Gradual Sampling and Mutual Information Maximisation for Markerless Motion Capture

Yifan Lu[1] Lei Wang[1] Richard Hartley[13] Hongdong Li[13] Dan Xu[2]

[1]Department of Information Engineering, CECS, Australian National University
[2]Department of Computer Science and Engineering, SISE, Yunan University
[3]Canberra Research Labs, National ICT Australia
{Yifan.Lu, Lei.Wang, Richard.Hartley, Hongdong.Li}@anu.edu.au, danxu@ynu.edu.cn

Abstract. The major issue in markerless motion capture is finding the global optimum from the multimodal setting where distinctive gestures may have similar likelihood values. Instead of only focusing on effective searching as many existing works, our approach resolves gesture ambiguity by designing a better-behaved observation likelihood. We extend Annealed Particle Filtering by a novel gradual sampling scheme that allows evaluations to concentrate on large mismatches of the tracking subject. Noticing the limitation of silhouettes in resolving gesture ambiguity, we incorporate appearance information in an illumination invariant way by maximising Mutual Information between an appearance model and the observation. This in turn strengthens the effectiveness of the better-behaved likelihood. Experiments on the benchmark datasets show that our tracking performance is comparable to or higher than the state-of-the-art studies, but with simpler setting and higher computational efficiency.

1 Introduction

Recent advances of markerless motion capture have produced a number of new methods and approaches. Balan et al. [1] extended a deformation scheme to recover the detailed human shape and pose from images. Corazza and Mundermann et al.'s work employs an articulated ICP based method to register body segments with a sequence of visual hulls. Meanwhile, high quality performance capture [2] presented by de Aguiar et al. proposes a new mesh-based framework which captures both gestures of the subject and recovers small-scale shape details. The similar study [3] from Starck and Hilton adopts a graph-cut global optimisation for highly realistic surface capture. Most of the above works reside on the same underlying concept, shape-from-silhouette [4]. A bounding geometry of the original 3D shape, so called Visual Hull, is determined by intersecting generalised cones formed by back-projecting each multi-view silhouette with camera parameters. Moreover, these studies employ a similar approach that adjusts the human template model to explicitly or implicitly fit a visual hull in order to find the best posture. Unfortunately, when only limited camera views and moderate quality data are available, approaches in this category often suffer from low

tracking accuracy and the lack of robustness as pointed out in [5]. This is due to gesture ambiguities raised from two major facts: 1) With the limited camera views, the visual hull is the maximal volume of the tracking subject which is a superset of the true volume. It does not determine a unique posture but multiple ones; 2) Silhouettes generated by image segmentation are not reliable but noisy. Visual hulls from noisy silhouettes are often corrupted and inconsistent. As a consequence, optimisation algorithms are more probably to be trapped in local maxima, leading to incorrect gestures. Moreover, the global maximum of the observation likelihood can be shifted because of the corrupted data. In the Annealed Particle Filtering based tracking framework, a sufficiently slow annealing schedule can be taken as a measurement to rescue particles from the attraction of local minima. With such an annealing schedule, the initial energy function is shaped in a way that local minima are flattened out whereas the global minimum becomes relatively more pronounced. Also, particles will have more evaluations and random perturbations to move into the neighbourhood of the global mode. However, a sufficiently slow annealing schedule is usually computationally intractable in practice.

Besides designing more efficient annealing schedules, a better-behaved observation likelihood function is also vital important to guarantee that optimisation converges to the true posture. In this work, we propose an efficient and robust observation likelihood function for human motion tracking: 1) It employs an appearance-based likelihood evaluation with an illumination-invariant Mutual Information (MI) [6] criterion (in Section 3.1) to supersede the silhouette-based evaluation. In doing so, the likelihood evaluation can avoid the artifacts or noises introduced by silhouette segmentation. Also, with extra appearance information introduced, the landscape of the observation likelihood function becomes well modelled. These make the true posture is more likely to correspond to the global mode of the likelihood function; 2) More importantly, a gradual sampling scheme is developed (in Section 3.2) for the Annealed Particle Filtering (APF) based tracking framework. It is a smart selection of the observed image data to be included in the evaluation of the observation likelihood function. The selection is based on the error distribution over the observed image in the previous exploration. This scheme works in an error-oriented fashion, allowing the evaluations of observation likelihood to concentrate on major and relevant errors. It is able to produces better tracking performance and higher computational efficiency with a common annealing schedule. The excellent tacking performance obtained by our approach will be verified by the experiments on the benchmark data sets and compared with that given by the state-of-the-art ones in the literature.

2 Our Appearance-based Body Template

Our appearance-based body template incorporates the colour texture information as priori into tracking. This allows our approach to avoid less robust silhouette fitting procedure and achieve better tracking performance. The textured body template in our work uses a standard articulated-joint parametrisation to

Fig. 1. From left to right: the articulated skeleton parameterised by 25 DOF, the visual hull of HumanEvaII Subject 4 constructed by Octree-based voxel colouring and the textured template model after manual refinements.

describe the human pose, further leading to an effective representation of the human motion over time. The articulated skeleton consists of 10 segments and is parameterised by 25 degrees of freedom (DOF) in Figure 1. It is registered to a properly scaled template skin mesh by Skeletal Subspace Deformation (SSD)[7]. Then, shape details and texture are recovered by the following procedure: Initially, Octree-based volumetric reconstruction with photo consistency [8] is used to recover the textured visual hull as shown in Figure 1. The vertices of the skin mesh are then registered to voxels by using Iterative Closest Point (ICP) method [9] with manual interactions to adjust misalignments. Colour textures are also assigned by corresponding skin vertices to textured voxels. The colour of a vertex that is invisible from all views is assigned to the same colour as the nearest visible vertex. At last, the template model is imported to commercial software to be finalised according to the real subject. The example of the final template model is illustrated in Figure 1.

3 Proposed Illumination Invariant and Error Oriented Evaluation

The proposed approach resides on the APF framework that is first introduced in markerless motion capture by Deutscher et al. [10]. Markerless motion capture can be formulated as a Recursive Bayesian filter framework [11]. It estimates a probability distribution recursively over time using incoming observations and temporal dependencies. Mathematically, this process can be expressed as

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} \tag{1}$$

where $\mathbf{x}_t$ denotes a pose configuration at time t, $\mathbf{y}_t$ an observation at time t, $\mathbf{x}_{t-1}$ the previous state, and $\mathbf{y}_{1:t}$ the collection of observations from time 1 to time t. Because the above integral does not have a closed form solution in general, the posterior distribution is often approximated by Sampling techniques. Simulated annealing [12] is incorporated in APF to maximise the observation likelihood

$p(\mathbf{y}_t|\mathbf{x}_t)$ that measures how well a particle (a pose configuration) $\mathbf{x}_t$ fits the observation $\mathbf{y}_t$ at time t. The observation likelihood is often formulated in a modified form of the Boltzmann distribution:

$$p(\mathbf{y}_t|\mathbf{x}_t) = \exp\{-\lambda E(\mathbf{y}_t, \mathbf{x}_t)\} \tag{2}$$

where the annealing variable λ is defined as $1/(k_B T_t)$, an inverse of the product of the Boltzmann constant k_B and the temperature T_t at time t. Maximising the observation likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ becomes minimising an energy function $E(\mathbf{y_t}, \mathbf{x_t})$. The optimisation of APF is iteratively done according to a predefined M-phase annealing schedule $\{\lambda = \lambda_1, ..., \lambda_M\}$, where $\lambda_1 < \lambda_2 < ... < \lambda_M$. At time t, considering a single phase m, initial particles are outcomes from the previous phase $m-1$ or drawn from the temporal model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$. Then, all particles are weighted by the observation likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$, and are resampled probabilistically to identify good particles that are highly likely to be near the global optimum. Subsequently, particles are perturbed to new positions for the next iteration by a Gaussian noise with a covariance matrix P_m. Eventually, particles are expected to concentrate on the neighbourhood of the global optimum of the posterior distribution, or equally, the energy function. As seen, the energy function $E(\mathbf{y}_t, \mathbf{x}_t)$ plays a critical role. The following sections, we improve the robustness, accuracy and computational efficiency in optimising $E(\mathbf{y}_t, \mathbf{x}_t)$ through Mutual Information maximisation and error-oriented gradual sampling.

3.1 Maximisation of Mutual Information

Fig. 2. From left to right: an observed image, and a synthesised image

Given multi-view image observations, the estimated pose, and the pre-built template human model, the likelihood evaluation is performed by comparing the observed images against the synthesised images. A synthesised image is obtained by projecting the template human model onto a mean static background image based on camera calibration parameters and a given pose configuration. Figure 2 shows an observed image and a synthesised image from a tracked example in the HumanEvaII dataset [13]. A direct comparison of the two images in the commonly used RGB colour space will not be robust and is often affected by lighting

conditions and appearance differences between the template model and the real subject. In this work, we employ robust image similarity metrics developed in the literature to overcome this problem. Especially, the work by Viola et al. [14] suggests that MI metrics are reliable for evaluating models with substantial different appearances and even robust with respect to variations of illumination. Mutual Information is a quantity that measures the mutual dependence of the two variables. Considering two images Ψ, Ω, and their pixels ψ, ω as random variables, Mutual Information can be expressed in terms of entropy as:

$$\begin{aligned} I(\Psi;\Omega) &= H(\Psi) + H(\Omega) - H(\Omega,\Psi) \\ &= -\sum_{\psi\in\Psi} p(\psi)\log p(\psi) - \sum_{\omega\in\Omega} p(\omega)\log p(\omega) + \sum_{\omega\in\Omega}\sum_{\psi\in\Psi} p(\psi,\omega)\log p(\psi,\omega) \\ &= \sum_{\omega\in\Omega}\sum_{\psi\in\Psi} p(\psi,\omega)\log\left(\frac{p(\psi,\omega)}{p(\psi)\,p(\omega)}\right) \end{aligned}$$

where $H(\Psi)$ and $H(\Omega)$ denote the marginal entropies, $H(\Psi,\Omega)$ the joint entropy. and $p()$ the probability density function. In our case, $p()$ is approximated by using the Parzen Window method with the Gaussian functions:

$$p(\psi) \approx \frac{1}{N}\sum_{\psi_i\in W}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\psi-\psi_i)^2}{2\sigma^2}\right)$$

$$p(\psi,\omega) \approx \frac{1}{N}\sum_{\psi_i\in W_\psi,\omega_i\in W_\omega}\frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}}\exp\left(-\frac{1}{2}\begin{bmatrix}\psi-\psi_i\\ \omega-\omega_i\end{bmatrix}^T\mathbf{\Sigma}^{-1}\begin{bmatrix}\psi-\psi_i\\ \omega-\omega_i\end{bmatrix}\right)$$

where N denotes the number of samples in the window W or W_ψ, W_ω. $\mathbf{\Sigma}$ is assumed as a diagonal covariance matrix. To be more robust to lighting conditions, MI is computed in the CIELab colour space in our work. Overall, the energy function $E(\mathbf{y}_t,\mathbf{x}_t)$ can be summarised as:

$$E(\mathbf{y}_t,\mathbf{x}_t) = \frac{1}{N_{view}}\sum_{i=1}^{N_{view}}\frac{1}{k_L I_L(IM^i_{\mathbf{y}_t};IM^i_{\mathbf{x}_t}) + k_a I_a(IM^i_{\mathbf{y}_t};IM^i_{\mathbf{x}_t}) + k_b I_b(IM^i_{\mathbf{y}_t};IM^i_{\mathbf{x}_t})} \quad (3)$$

where $IM^i_{\mathbf{y}_t}$ denotes the ith view observed image $\mathbf{y}_t$ at time t, $IM^i_{\mathbf{x}_t}$ the ith view synthesised image produced by projecting the estimate state $\mathbf{x}_t$ at time t, and $I_L()$, $I_a()$ and $I_b()$ the MI criterion values calculated in the channel L, a and b, respectively. Also, k_L, k_a and k_b denote the coefficients that control the weights of the L, a and b channels. Usually, k_L is set to be small in order to suppress the illumination influence.

3.2 Gradual Sampling Scheme for Annealed Particle Filter

The key ideas of our gradual sampling scheme are: 1) the precision of the energy function evaluations changes adaptively with the process of annealing. We blur observed and synthesised images at the early layers and use the blurred versions

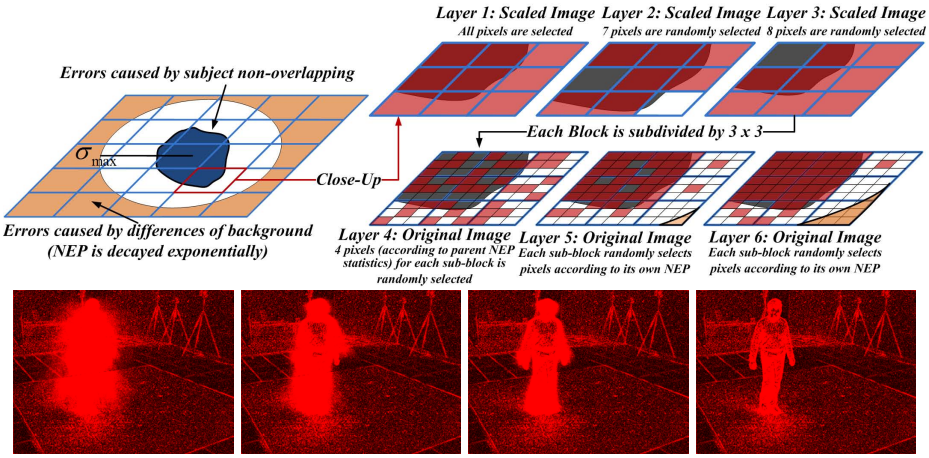

Fig. 3. The top diagram shows the error-oriented gradual sampling with 6 layers, and the bottom pictures illustrate that large errors among 200 particles gradually concentrate on the tracking subject through optimisation.

to evaluate the energy function. We observe that this is able to flatten the shape of the energy function, allowing a large number of particles to survive and encouraging a broader exploration. Then we gradually increase the resolution of the observed and synthesised images and use them in function evaluation at the later layers. This will provide more precise information to correctly single out those close to the global optimum; 2) With the increase of layers, the majority of large errors will progressively concentrate on the tracking subject and even only on some body areas where the observed and synthesised images have not been well aligned, as shown in the bottom of Figure 3. Taking this situation into account, we proposed a "smart" selection of the image areas (or extremely, the pixels) to be used in energy function evaluation. We weight different image areas based on the magnitude of the error from them in the last layer. Two criteria are used: i) if the error in an area is small (it means that the observed and synthesised have been well aligned in this region), we lower its weight; ii) if the area is far from the centre of the projection of the human body, we lower its weight. For an area with lower weight, we will sample a less number of its pixels to be included in the energy function evaluation. This is where the name "Gradual Sampling" comes from.

Following the above ideas, we partition the synthesised and observed images into a number of small-sized blocks, and compute the distribution of error over them. The error-oriented block selection and pixel sampling use the distribution from the last layer to guide the current function evaluation. This allows the evaluation to focus more on the large-error areas and at the same time reduce the number of pixels in evaluation. The whole process relies on a measurement

called Number of Effective Pixels [1] (NEP in short) defined for the ith block. In the layer m, the NEP for the ith block is expressed as:

$$NEP_{m,i} = \begin{cases} N_i \cdot \eta_{m,i} & ||c_i - C_{bo}||_2 < \sigma \\ N_i \cdot \eta_{m,i} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|c_i - C_{bo}\|_2^2}{2\sigma^2}\right) & ||c_i - C_{bo}||_2 \geqslant \sigma. \end{cases} \quad (4)$$

Here, $\eta_{m,i}$ is one factor controlling the number of pixels sampled from the ith block. It is proportional to the ratio of the average error from the ith block (denoted by $err_{m,i}$) to the maximal average error over all blocks (denoted by $\max_i(err_{m,i})$). In doing so, NEP_i for the block with smaller error will decrease relatively fast, whereas NEP_i for the block with greater error will maintain relatively large and decrease more slowly. This helps to gradually concentrate on misalignments. Also, considering that misalignments between the template model and real subject are gradually reducing with the annealing, we will progressively decrease the total number of pixels involved in function evaluation. This speeds up function evaluation and the tracking speed. To realise this, $\eta_{m,i}$ is also proportional to a predefined β_m, which decreases with the number of the layer, m. Therefore, $\eta_{m,i}$ is mathematically expressed as

$$\eta_{m,i} = \beta_m \cdot \frac{err_{m,i}}{\max_i(err_{m,i})}. \quad (5)$$

Equation (4) also takes the distance of the ith block from the tracking subject into account. We use a Gaussian distribution based weighting scheme to exponentially decay the importance of the ith block. The farther the centroid of the ith block (denoted by c_i) from the centroid of the human body template projected onto the image plane (denoted by C_{bo}), the less important this block is and the less the number of sampled pixels is. These makes the energy function evaluation can well concentrate on the error around the tracking subject. σ denotes the standard deviation of this Gaussian distribution. It will be empirically set at the beginning of the tracking based on the scattering radius of the projection of the human body template in a "T" gesture. We illustrate a typical procedure of the error-oriented pixel sampling in the top of Figure 3. Our gradual sampling procedure at time t is outlined in Algorithm 1.

4 Experiments and Discussion

Experiments are performed on the benchmark dataset [13], HumanEvaI that contains 4 grayscale and 3 colour calibrated video streams synchronised with Mocap data at 60Hz, and HumanEvaII that contains 4 colour calibrated image sequences synchronised with Mocap data at 60Hz. The tracking results are evaluated against the groundtruth Mocap data to obtain the absolute mean joint centre position errors and standard deviations. The experimental results with 50

[1] Pixels with large errors often correspond to misalignments of the subject, and they contribute to "effective" measurements.

Algorithm 1 Gradual Sampling for a typical frame at time t

Require: The survive rate α_m in APF [10], a set of predefined β_m, observation $\mathbf{y}_t$, the total number of layers M, and the initial covariance matrix $\mathbf{P}_0$.

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

1: Initialise N particles $\mathbf{x}_t^1, \cdots, \mathbf{x}_t^N$ from the last layer or the temporal model;

2: Evaluate $E(\mathbf{y}_t, \mathbf{x}_t)$ for each particle with the NEP_i pixels sampled from each block of blurred/original images. Average the error statistics over all particles;

3: Compute NEP_i for each block with the error statistics and equation (4)

4: Calculate λ_m by solving $\alpha_m N \sum_{i=1}^{N} (w_{t,m}^i)^2 = \left(\sum_{i=1}^{N} w_{t,m}^i\right)^2$ where $w_{t,m}^i = \exp\{-\lambda_m E(\mathbf{y}_t, \mathbf{x}_{t,m}^i)\}$ and N is the number of particles;

5: Update weights for all particles using $\exp\{-\lambda_m E(\mathbf{y}_t, \mathbf{x}_t)\}$.

6: Resample N particles from the updated importance weight distribution.

7: Perturb particles by Gaussian noise with covariance $\mathbf{P}_m = \mathbf{P}_{m-1}\alpha_m$.

end for

Study	Particles	Layers	Errors($ave \pm std$)mm	comments
[15]	250	15	32 ± 4.5	two-pass optimisation with smoothing
Ours	**200**	**10**	$\mathbf{54.6 \pm 5.2}$	**MI and Gradual sampling**
[16]	800	10	50-100	hierarchical approach
[17]	200	5	80 ± 5	Bi-directional silhouette-based
[18]	N/A	N/A	over 170	mix learning and tracking

Table 1. Absolute mean joint position errors on HumanEvaII Subject 4 from different research groups

particles and 10 layers on the 443-frame trail 1 of the subject 3 walking sequence in HumanEvaI is plotted in the top of Figure 4. The proposed method using only 3 colour video streams is able to maintain $64.5 \pm 8.2mm$ (using both MI and GS), $69.4 \pm 9.9mm$ (without using MI) and 68.6 ± 8 (without using GS). The difference among them demonstrate the effect of incorporating MI and GS in human motion tracking. All of the three methods outperform the silhouette-based method[2] $78 \pm 12.8mm$ using 7 video streams. To better verify the accuracy of our tracking result, we project the estimated poses onto another 4 novel views unused in our tracking algorithm. As shown in Figure 6, the projections match the figures of the real subject very well.

Another experiment uses 200 particles and 10 layers for the longer 1257-frame combo sequence of the subject 4 in HumanEvaII. The ground truth Mocap data is withheld by the data set owner and only available for online evaluations. We submitted our tracking result and obtained the online evaluation results as follows. As shown, in the middle of Figure 4 the proposed method is able to achieve $54.6 \pm 5.2mm$. Without the support of MI and GS, errors rise to $64.3 \pm 12.2mm$ and $59.38 \pm 5.5mm$, respectively. In contrast, the silhouette-based method with the same settings can only achieve $90.7 \pm 16.7mm$ and the maximum error reaches about $170mm$. This shows the advantage of our appearance-based likeli-

[2] The silhouette based method uses only the silhouette feature.

hood evaluation function over a silhouette-based one. Although jogging from the frames 400 to 800 is more difficult to track than slow movements of walking and balancing, the proposed method using MI can still track stably when other two methods without MI experience drastically fluctuation. As illustrated in Table 1, several research groups [15–18] have evaluated their results against the subject 4 of HumanEvaII. Our results achieve the second best performance overall. The work in [18] proposed a learning based approach with errors over $170mm$, which is relatively inaccurate when compared with APF based approaches. The work in [17] utilised bi-directional silhouette-based evaluation and achieved $80 \pm 5mm$. However, it relies on the quality of silhouette segmentation. The work in [16] proposed a hierarchical approach that employs a relative large number of evaluations to achieve errors within $50 - 100mm$. The method in [15] can be expected to perform better than ours because they utilise two-pass optimisation. In the second pass, a smoothing process with respect to future frames is used. These are not taken in our approach because two-pass optimisation incurs more computational overhead and limits its applicability to real-time tracking. Moreover, as pointed out in [16, 5, 18], when the error is less than $50mm$, the actual tracking errors will not be measurable because of the limited precision of the joint centres' positions estimated from the Mocap data, which is considered as ground truth, and the error between the human model and the real subject[3]. Hence, considering this context, our performance of $54.6 \pm 5.2mm$ has almost been the best possible. Also, this context explains why there has been about $50mm$ errors for our initial pose although it is accurately given. More results are presented in Figure 5 and supplementary material.

The performance experiment is set up on a dual core Windows system with 2.8GHz CPU and 4G RAM. The average one frame computational time is compared with the benchmark baseline algorithm[4] [17] which is the only publicly available implementation for the HumanEva dataset. We run both algorithms in different combinations of the layers and particles and the results are shown in the bottom of Figure 4. Despite of the extra computational overhead due to the use of Mutual Information criterion, our method with gradual sampling can still achieve almost 10 times faster than the baseline algorithm in [17].

5 Conclusion and Future Work

Our method introduces 1) A robust mutual information maximisation that utilises more visual information, reduces influences from illumination variations and supersedes unreliable image segmentation from the tracking pipeline; 2) A gradual sampling scheme for APF-based tracking framework to gradually focuses on resolving the major mismatches around the tracking subject and effectively allocate the computational resource accounting for the error distribution and

[3] Note that there are no markers corresponding to actual joint centres in the Mocap data. As a result, the joint centres' positions cannot be recovered very accurately from the Mocap data.

[4] Available online via http://vision.cs.brown.edu/humaneva/baseline.html

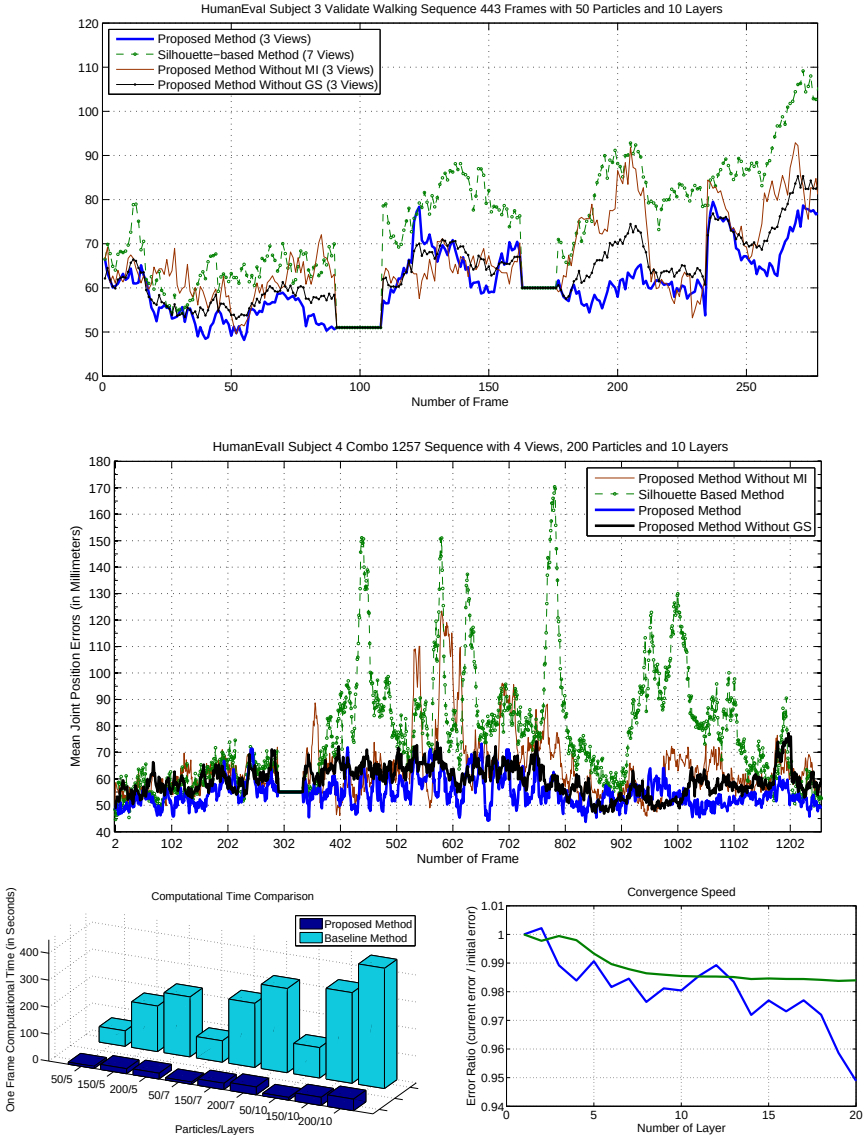

Fig. 4. From the top to bottom, 1) tracking results on HumanEvaI the Subject 3 Walking Sequence, 2) the Subject 4 of HumanEvaII Combo Sequence, 3) computational time comparison for the proposed and baseline method with different particles and layers, and 4) the convergence speed of Gradual Sampling versus the Common APF method. Note that the ground truth data is corrupted at 91-108 and 163-176 frames on HumanEvaI and at 298-335 frames on HumanEvaII.

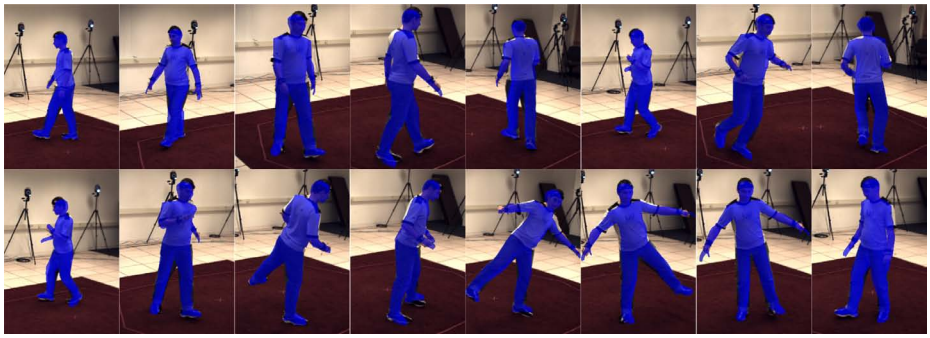

Fig. 5. Accurate tracking results from HumanEvaII Subject 4. The average Euclidean error of 3D joint positions is less than 55mm comparable to the best result in [15].

the hierarchical order of the human anatomy. Experiments with the benchmark datasets demonstrate that our method has very competitive performance among the other state-of-the-art studies from different research groups. Markerless motion capture in our research can be further improved in future work. The acquisition of the appearance based human body model using a few multi-view images currently heavily relies on manual interactions. Also, the global optimisation is one major bottleneck preventing markerless motion capture from real time applications. These open problems indicate the directions of our future research.

Acknowledgement. Authors would like to thank the support from National ICT Australia, and Leonid Sigal from Brown University mades the HumanEva dataset available.

References

1. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: CVPR. (2007)
2. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: SIGGRAPH, ACM (2008) 1–10
3. Starck, J., Hilton, A.: Surface capture for performance based animation. IEEE Computer Graphics and Applications **27(3)** (2007) 21–31
4. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE TPAMI **16** (1994) 150–162
5. Corazza, S., Mndermann, L., Gambaretto, E., Andriacchi, T.P.: Markerless motion capture through visual hull, articulated icp and subject specific model generation. International Journal of Computer Vision **87** (2010) 156–169
6. Shannon, C.E.: A mathematical theory of communication. Bell Systems Technical Journal **27** (1948) 379–423 Continued 27(4):623-656, October 1948.
7. Magnenat-Thalmann, N., Laperrière, R., Thalmann, D.: Joint-dependent local deformations for hand animation and object grasping. In: Proceedings on Graphics interface '88, Canadian Information Processing Society (1988) 26–33

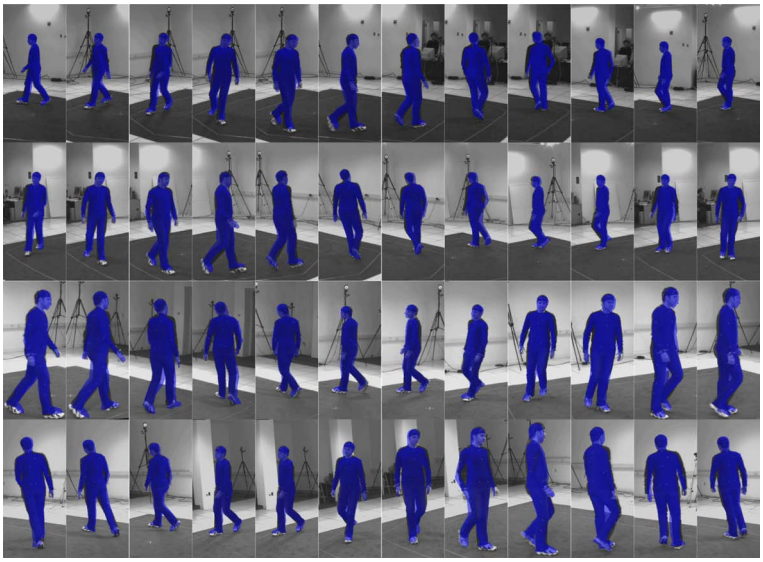

Fig. 6. Visual results from HumanEvaI Subject 3 show 4 novel camera views which are not used in tracking.

8. Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. In: CVPR, IEEE Computer Society (1997) 1067
9. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE TPAMI **14** (1992) 239–256
10. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR. Volume 2. (2000) 126–133 vol.2
11. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. Statistics and Computing **10** (2000) 197–208
12. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science, Number 4598, 13 May 1983 **220, 4598** (1983) 671–680
13. Sigal, L., Black, M.J.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, Department of Computer Science (2006)
14. Viola, P., III, W.M.W.: Alignment by maximization of mutual information. International Journal of Computer Vision **24** (1997) 137–154
15. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture - a multi-layer framework. IJCV **87** (2010) 75–92
16. Bandouch, J., Beetz, M.: Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In: IEEE ICCV Workshop on Human-Computer Interaction (HCI). (2009)
17. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision (2010)
18. Cheng, S.Y., Trivedi, M.M.: Articulated human body pose inference from voxel data using a kinematically constrained gaussian mixture model. In: CVPR Workshop on EHUM2, IEEE (2007)