical structure of the model will also be upgraded, and rendering will be extended to produce silhouettes that resemble more accurately the human body. Ellipsoids of variable principal axis radii will be used for the simulation of major muscle groups as in [9] as opposed to simply modelling limbs with ellipsoids. Creating a more realistic model will validate the assumptions made about the Gaussian nature of N_k. More advanced tracking strategies using Kalman Filtering will also be considered. Finally, we will address the issue of state initialisation, perhaps with a Particle Filtering based approach as in [2, 3, 4].

6. REFERENCES

[1] R. Urtasun and P. Fua, "3d human body tracking using deterministic temporal motion models," in *Proc. European Conference on Computer Vision (ECCV)*, May 2004, vol. III, pp. 92–106.

[2] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2000.

[3] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[4] K. Choo and D. J. Fleet, "People tracking using hybrid monte carlo filtering," in *Proc. International Conference on Computer Vision (ICCV)*, 2001.

[5] P. Fua, A. Gruen, N. D'Apuzzo, and R. Plankers, "Markerless full body shape and motion capture from video sequence," in *Proc. International Archives of Photogrammetry and Remote Sensing*, 2002, vol. 34, pp. 256–261.

[6] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.

[7] S.L. Altmann, *Rotations, Quaternions, and Double Groups*, Clarendon Press, Oxford, UK, 1986.

[8] T. Y. Zhao, "Articulated models for motion tracking," MEng Thesis, Cambridge University Engineering Department, June 2001.

[9] N. D'Apuzzo, "Motion capture by least squares matching tracking algorithm," in *Proc. AVATARS2000*, November 2000.