Multi-view Human Motion Capture with An Improved Deformation Skin Model

Yifan Lu[1], Lei Wang[1], Richard Hartley[12], Hongdong Li[12], Chunhua Shen[12]
[1]Department of Information Engineering, CECS, ANU [2]NICTA, Canberra Research Lab
{Yifan.Lu, Lei.Wang, Richard.Hartley, Hongdong.Li, Chunhua.Shen}@anu.edu.au

Abstract

Markerless human motion capture has received much attention in computer vision and computer graphics communities. A hierarchical skeleton template is frequently used to model the human body in literature, because it simplifies markerless human motion capture as a problem of estimating the human body shape and joint angle parameters. The proposed work establishes a skeleton based markerless human motion capture framework, comprising of 1) an improved deformation skin model suitable for markerless motion capture while it is compliant with the computer animation standard, 2) image segmentation by using Gaussian mixture static background subtraction and 3) non-linear dynamic temporal tracking with Annealed Particle Filter. This framework is able to efficiently represent markerless human motion capture as an optimisation problem in the temporal domain and solve it by the classic optimisation scheme. Several experiments are used to illustrate its robustness and accuracy comparing with the existing approach.

1. Introduction

Human motion capture also known as 3D posture estimation is a task of acquiring motion information from a moving performer. It is a problem of estimating the parameters of the human body model from the acquired data as the position and configuration of the tracked performer change over time. One class of applications are those where the estimated body model parameters are used directly, for example to interact with a virtual world, drive an animated avatar in a video game or for computer graphics character animation. Another class of applications use estimated parameters to classify and recognise people, gestures or motions, such as surveillance systems, intelligent environments, or advanced user interfaces (sign language translation, gesture driven control, gait, or pose recognition). Finally, the motion parameters can be used for motion analysis in applications such as personalised sports training, choreography, or clinical studies of orthopedic patients. Currently available commercial systems for motion capture require the subject to wear special markers, body suits or gloves. The dedicated hardware is not often affordable for individuals. In contrast to such marker based motion capture, finding an economical solution, which utilises markerless, unconstrained posture estimation using only cameras, has gained increasing attention in computer vision. Recent efforts [22, 19, 1, 3, 14, 25] have been focused on markerless human motion capture in order to realise a cost-effective and easily deployed motion capture system.

In markerless human motion capture, the silhouette is often used to describe the shape of the human body, since it is more robust to illumination variation and easier to be matched and corresponded than colour and texture features. However, when lacking colour and texture information, essentially, the silhouette describes an image no more than a contour line which only contains partial information of the original image. Shape ambiguities can raise along the depth direction. Considering an example of human body reconstruction from multi-view silhouette images, the shape-from-silhouette [13] is often used to achieve reconstruction by computing a spatial intersection "visual hull" of silhouettes. Strictly speaking, the true visual hull is the maximal volume constructed from all possible silhouettes (infinite number of silhouettes). It is not computable in almost any practice circumstances. Alternatively, an "inferred" visual hull is computed with respect to only a finite number of silhouettes. The inferred visual hull usually has greater volume than the actual human body. This can result in multiple solutions when one attempts to fit the original human body to the inferred visual hull. In fact, this fitting exercise is a central part to solve markerless human motion capture. Furthermore, human motion is a complex process concerned on the human anatomy structure which involves interactions of multiple bones and muscles, external forces and other factors. To model such a complicated human anatomy structure, the high-dimensional variable has to be used to parameterise the human motion and posture. Overall, it turns out that the solution of markerless human motion capture is subject to a non-convex and multi-modal optimisation in a high dimensional space.

The proposed approach integrates a generic parameterisable human skeleton template with the deformable skin which encourages the more accurate silhouette based likelihood evaluation, performs run-time Gaussian mixture static background subtraction to segment silhouettes from multi-view images and finally incorporates parameterised human pose as state variables and silhouettes as observations into the Annealed Particle Filter framework to solve a global optimisation problem in high dimensional space.

2. Related Works

Focusing on difficulties in the high dimensional optimisation, Sminchisescu et al [23] proposed a global search method that combines a local optimisation with a long escape perturbation, while iteratively scaling the search space by adjusting the perturbation covariance. Since their method can effectively scale the search space and explore a relatively large area, it is able to work even in the monocular image case. On the other hand, continuing the success of the particle filter [8, 2] in non-linear generic object tracking, many works (e.g. [6]) attempt markerless human motion capture by utilising the particle filter. However, when dimensionality increases, the particle filter does not scale very well and often fails to approximate the posterior distribution. This has been addressed in the method [7] proposed by Deutscher et al. It firstly extends the particle filter by augmenting simulated annealing [12] in the context of human motion capture, known as the Annealed Particle Filter (APF). A similar approach annealed importance sampling was stated in a statistics publication [17]. APF, which inherits advantages of simulated annealing, is able to escape from local minima and find the global minimum, but good tracking depends on a slow annealing schedule. In such skeleton based approach, many studies[3, 26, 20, 7] use simple geometry primitives (e.g. cylinder and box) to approximate the skin of the human body. This often results in the mismatch between the silhouette and the human model, and therefore it possible leads to the inaccurate silhouette based likelihood evaluation.

Departing from the above point of view, many excellent attempts have also tried on the learning based method. Agarwal et al's [1] method recovers human pose from a single image by evaluating both regularised least squares and the relevance vector machine regressor over both linear and learned basis. Their method does not depend on pre-built human model and label images, which usually is a general assumption for other existing approaches. In Wang et al's work [25], the Gaussian process latent variable model is used to learn the lower dimensional representation from motion capture data, resulting in a non-parametric dynamical model that accounts for uncertainty in the model. The learning based approach can be generalised well to activities similar to the learned one, however, it has inability to cope with unknown type activities.

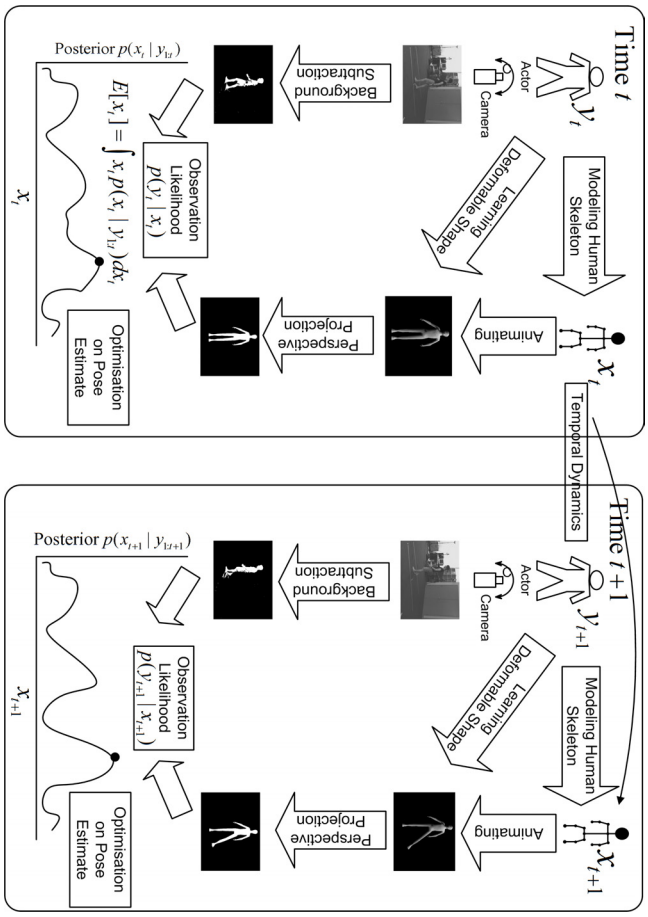

Figure 1: Dynamical System Architecture

3. Framework Overview

The framework described in this section shares common components with those tracking frameworks proposed in literature. Intuitively, it is constructed on the basis of the dynamical system that, given a pair of a hidden state and an observation corresponding to a certain point of time, explicitly characterises the causality between the state and the observation, and the dependency between the state and the prior state in the temporal domain. Provided the state is independent of other states, the first-order Hidden Markov Model [5] is sufficient to capture the sequential characteristics of states. The first order Hidden Markov Model assumes only a dependency between the current state and the previous state. All other states are ignored. Therefore, estimating the current state no longer requires storing all historical states.

From the dynamical system point view, human motion can be considered as a sequence of states (human poses) and signals (associated observations) emitted from these states. The above dynamical system can be contextualised by human motion capture. At a certain point of time t, there is an observation y_t that is the observable evidence of the human pose, and a hidden state x_t that is an underlying true pose. The goal is to find a true state, given the current and historical observations. In reality (the solution within the dynamical system is often analytically intractable), it is impossible to obtain the exact value of the true state x_t. Hence, a Monte Carlo estimate $\widehat{x_t}$ is calculated instead. From the computer vision literature, a recursive Bayesian formulation [8, 2], which recursively calculates expectation of x_t over the posterior $p(x_t|y_{t:1})$, has been proven to be a reliable estimate and is widely employed. It starts with the previous posterior distribution, maximises a posterior by considering the product of the observation likelihood and the prior in the sense of the Bayesian paradigm. The optimal estimate is found when the posterior probability is maximised. Maximising the posterior probability substantially relies on a likelihood function that measures how well an estimate state $\widehat{x_t}$ fits observation y_t. The selection and design of such a likelihood function could be very flexible, but it should obey the principle of simplicity since most computational loads will be spent in evaluating the likelihood function.

The Contextualised Dynamical Architecture illustrated in Figure 1 depicts the overview of the framework. Initially, the action of an actor at time t is captured as a digital image. Subsequently, background subtraction is applied to this image to generate a silhouette image represented by a bit mask. In parallel, the pose is described by an articulated skeleton associated with a parameter vector. A pseudo silhouette image can be rendered by applying perspective projection [9] with known camera calibration parameters. The observation likelihood can then be evaluated by comparing the overlapping area between the silhouette image and the pseudo silhouette image of the estimate state $\widehat{x_t}$. The deformable virtual human model is learned from the real actor at the training stage, and prior knowledge about how the pose evolves from time t to $t+1$, is incorporated by employing a temporal dynamical model.

4. Skeletal Deformation

As a strong prior, the articulated human body model is widely used in literature (e.g. [4, 7]) in order to reduce unnecessary complexities of human motion capture. It provides a parametrisation[1] of human pose, and allows markerless human motion capture to be formalised as tracking of

[1] Such parametrisation has been extensively studied in computer animation and graphics. It has already appeared as the standard in industry.

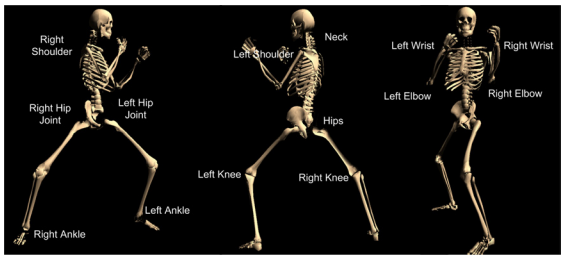

Figure 2: Human Skeleton and Joint Angles

a hierarchical kinematic structure which is relative simple and well studied.

Up to date, there are two primary standards describing modelling of the human body in computer animation, H-Anim (Humanoid Animation) 1.1 standard [10] and Body Animation MPEG4 standard [11], and three major motion capture formats C3D, BVH/BVA, ASF/AMC which are used to store and retrieve the motion capture data. These two standards and three data formats share commons in a way of defining a generic human skeleton structure. This is not a coincidence. Any simple and compact representation of the human body should naturally fit to the anatomy of the human body. The proposed human body skeleton is therefore designed to be conformed to H-Anim standard and the ASF/AMC format. This consideration enables a natural integration to the skeleton-based animation scheme in the computer animation standard.

The human model used in this work is based on the skeleton illustrated in Figure 2, which has total 27 segments and 167 degrees of freedom (DOFs). Avoiding too complicated representation, only 10 articulated segments (the ankle and wrist joints are optional) and 25 DOFs are considered as important and modelled for tracking. The translation and orientation of the entire model are described by 6 DOFs. The rest of 19 DOFs is used to describe the joint angles of limbs. Thereby, any point ${}^bP$ in a local kinematic coordinate frame b can be transformed to the point ${}^wP$ in the world coordinate frame w by:

$$ {}^wP = \prod_i^N T(\theta_i) {}^bP $$

where, N is the number of transformations. $T(\theta_i)$ is a homogeneous transformation matrix specified by θ_i a particular set of joint angels and a translation.

Because of conforming to the humanoid skeleton standard, rather than using geometry primitives to approximate the skin, the skin can be imported from any 3D object format. The 3D skin mesh can be then associated with the hierarchical skeleton by assigning a group of vertices to each bone. This is sometimes referred to as rigging. Each vertex in the mesh is associated and controlled by multiple bones

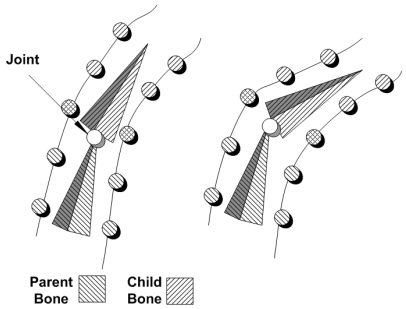

Figure 3: Vertex blending. The bones are drawn in triangular solids, vertices are drawn in circles. Vertices are shaded according to its associated bones. The movement of bones drive the vertices to be transformed up to the scale of vertex weights, ultimately leading to skin deformation.

with a scaling factors called vertex weights[2]. As a result, Portions of the skin can be deformed to account for transformations of multiple bones. Instead of animating each vertex individually, the skeleton is manipulated, and the skin is deformed automatically. As an example illustrated in Figure 3, vertices are assigned to the bones according to geometrical distances. As the child bone is rotated, its associated vertices are transformed up to the scale of vertex weights. Therefore, the vertices which are far from the parent bone are transformed further. Conversely, the vertices close to the parent bone almost remain as before.

This is formally stated in the Skeletal Subspace Deformation (SSD) algorithm [16] which is based on the weighted blending of an affine transformation of each joint by:

$$v_d = (\sum_{i=1}^{M} w_i T(\theta_i)) v_0$$

where, M is the number of joints, v_d is a vertex after deformation, w_i is vertex weights and v_0 is a vertex in the registered initial pose. Although SSD suffers from inherent limitations of the linear blending [15] (known as"collapsing joints" and "twisting elbow problem" which in general, are the mesh deformed by SSD loses volume as the joint rotation increases), this simple algorithm still remains the most popular deformation scheme because it is computationally efficient.

Figure 4c and 4d show the silhouette matching examples for a already tracked posture using the proposed model and Balan et al's model in [4]. The non-overlapping area is significantly reduced by using the SSD scheme, resulting in the more accurate matching evaluation.

[2]Vertex weights are often assigned by the computer graphics software.

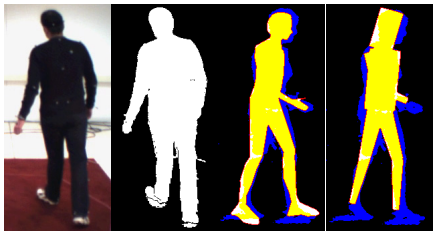

Figure 4: Image segmentation by static background subtraction and silhouette matching examples (The original silhouettes are coloured by blue, the generated model silhouettes are coloured by yellow)

5. Static Background Subtraction

The silhouette is a suitable shape descriptor for markerless human motion capture. Its generation and quality are vital important for subsequent processes. Any noise appearing in silhouettes will eventually remain in the inferred visual hull and corrupt the shape of the original body. Poor silhouette extraction usually is a primary reason leading to the failure of tracking.

In this work, static background subtraction used is in a way similar to the method described in [24]. Initially, the static background statistics are modelled by the pixel-wised Gaussian mixtures which account for general variations in the static background. As the static background statistics are collected in the three periods, each of them can be modelled by a Gaussian distribution. Therefore, each pixel-wised Gaussian mixture model has three temporal components. When the object appeared, any pixel with the probability deviating from the normal static range, which is a pre-defined threshold, is labelled as foreground, otherwise is labelled as background. The background pixel value x should satisfy the following criterion:$p(x) = \sum_{i=1}^{k} \eta_i N(x, u_i, \sigma_i^2) > C_{thres}$ where, u_i is the mean value of the static background, σ_i^2 is the variance of the static background, C_{thres} is the pre-defined threshold and η_i is component coefficients for Gaussians. The component coefficients and a normal static threshold can be learned from correctly segmented images or determined according to the feedback from an interactive segmentation approach. This method is robust, fast and easy to be incorporated into the markerless human motion capture framework. An example of the segmentation result is shown in Figure 4a and 4b.

6. Optimisation by Annealed Particle Filter

6.1. Particle Filter

Particle filter is built on the basis of the recursive Bayesian filter which is firstly formulated in [8, 18], which generalises the temporal dependencies of a sequential dynamical system by First-order Hidden Markov Model in the sense of the Bayesian paradigm. The mathematical formulation is given by:

$$p(x_t|y_{1:t}) \propto p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}$$

Intuitively, above formula states that the predictive posterior is dependent upon likelihood-weighted expectation of temporal dynamics/transition priori $p(x_t|x_{t-1})$ with respect to the previous posterior $p(x_{t-1}|y_{1:t-1})$.

Combining recursive Bayesian filter with importance sampling, particle filter can be briefly described below:

Let $\{x_t^i, w_t^i\}_{i=1}^N$ denotes a set of N random samples x_t^i with associated normalised importance weights $w_t^i (\sum_{i=1}^N w_t^i = 1)$ at time t. Providing that the N number of samples is reasonable large with respect to the dimensions of the state vector x, an empirical estimate of posterior $p(x_t|y_{1:t})$ at time t can be approximated as:

$$p(x_t|y_{1:t}) \approx \sum_{j=1}^{N} w_t^j \delta_{x_t^j}(x_t)$$

where, $\delta_{x_t^j}(x_t)$ is Kronecker delta function. Further, the estimate state $\hat{x_t}$ can be evaluated by the expectation of the posterior probability:

$$\hat{x_t} = E[x_t]_{p(x_t|y_{1:t})} = \int x_t p(x_t|y_{1:t})dx_t \approx \sum_{i=1}^{N} w_t^i x_t^i$$

Assuming that the current state and observation are dependent solely upon the immediate previous state and current observation, the primary task of the algorithm is to determine importance weights w_t^i. This is done by iteratively updating importance weights by:

$$w_t^i \propto w_{t-1}^i \frac{p(y_t^i|x_t^i)p(x_t^i|x_{t-1}^i)}{\pi(x_t^i|x_{t-1}^i, y_t^i)} \quad (1)$$

where, $p(y_t^i|x_t^i)$ is the observation likelihood, $p(x_t^i|x_{t-1}^i)$ is the temporal dynamics, and $\pi(x_t^i|x_{t-1}^i, y_t^i)$ is the important distribution.

6.2. Annealed Particle Filter

APF [7] incorporates simulated annealing for minimising the energy function $E(y_t, x_t)$ or, equivalently, maximising the observation likelihood $p(y_t|x_t)$ in the particle filter. The observation likelihood is essential to approximate the posterior distribution and often formulated in a form of the Boltzmann distribution:

$$p(y_t|x_t) = \exp\{-\lambda E(y_t, x_t)\} \quad (2)$$

where, $E(y_t, x_t)$ is an energy function between y_t and x_t. λ annealing variable is an inverse of the product of the Boltzmann constant and the temperature.

One simple energy function $E(y_t, x_t)$ can be defined as:

$$E(y_t, x_t) = \frac{1}{N_v} \sum_{i=1}^{N_v} D_s(y_t, x_t) + \alpha D_c(y_t, x_t) \quad (3)$$

where, N_v is the number of views. $D_s(y_t, x_t)$ measures differences between the observed silhouette y_t and the silhouette generated by the particle x_t. $D_c(y_t, x_t)$ measures differences between contours which is used to emphasise the shape consistency. α is used to balance the silhouette and contour term.

The optimisation of APF is iteratively done according to a predefined M-phase schedule $\{\lambda = \lambda_1, ..., \lambda_M\}$, where $\lambda_1 < \lambda_2 < ... < \lambda_M$, known as the annealing schedule. At time t, considering a single phase m, initial particles are outcomes from the previous phase $m-1$ or drawn from the temporal model $p(x_t^i|x_{t-1}^i, y_t^i)$. Then, all particles are weighted by their observation likelihood $p(y_t|x_t)$, resampled probabilistically to select good particles which are highly likely to near the global minimum. Finally, particles are perturbed by a Gaussian noise with zero mean and the diagonal covariance matrix.

Besides λ_m, another two important parameters a survival rate α_m and a perturbation covariance matrix P_m control and tune a pace how samples are superseded and perturbed to concentrate on the minimum of the energy function. Given the survival rate α_m and particles at the current phase, λ_m can be determined as suggested in [7] by:

$$\alpha_m N \sum_{i=1}^{N} (w_{t,m}^i)^2 = \left(\sum_{i=1}^{N} w_{t,m}^i \right)^2 \quad (4)$$

where, N is the number of particles, $w_{t,m}^i = \exp\{-\lambda_m E(y_t^i, x_{t,m}^i)\}$.

Overall, the APF algorithm for a typical frame can be summarised in the Algorithm 1.

7. Experiments

Experiments were conducted on the publicly available MOCAP (Synchronized Video and MOCAP dataset) and HumanEva-I dataset [20, 4, 21] from Brown University. These datasets are aimed to quantitative evaluation for articulated human motion, and contain multiple videos, multiple

Algorithm 1 Anneal Particle Filter for a typical frame at time t

Require: appropriate α_m is defined, previous particles $\mathbf{x_{t-1}}$, observation y_t, the number of phases M and the initial covariance matrix P_0 are given

for $m = 1$ to M **do**

1) Initialise N particles x_t^i from the previous phase or the temporal model $p(x_t^i|x_{t-1}^i, y_t^i)$.

2) Calculate the energy $E(y_t, x_t)$ for all particles using the equation (3).

3) Find the λ_m by solving the equation (4).

4) Update weights for all particles using the equation (2).

5) Resample N particles from the important distribution.

6) Perturb particles by Gaussian noise with covariance $P_m = \alpha_m P_{m-1}$ and mean $\mu = 0$.

end for

subjects with properly documented body parameters, camera calibration parameters and motion capture ground truth. The motion capture data was captured simultaneously using a calibrated marker-based motion capture system and multiple high-speed video capture systems.

The first experiment is on the MOCAP dataset that contains a walking subject with total 529 frames, 4-view grayscale images synchronised in hardware at 60 Hz and motion capture data which was collected using a 6 camera Vicon system at 120 Hz. Mocap and image data was synchronised in software by sub-sampling and aligning the two data streams. The results of the proposed method are compared with the results of Balan et al's Annealed Particle Filter. The position error in 10 centimetres and body orientation error in degrees against the ground truth data are plotted in Figure 5. Since the initial pose comes from ground truth, the position errors of both Balan et al 's and the proposed method appear relatively small. As the tracking process proceeds, both methods deviate from the ground truth data and experience the larger errors. However, the proposed method is able to maintain the errors on average fewer than 8 centimetres for the position error. The body orientation error fluctuates around 8 degrees. Although it appears less accurate than Balan et al 's method, it still provides comparable results. Particularly, Balan et al 's method mistracked the right arm (highlighted in yellow) at the frame 35 (illustrated at last row and left 4 columns in Figure 6) and the left leg (highlighted in light blue) at the frame 135 (illustrated at the last row and right 4 columns in Figure 6). Since the more human-like model is used, the silhouette based likelihood evaluation becomes more accurate, the proposed method is able to track correctly in both situations.

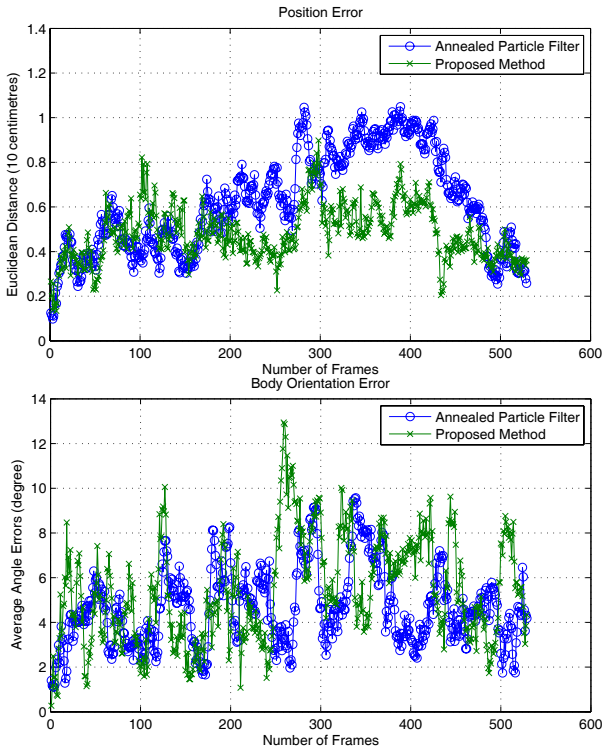

Figure 5: The position error and body orientation error against the ground truth data

The second experiment was conducted on the HumanEva-I dataset that contains software synchronised data from 7 video cameras (4 grayscale and 3 colour) and motion capture data[3]. Figure 7 illustrates the tracking at the frame 17, 91, 198 and 418, respectively. For every two rows, the top row shows the original image data from 4 greyscale and 3 colour cameras. The bottom row shows the tracking results with the human skeleton model projected on the original images. Overall, the tracking result is robust, except that the orientation of the head is not quite accurate. The reason for this is because the shape of the head is close to a sphere which is central symmetry and has a high possibility to raise ambiguities.

8. Conclusion and Future work

This paper has demonstrated a skeleton-based markerless human motion capture framework that utilises a skeletal deformation scheme to improve the likelihood evaluation accuracy, static background subtraction and the global optimisation by the annealed particle filter as its basic components. It has outlined a systemic way to understand the

[3]Since this project is undergoing, the quantitative evaluation against motion capture data is not available and only visual results are provided at this stage.

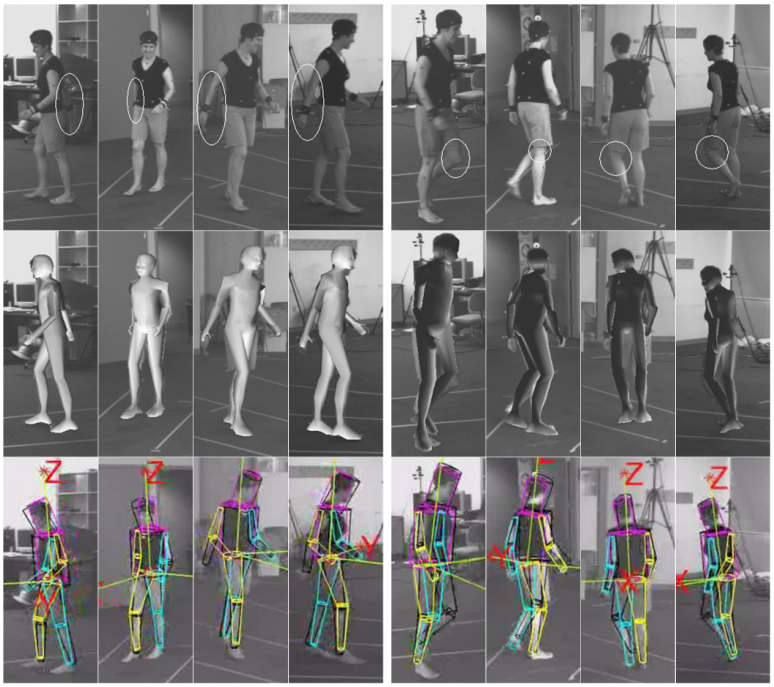

Figure 6: Left and right 4 columns are tracking results for frame 35 and 135, respectively. The rows from top to bottom correspond to the original images, the proposed method's and Balan et al's results (Ground truth is coloured in black), respectively. Note the circled parts are mistracked in Balan et al's results and correctly tracked in the proposed method's results.

nature of complex markerless human motion capture problem and solve it within a well-defined framework. Experiments have also shown the proposed approach is able to solve the problem efficiently on the real data and perform more robust than the existing approach on some situations.

Future work will continue to investigate the markerless human motion capture with more emphasis on the temporal dependencies. It may employ machine learning methods to encode temporal dependencies in human motion in order to help the global optimisation.

9. Acknowledgement

The authors would like to thank Balan and Sigal at Brown University for providing access to their dataset and answering the questions. This work is funded by Australia Research Council and NICTA.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.

[2] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb. 2002.

[3] A. Balan and M. Black. An adaptive appearance model approach for model-based articulated object tracking. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:758–765, 17-22 June 2006.

[4] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 349–356. IEEE Computer Society, 2005.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[6] O. Bernier. Real-time 3d articulated pose tracking using particle filters interacting through belief propagation. *International Conference on Pattern Recognition*, 1:90–93, 2006.

[7] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2):185–205, 2005.

[8] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

Figure 7: Tracking results on HumanEva I. Every two rows from top to bottom corresponds the frame 17, 91, 198 and 418, respectively.

[10] Human Animation Working Group. Information technology computer graphics and image processing humanoid animation (h-anim). *ISO/IEC FCD 19774:200x*, version 1.1.

[11] ISO/IEC Moving Picture Experts Group. Information technology – coding of audio-visual objects – part 2: Visual. *ISO/IEC 14496-2:2004/Amd 4:2008*, 2008.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[13] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[14] C.-S. Lee and A. M. Elgammal. Carrying object detection using pose preserving dynamic shape models. In *Articulated Motion and Deformable Objects*, pages 315–325, 2006.

[15] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, pages 165–172, 2000.

[16] N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings on Graphics interface '88*, pages 26–33. Canadian Information Processing Society, 1988.

[17] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[18] D. P. Sequential monte carlo methods in practice. *Journal of the American Statistical Association*, 98:496–497, January 2003.

[19] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision*, 73(3):243–262, 2007.

[20] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 1:I–421–I–428 Vol.1, June-2 July 2004.

[21] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, Department of Computer Science, 2006.

[22] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Bme : Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):2030–2044, Nov. 2007.

[23] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *The International Journal of Robotics Research*, 22(6):371, 2003.

[24] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 2:–252 Vol. 2, 1999.

[25] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, Feb. 2008.

[26] X. Xu and B. Li. Learning motion correlation for tracking articulated human body with a rao-blackwellised particle filter. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.