Tracking Generic Human Motion via Fusion of Low- and High-Dimensional Approaches

Yuandong Xu
xuyd@cis.pku.edu.cn
Jinshi Cui
cjs@cis.pku.edu.cn
Huijing Zhao
zhaohj@cis.pku.edu.cn
Hongbin Zha
zha@cis.pku.edu.cn

Key Laboratory of Machine Perception
(The Ministry of Education, China)
Peking University
Beijing, China

Abstract

Tracking generic human motion is significantly challenging because of the high-dimensional state space as well as various motion types. In order to deal with the challenges, we propose a fusion formulation to integrate the low- and high-dimensional tracking approaches into one framework. The low-dimensional approach successfully overcomes the high-dimensional problem on tracking the motions with available training data by learning motion models. On the other hand, the high-dimensional approach is employed to recover the motions without learned models by sampling directly in the pose space. Within the framework, the two parallel approaches are fused by a set of criteria at each time step. The fusion criteria ensure that the overall performance of the system is improved by concentrating the advantages of the two approaches and avoiding their weak points. Experimental results with qualitative and quantitative comparisons demonstrated that the proposed formulation can fully display the advantages of different algorithms and effectively track generic human motion.

1 Introduction

3D human motion tracking technology has gained a lot of attentions in recent years because of its potential applications on smart surveillance systems, advanced human-computer interfaces, markerless motion capture, etc. In this research, there are two primary challenges. The algorithm should be able to cope with the high-dimensional state space as well as to recover complex postures with various motion types and styles. Many approaches have been proposed to address these problems [3, 4, 8, 15, 17, 18, 19]. One kind of low-dimensional approaches that learn motion models by dimensionality reduction can successfully deal with the high-dimensional problem, but it only works on specific motion types with available training data. Other approaches which employ smart sampling directly on high-dimensional pose space don't have that limitation. However, this kind of methods is not robust enough and has high computational cost.

BMVC 2011 http://dx.doi.org/10.5244/C.25.57

In order to solve the aforementioned problems simultaneously, we propose a fusion formulation to integrate the two kind of approaches into one framework. Within the framework, two independent trackers with different algorithms proceed in parallel in different state spaces, and are fused according to a set of criteria at each time step. The essence of this fusion is that the two trackers could cooperate and complement each other to deal with various problems in generic motion tracking. The main contribution of this paper is that we solved the two following issues in the fusion procedure: (1) How to adapt algorithms automatically according to the motion sequence and make them fully display their respective advantages. (2) How to make multiple algorithms complement each other to improve the overall tracking performance. The fusion of multiple approaches makes our tracking system outperform any system that uses single approach.

The remainder of this paper is organized as follows: in the next section, related work is briefly reviewed and discussed. Section 3 and 4 introduce the high-dimensional tracking approach as well as the low-dimensional approach respectively. In Section 5, details of our fusion framework are provided. Experimental results are presented in Section 6 and the paper is finally summarized in Section 7.

2 Related Work

The related works will be mainly divided into three categories for reviewing. One category of approaches employed sampling-based tracking techniques directly in the high-dimensional pose space, for example, particle filtering [15], partition sampling [11] and annealed particle filtering [3]. This kind of approaches can recover 3D human poses without restriction on motion types. However, it requires a large number of particles to search the optimal mode in the pose space as well as it's easy to fail under conditions of a loose-fitting body model and noisy background lacking rich observations [12].

Another category of methods employ motion models to guide human tracking. For this kind of approaches, a low-dimensional space is learned by applying dimensionality reduction on the Mocap training data and the human poses are recovered by searching in the low-dimensional space. A variety of existing dimensional reduction algorithms are considered, such as Principal Component Analysis (PCA)[19], Locally Linear Embedding (LLE)[8], Gaussian Process Latent Variable Model (GPLVM) [18], Gaussian Process Dynamical Model [17] and dynamical binary latent variable models [4]. Tracking with learned motion models can successfully recover 3D human poses, but the drawback is that it cannot be generalized out of the set of training data. As improved solutions, the works of [2] and [10] employ a hierarchical decomposition of the motion model by H-GPLVM. Hierarchical searching is used through the latent spaces for finding the optimal human poses. This method can track more types of motions given a set of "basic activity" training data. However, we cannot make the algorithm generalization by just increasing the complexity of the motion models or increasing the types of motion models, because of the following limitations: (1) The motions with available training data are always limited. (1) The computational cost would increase dramatically as excessive competition from the redundant models.

As the third kind of approach, the recent work [9] combines low- and high-dimensional motion models together to track human poses. In its framework, a low-dimensional joint-activity space is learned with the training data, and the other "unknown" motion types without training data are modeled in high-dimensional space. Variable particle numbers are allocated for each motion model, and the annealed particle filtering is employed for searching the optimal mode. This work obtains good performance on tracking generic motion sequences. Nevertheless, the dimensionality of the joint-activity space must grow and the computational cost will increase with the number of "known" activities. Moreover, there's no re-initialization mechanism to guarantee that the particles in the joint space are in the proper positions when the pose transfers from the high-dimensional model of an "unknown" activity to the low-dimensional model of a "known" one.

That is to say fusion of multiple techniques is a promising way to solve tracking problems. Instead of uniting various motion models together [9], we propose a formulation to integrate the low- and high-dimensional approaches into one framework. Within the framework, the power of each approach can be exploited. Details of the proposed method will be introduced in the following sections.

3 Tracking in the high-dimensional pose space

Tracking human motions directly in the original pose space can handle unconstraint types of motions, while the high dimensional problems should be addressed. Many smart sampling approaches have been proposed to deal with this problem. Among these methods, the annealed particle filtering (APF)[3] obtains a relatively good performance and is often used as a baseline algorithm [10, 16].

The APF employs a set of N weighted particles, $\{(x^1,w^1),(x^2,w^2),\dots,(x^N,w^N)\}$ to approximate the target distribution over full 3D pose space. For each time step, the APF attempts to find optimal mode by "cooling" the target distribution and then "warming" it gradually through a number of successive re-sampling iterations (or layers).

At each layer $m = M, M-1, \dots, 1$, the particles are dispersed by a dynamical model $p(x_t|x_{t-1})$, and evaluated against the observation y by a weighting function $w_m(x,y)$, where

$$w_m(x,y) = w(x,y)^{\beta_m}, \tag{1}$$

for $\beta_1 > \beta_2 > \dots > \beta_M$. A large β_m produces a peaked weighting function w_m, resulting in a high rate of annealing. Small values of β_m have the opposite effect. The whole effect is to gradually concentrate particles into the globally optimal mode of the target distribution.

The standard APF often uses the addition of Gaussian noise to approximate the dynamical model $p(x_t|x_{t-1})$. Since the state space is quite large and high-dimensional while the searching range permitted is quite small, the tracking could be easy to fail and hard to recover if the standard APF falls into incorrect mode [1].

4 Tracking in the low-dimensional latent space

Reducing the dimensionality of the state space is a successful approach to deal with the high dimensional problems of human motion tracking. The idea is based on the facts that: the space of possible human motions is intrinsically low-dimensional [6, 14] and the set of typical human poses is far smaller than the set of kinematically possible ones [5]. Many recent works proposed to learn low-dimensional motion models and track human motions in the low-dimensional space.

4.1 Tracking specific activities by learning motion models

For successfully tracking a specific human motion, the learned models should contain rich prior of the high-dimensional poses and dynamics. Many dimensionality reduction techniques have been employed to learn motion models. Among all of them, the GPDM is one of the most effective approaches because it has smooth latent embedding and generalizes gracefully to motions outside the training dataset [17].

The motion model learned by the GPDM contains a temporal dynamics in the latent variable space,

$$z' = f_{gp\_dyn}(z), \tag{2}$$

and a static mapping for the high-dimensional pose recovery from the latent variable space,

$$x = f_{gp}(z) \tag{3}$$

where z represents the position in the latent space and x represents the corresponding high-dimensional pose.

In the tracking step, we take a similar framework as the GP-APF [13]. A set of N weighted particles $(z^1, w^1), (z^2, w^2), \ldots, (z^N, w^N)$ is used to approximate the posterior distribution over the latent pose positions. The proposal distribution of the temporal dynamics is defined as

$$p(z_t | z_{t-1}) = N(z_{t-1} + (f_{gp\_dyn}(z_{t-1}) - z_{t-1})\Delta_T,\ \sigma_{dyn}) \tag{4}$$

where Δ_T is the time interval which satisfied a Gaussian distribution, and σ_{dyn} is the variance of the prediction errors, computed by the GPDM dynamics. After the dynamical prediction, the pose hypotheses are recovered by the GPDM static mapping of Equation (3), and then evaluated by the weighting function.

In order to give an estimation as close as possible to the actual pose, we add an additional annealing layer at the last step to optimize the poses in the original pose space, as [13]. Because the models learned by the GPDM provides rich motion prior and the searching space is relatively low-dimensional, only a small amount of particles can achieve satisfactory tracking performance.

4.2 Switching Multiple Motion Models by mixed-state CONDENSATION

For tracking the motion sequence containing several activities, the motion models are learned respectively for each activity by the GPDM. We need a mechanism to support multi-model switching when the motion transfers from one activity to another, because each motion model has its distinct latent variable space. As a solution, the mixed-state CONDENSATION [7] provides an effective approach for tracking with multiple dynamical models.

For the first step, we should learn the transition mappings between any two motion models. For any two activities, e.g. walking and jogging, two sets of training pairs are generated by looking for the most similar poses in one dataset, given a pose belonging to the other dataset and vice versa. A pair of transition mappings between the latent spaces of the two motion models are learned using Relevant Vector Machine (RVM). The similarity between two poses is defined as proportional to the cosine similarity of the pose vectors, and inversely proportional to the Euclidean distance of the pose vectors. In practice, we find that better matching result would be achieved if considering the angles of key joints only and remove leaf joints from the original pose vector when computing the similarity.

We define our mixed-state space as $[z, a]$, where z denotes the latent position and a denotes the activity label. The dynamical propagation is decomposed as follows,

$$p(z_t, a_t | z_{t-1}, a_{t-1}) = p(z_t | a_t, z_{t-1}, a_{t-1}) p(a_t | z_{t-1}, a_{t-1}) \tag{5}$$

$$p(a_t | z_{t-1}, a_{t-1}) : p(a_t = i | z_{t-1}, a_{t-1} = j) = T_{ij}(z_{t-1}). \tag{6}$$

where the T_{ij} are transition probability from activity i to activity j. If we assume that T_{ij} is independent of the pose, the transition matrix is invariant and can be determined by statistical estimation at the beginning of the tracking.

For every tracking step, the activity labels are determined by sampling from the transition matrix first. If any particle changes its activity label, the latent position is reset in the new motion space by the transition mappings. Then, the dynamical propagation in the latent space is proceeded as the same process as tracking with single motion model. The mixed-state CONDENSATION can adapt the number of particles belonging to each motion model by weighting and re-sampling.

5 Fusion of Low- and High-Dimensional Approaches

Though the learned motion models can reduce motion ambiguities and enhance tracking accuracy and stability, a limited number of motion models cannot get good results due to the complexity and uncertainty of human motion. On the other hand, annealed particle filtering could track motions of unconstraint types, but it is lack of robustness, with high computational cost, and hard to recover from failures. Therefore, we propose to a method to integrate the model learning approach with the standard APF into one framework. Our goal is to track generic human motions without type constrains as stably and effectively as possible.

5.1 Fusion Strategy Overview

Figure 1 shows an overview of our fusion framework. Within the framework, two independent trackers with different algorithms run in parallel and are fused by a set of criteria at each time step.

The first tracker (denoted by the GPDM-APF) employs the learned motion models to track human motions. A smaller amount of particles are allocated for it because the latent state space is low-dimensional. The second tracker takes the standard APF algorithm to recover human poses in the high-dimensional pose space, with a larger amount of particles. After each time step, the quality of the performance is evaluated for each tracker by applying the cost function on each expectation output. The criteria is established to take the result of the winning tracker as the current output of the system and update the state of the other route if necessary. Therefore, when the body performs a trained activity, the system prefers tracking with learned motion models, and the state of the standard APF will not deviate beyond a certain range. When the body performs motions of un-trained motions, the standard APF takes over the tracking and supplements the tracker with learned motion models if necessary.

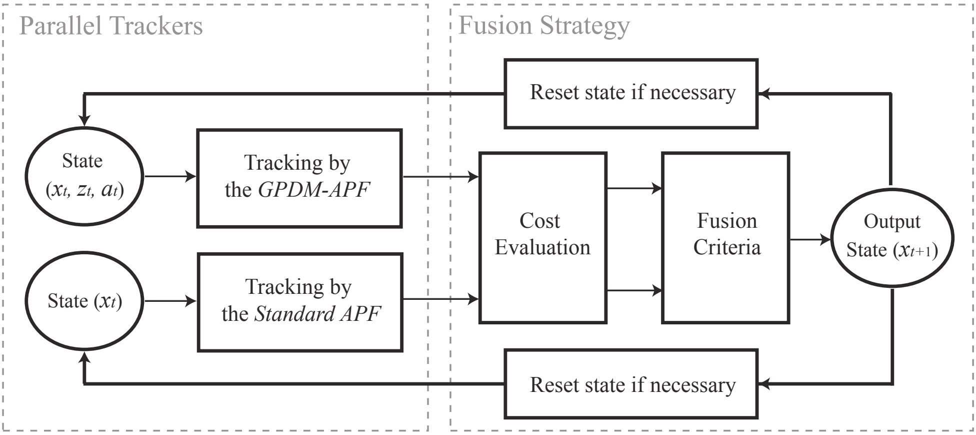


Figure 1: **An overview of the fusion framework.** We integrate the two parallel trackers (i.e. the GPDM-APF and the standard APF) and make them complement each other by the fusion criteria.

5.2 Cost Function and Criteria

The cost function is used for the comparison of pose hypotheses with image observations. The particle weights w are also calculated by exponentiation on the values of the cost function. The definition of cost function is taken from the symmetrical silhouette likelihood used by Sigal *et al.* [16], which penalizes non-overlapping regions for both silhouettes of foreground and pose projection.

Let $F(p)$ represent the observation foreground and $M(p)$ the silhouette map of projection model. The cost is computed as,

$$Cost = \frac{1}{N}\sum_{n}\left(\frac{\sum_p (F(p)(1-M(p)))}{\sum_p (F(p))} + \frac{\sum_p (M(p)(1-F(p)))}{\sum_p (M(p))}\right). \tag{7}$$

where N is the number of camera views .

During the tracking, we apply the cost function on the pose expectation of each tracker after each time step, and the cost value is used to evaluate the tracking quality of the tracker for the current frame. Let $cost_{(SAPF)}$ denote the cost of the standard APF and $cost_{(GPDM-APF)}$ the cost of tracking with the GPDM models, the criteria for choosing the output and updating the trackers is set as follows:

If $cost_{(SAPF)} \geq cost_{(GPDM-APF)}$, the output is set as the expected pose given by the GPDM-APF. Conversely, the output is set as the expected pose given by the standard APF.

If $cost_{(SAPF)} - cost_{(GPDM-APF)} > \delta$, the deviation of the standard APF exceeds the threshold δ. The state of the particles from the standard APF is set to the expected pose of the GPDM-APF, and the corresponding weights are all reset as equal.

If $cost_{(GPDM-APF)} - cost_{(SAPF)} > \delta'$, the deviation of the GPDM-APF exceeds the threshold δ'. This is often the case when an un-trained activity occurs and the learned GPDM models cannot handle it. We need to reset the latent positions of the particles in the GPDM-APF in order to guarantee that the GPDM-APF can work when the motion transfers back to the trained activities. Suppose the expected pose of the standard APF is denoted by x_{APF}, the new particle set of the GPDM-APF, which contain poses, latent positions and activity labels, are created by selecting the top k nearest neighbor of x_{APF} from the training dataset of each activity. The weights for the particles are set to the similarity between the corresponding poses and x_{APF}.

Since the GPDM-APF is preferred for tracking the sequences of trained activities, we set that $\delta' \geq \delta$ in the criteria. In the following experiments, the values of the thresholds δ and δ' are determined manually. However, we find that the performance are better than that of using just standard APF or just GPDM-APF even when both thresholds are set equal to zero.

6 Experiments

In order to investigate the performance of our technical fusion approach to generic motion tracking, we design two experiments to track the *HumanEva-II Combo* sequences. The results are evaluated by the online evaluation system using the 3D absolute error defined in [16]. The cost of the parallel trackers at each frame are also provided to prove the correctness of our fusion criteria. The experiments also conducted quantitative comparisons with the methods using only the standard APF or the GPDM-APF.

6.1 Basic Test on Generic Motion Sequences

In the first experiment, we test our fusion approach on the *HumanEva-II S2 Combo* motion sequence. This sequence contains three activities, i.e. walking, jogging and balancing. For the first two activities, the training data of *HumanEva-I S2* are used to learn the motion models. However, no training data is available for the balancing motion.

For the experimental settings, 4 camera views are used for computing the likelihood. We assigned 80 particles with 3 annealing layers for the GPDM-APF and 150 particles with 4 layers for the standard APF. The distribution of the time interval Δ_T in the GPDM-APF is set to $N(1, 0.3)$. The sampling covariance of the standard APF is learned using the *HumanEva-I* training data. For the fusion criteria, we set $\delta = 0$, $\delta' = 0.01$ and $k = 10$. The cost, as the quality representation for each tracker is represented in Figure 2.

Figure 3 shows the tracking performance of our fusion approach, with quantitative comparison to the other two methods using only the standard APF or the GPDM-APF. The output poses are shown on the images of *HumanEva-II S2* camera *C1* in Figure 4. As the results demonstrated, our method could work properly on the kindred motion sequences partly with trained models.

6.2 Extended Test on Generic Motions with Various Styles

In the second experiment, we considered the extensibility of our fusion formulation on tracking motions with various styles. The test sequence is from *HumanEva-II S4 Combo*, which contains the activities of walking, jogging and balancing. However, no training data is available for subject *S4*. We used the learned walking and jogging models of subject *S2* instead. Note that the activity styles of $S4$ are very different from the ones of $S2$.

The experiment is conducted under the same parameters as the first one. The cost of each tracker during the tracking process is represented in Figure 5. The quantitative comparisons between different approaches are shown in Figure 6, and the output poses are visualized on the images of *HumanEva-II S4* camera *C1* in Figure 7.

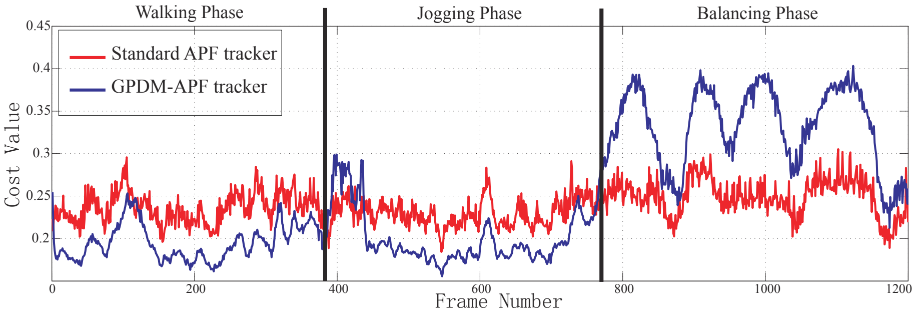


Figure 2: **Cost of the two parallel trackers in the tracking of** $S2$**.** We choose the tracker with minimum cost at each frame, and update the state of the other one. Note that the standard APF tracker has lower cost during the transition from *Walking* to *Jogging*. This not only produces smooth pose output, but also helps the GPDM-APF tracker switch motion models.

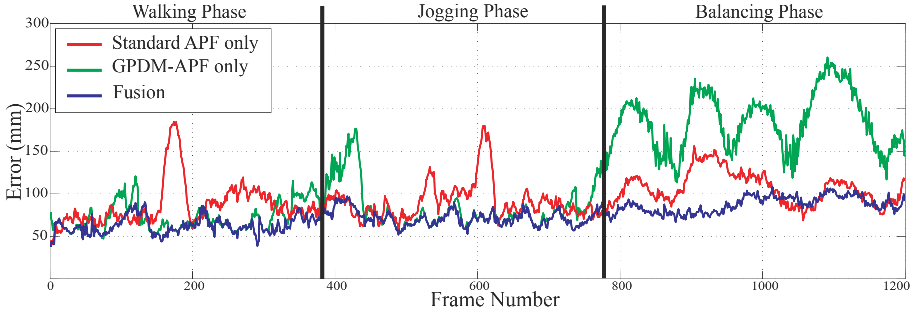


Figure 3: **Performance comparison between the fusion approach and the other two methods on tracking** $S2$**.**

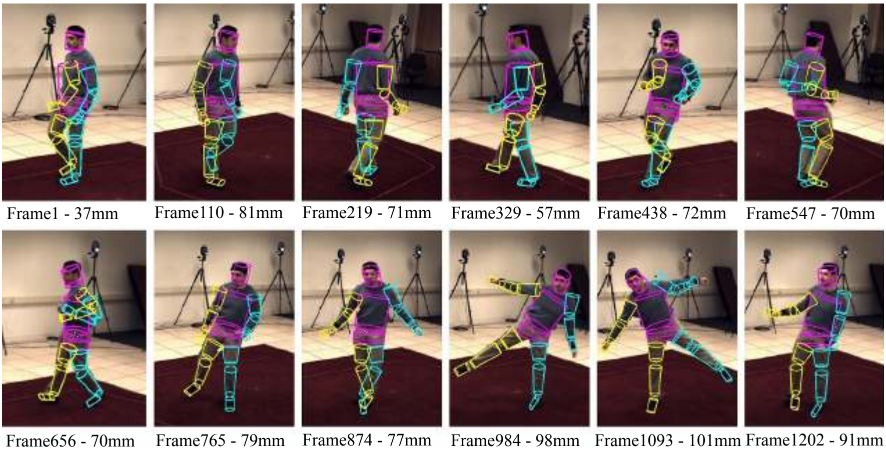

Frame1 - 37mm
Frame110 - 81mm
Frame219 - 71mm
Frame329 - 57mm
Frame438 - 72mm
Frame547 - 70mm
Frame656 - 70mm
Frame765 - 79mm
Frame874 - 77mm
Frame984 - 98mm
Frame1093 - 101mm
Frame1202 - 91mm

Figure 4: **Results of tracking** $S2$ **by the fusion approach.** The recovered body model is shown projected into the images, with the corresponding 3D error shown underneath.

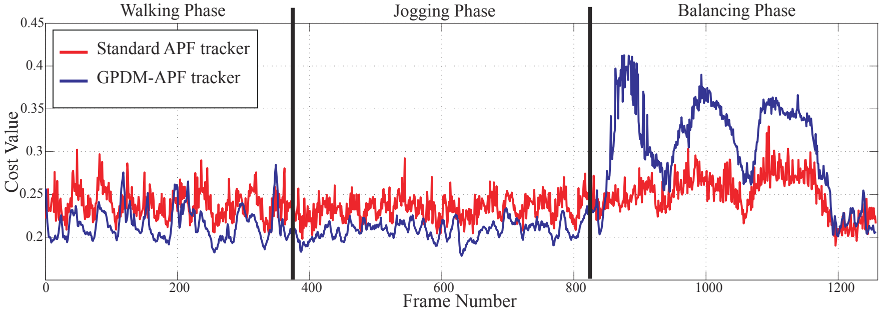


Figure 5: **Cost of the two parallel trackers in the tracking of *S*4.** Since the motion styles of the learned model are different from subject *S*4, the averaging gap between the cost of the two trackers in the walking and jogging phases are smaller than that in figure 2. At the end of the sequence, the cost of the GPDM-APF tracker is lower than the other. That is reasonable because the subject stops balancing and starts walking out of the scene.

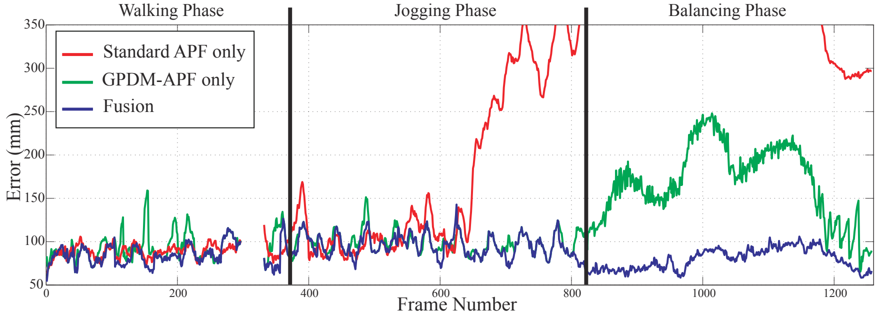
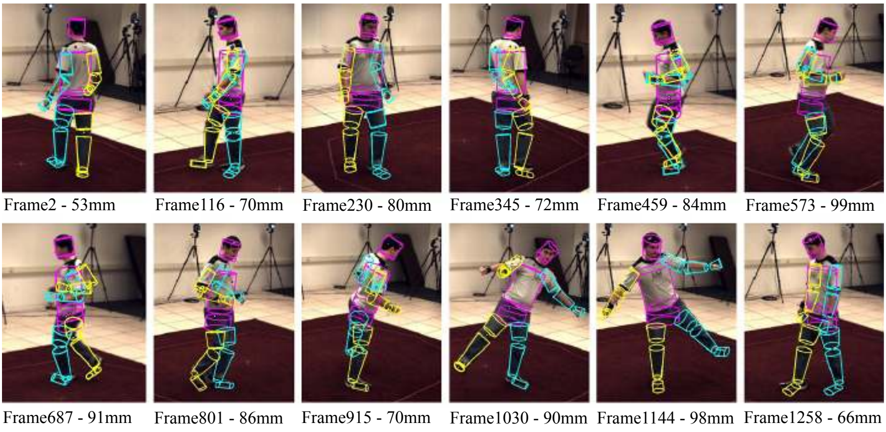


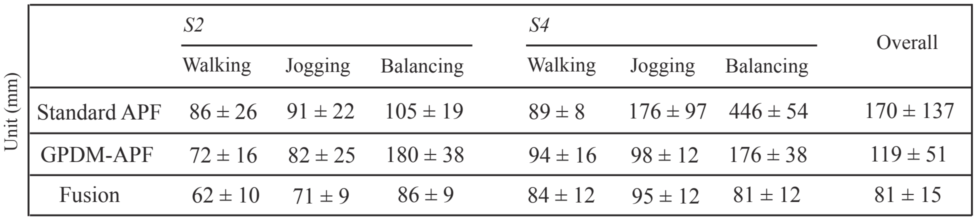
Figure 6: **Performance comparison between the fusion approach and the other two methods on tracking *S*4.**



Frame2 - 53mm Frame116 - 70mm Frame230 - 80mm Frame345 - 72mm Frame459 - 84mm Frame573 - 99mm

Frame687 - 91mm Frame801 - 86mm Frame915 - 70mm Frame1030 - 90mm Frame1144 - 98mm Frame1258 - 66mm

Figure 7: **Results of tracking *S*4 by the fusion approach.** The recovered body model is shown projected into the images, with the corresponding 3D error shown underneath.



Unit (mm)	*S2*			*S4*			Overall
	Walking	Jogging	Balancing	Walking	Jogging	Balancing	
Standard APF	86 ± 26	91 ± 22	105 ± 19	89 ± 8	176 ± 97	446 ± 54	170 ± 137
GPDM-APF	72 ± 16	82 ± 25	180 ± 38	94 ± 16	98 ± 12	176 ± 38	119 ± 51
Fusion	62 ± 10	71 ± 9	86 ± 9	84 ± 12	95 ± 12	81 ± 12	81 ± 15

Figure 8: **The average error and standard deviation for the sequences produced by the standard APF, the GPDM-APF and the Fusion approach.**

Finally, Figure 8 compares the average error and standard deviation for the testing sequences produced by our fusion approach and the other two methods.

7 Conclusion and Future Work

In this paper, we presented a novel fusion formulation to integrate the low- and high-dimensional approaches into one framework. The proposed formulation not only incorporates the respective advantages of the two approaches, but also overcome their weakness. The experimental results demonstrate that our approach can effectively track generic human motion with various types and styles. For the computational cost, the overall efficiency only depends on the slower tracker, i.e. the standard APF, because the time consumption for the fusion is very small and the two trackers are in parallel. We consider that the computational cost can still be improved in the future.

The fusion criteria are very easy to be extended to support a variety of rules and strategies. In our experiments, basic rules are defined based on the results of the cost function and manually set thresholds. However, more flexible fusion criteria could be probably embedded by considering the tracking history. In the future, an online learning module will be provided for better fusion guidance.

References

[1] Alexandru O. Balan, Leonid Sigal, and Michael J. Black. A quantitative evaluation of video-based 3d person tracking. In *International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 349–356, 2005.

[2] John Darby, Baihua Li, Nicholas Costen, David Fleet, and Neil Lawrence. Backing off: Hierarchical decomposition of activity for 3d novel pose recovery. In *British Machine Vision Conference (BMVC)*, 2009.

[3] Jonathan Deutscher and Ian Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61:185–205, 2005.

[4] David J. Fleet Graham W. Taylor, Leonid Sigal and Geoffrey E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *International Conference on Computer Vision (ICCV)*, pages 641–648, 2003.

[6] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popovich. Style-based inverse kinematics. In *SIGGRAPH*, 2004.

[7] Michael Isard and Andrew Blake. A mixed-state condensation tracker with automatic model-switching. In *International Conference on Computer Vision (ICCV)*, 1998.

[8] Tobias Jaeggli, Esther Koller-Meier, and Luc Gool. Learning generative models for multi-activity body pose estimation. *International Journal of Computer Vision*, 83: 121–134, 2009.

[9] Baihua Li John Darby and Nicholas Costen. Tracking human pose with multiple activity models. *Pattern Recognition*, pages 3042–3058, 2010.

[10] Michael Rudzsky Leonid Raskin and Ehud Rivlin. 3d human body-part tracking and action classification using a hierarchical body model. In *British Machine Vision Conference (BMVC)*, London, 2009.

[11] John Maccormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2000.

[12] Patrick Peursum. On the behaviour of body tracking with the annealed particle filter in realistic conditions. In *Technical report*, 2006.

[13] Leonid Raskin, Ehud Rivlin, and Michael Rudzsky. Using gaussian process annealing particle filter for 3d human tracking. *European Journal of Advanced Signal Process (EURASIP)*, pages 1–13, 2008.

[14] Alla Safonova, Jessica K. Hodgins, and Nancy S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Trans. Graph.*, 23:514–521, 2004.

[15] Hedvig Sidenbladh, Michael J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision (ECCV)*, pages 702–718, 2000.

[16] Leonid Sigal, Alexandru Balan, and Michael Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2010.

[17] Raquel Urtasun. 3d people tracking with gaussian process dynamical models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[18] Raquel Urtasun and David J. Fleet. Priors for people tracking from small training sets. In *International Conference on Computer Vision (ICCV)*, 2005.

[19] Raquel Urtasun, David J. Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer Vision and Image Understanding*, 104:157–177, 2006.