A Swarm Intelligence Based Searching Strategy for Articulated 3D Human Body Tracking

Xiaoqin Zhang[1,2], Weiming Hu[1], Xiangyang Wang[3], Yu Kong[4], Nianhua Xie[1], Hanzi Wang[5], Haibin Ling[6], Steve Maybank[7]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Beijing, China
[2]College of Mathematics & Information science, Wenzhou University, Zhejiang, China
[3]School of Communication and Information Engineering, Shanghai University, Shanghai, China
[4]School of Computer Science, Beijing Institute of Technology, Beijing, China
[5]School of Computer Science, University of Adelaide, Australia
[6]Department of Computer and Information Science, Temple University, Philadelphia, USA
[7]School of Computer Science and Information Systems, Birkbeck College, London, UK

Abstract

This paper proposes an annealed particle swarm optimization based particle filter algorithm for articulated 3D human body tracking. In our algorithm, a sampling covariance and an annealing factor are incorporated into the velocity updating equation of particle swarm optimization (PSO). The sampling covariance and the annealing factor are initiated with appropriate values at the beginning of the PSO iteration, and 'annealing' is carried out at reasonable steps. Experiments with multi-camera walking sequences from the Brown dataset show that: 1) the proposed tracker can effectively alleviate the problem of inconsistency between the image likelihood and the true model; 2) the tracker is also robust to noise and body self-occlusion.

1. Introduction

3D articulated human body tracking is the task of determining the location, orientation and scale of each body part (i.e. head, torso, upper/lower arms, and upper/lower legs) in an image sequence. It is important for many vision understanding applications, e.g. human-computer interaction, visual interactive gaming, immersive virtual reality, etc.

However, it remains a challenging task for the following reasons. 1) High dimensional and nonlinear state space (usually 30+ dimensions, see Fig.1[1]); the search for the optimal model in such a high dimensional state space, if not impossible, involves a huge computation time. 2) Unknown image background and presence of clutter. The above two difficulties cause the most challenging problem: **inconsistency between the global optimum of the image likelihood and the ground truth** (see Fig.2). As shown in Fig. 2, the likelihood value of the ground truth is far from the likelihood value of the global optimum. This phenomenon degrades the performance of traditional optimization methods in 3D articulated human body tracking, such as gradient descent [1], particle filter [2].

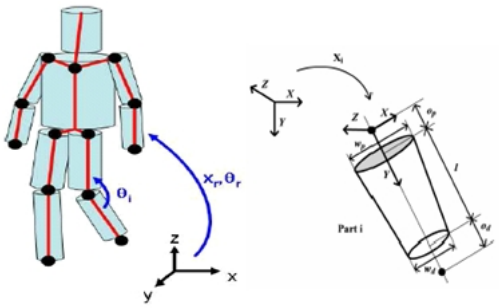

Figure 1. The human body model is represented as a 3D kinematic tree.

Generally, there are two major ways to alleviate the above problem. The first one is to define a more discriminative and robust observation model which is more consistent with the ground truth. The second one is to use more intelligent searching strategies and to confine the searching space to the region near the ground truth.

In [3], the visual cues such as face detection, head-shoulders contour matching and elliptical skin-blob detection are utilized to generate a discriminative likelihood model for 3D pose estimation. Balan and Black [4] develop a new image likelihood function based on based on the

[1]This figure is from [19]

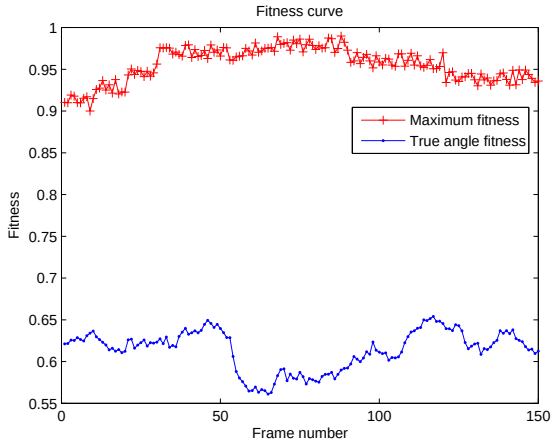

Figure 2. The image likelihood curve of ground truth and MAP (maximum a posteriori) state (here, the most commonly used likelihood model [9] is adopted)

Wandering-Stable-Lost appearance model [5]. These likelihood models are only useful in their constraint environments. However, it can not be extended to the general cases. Moreover, it is very difficult to obtain a likelihood model that is discriminative enough for every situation. Therefore, many efforts are made to use more intelligent searching strategies to confine the search space to a smaller region which contains the groud truth. Deutscher and Reid [6] propose an annealed particle filter (APF) algorithm, which adopts a deterministic annealing approach with stochastic sampling to confine particles to a reasonable search space. Meanwhile, a crossover operation is conducted to maintain the diversity of the particles. Compared with the regular particle filter (PF), APF greatly improves the tracking performance. Xu et al. [7] use a Rao-Blackwellised particle filter (RBPF) for articulated human body tracking. The motion correlation between the left and right parts of the body is learned by partial least squares regression. By incorporating the correlation model into the RBPF tracking framework, the sequential importance sampling is implemented in a reduced state space. As a result, both the accuracy and the efficiency of tracking are improved. Darby et al. [8] propose a human motion tracking algorithm using multiple pre-trained activity models for propagation of particles in annealed particle filtering. A quantitative evaluation of video-based 3D person tracking is conducted in [9]. In their work, the inference of different likelihood functions, the temporal dynamics and priors, the number of particles and the number of camera views are investigated. However, all the above work ignores the most important problem that degrades the performance: **the inconsistency between the global optimum of the image likelihood and the ground truth**.

An alternative solution to 3D articulated human body tracking is the manifold-based learning method: such as the GPLVM (Gaussian Process Latent Variable Model) [10, 11] and GPDM (Gaussian Process Dynamical Model) [12, 13, 14], which transform the high dimensional observation space into a low dimensional latent space. The latent space can reveal the intrinsic structure of human motion. However, the above inconsistency problem is inevitable even in the low dimensional space.

Based on the forgoing discussions, we first have a insight into the inconsistency problem and investigate the properties of the image likelihood distribution, and then propose an annealed particle swarm optimization based particle filter (APSOPF) algorithm for 3D articulated human body tracking, in which the sampling covariance and annealing factors are incorporated into the velocity updating equation of particle swarm optimization (PSO). The most important characteristic of this strategy is that the problem of inconsistency between the image likelihood model and the true model is effectively mitigated while the searching power of PSO is still maintained. The particles in APSOPF exchange information and communicate with each other, therefore the searching is more efficient than the crossover operation in APF [6].

The rest of the paper is organized as follows. The motivation is presented in Section 2. In Section 3, our algorithm is described in detail. Experimental results are shown in Section 4, and Section 5 is devoted to conclusion.

2. Motivation

In the following parts, we first introduce a human model and an image likelihood model which are widely used in 3D human body tracking [6, 9], and show by experiment that the maximum likelihood estimate of the human motion is not compatible with the ground truth.

2.1. Human Body Model

As shown in Fig.1, the articulated human body model is represented as a 3D kinematic tree with 31 parameters (or degrees of freedom (DOF)) comprising the position and orientation of the torso and the relative joint angles between limbs.

The body model is defined by a pair $M = \{X, L\}$, where X stands for the angles between the limbs and the global location of the body, and L for the length of limbs. The limb length L is the size of the tracked person and assumed to be constant in the tracking process, while the angles represent the body pose and therefore are dynamic. Thus, the whole tracking process is to estimate the angles and this is done by maximizing the image likelihood function.

2.2. Image Likelihood Model

For each candidate pose, a corresponding image likelihood needs to be computed to evaluate how well a given body pose matches the observed images. The most commonly used features are edges and silhouettes [6, 9]. Let $w(X, Y)$ be the image likelihood function, where X is the model's configuration (angles) and Y stands for the corresponding observed image. We introduce the the edge-based likelihood $w_e(X, Y)$ and the silhouettes-based likelihood $w_s(X, Y)$ in the following subsections.

2.2.1 Edge-based Likelihood

Image edges provide a good outline for visible body parts, such as arms and legs, and are invariant to color, clothing texture and lighting condition. For edge-based likelihood, we first use gradients that have been thresholded to obtain a binary edge map. This map is then convolved with a Gaussian kernel to yield an edge distance map which indicates the proximity of a pixel to an edge.

Each part is projected onto the image plane and samples of the hypothesized edges of human body model are drawn. A sum-squared difference (SSD) function is used to measure these samples:

$$w^e(X, Y) = \frac{1}{N_e} \sum_{i=1}^{N_e} (1 - p_i^e(X, Y))^2 \qquad (1)$$

where $p_i^e(X, Y)$ is the edge distance map, and N_e is the number of projected points.

2.2.2 Silhouette-based Likelihood

The second feature is the silhouette obtained by subtracting the background from the image. Silhouette maps are generated by learning a Gaussian mixture model for each pixel over 1000 background images and comparing the background pixel probability to that of a uniform foreground model. The foreground pixel map is calculated and the values of the foreground pixels are set to 1 (and 0 for the background pixels). By taking a number of visible points on each limb and projecting them into the image, the SSD between the predicted and observed silhouette values for these points is computed:

$$w^s(X, Y) = \frac{1}{N_s} \sum_{i=1}^{N_s} (1 - p_i^s(X, Y))^2 \qquad (2)$$

where $p_i^s(X, Y)$ are the values of N_s sample points in the foreground pixel map.

In order to calculate the total image likelihood function, the features are combined in the following way:

$$w(X, Y) = exp(-\sum_{i=1}^{C} (w_i^e(X, Y) + w_i^s(X, Y)) \qquad (3)$$

where C is the number of camera views.

2.3. Inconsistency Between Observation Model and Ground Truth

As stated above, the tracking process is to find the state X maximizing the image likelihood function. It is ideal that the likelihood value of the ground truth is equal to the maximum value of the image likelihood function. However, as shown in Fig. 2, the likelihood value of the ground truth lies in the region [0.55, 0.66], which is far from the maximum value of the image likelihood function. Such phenomenon means that there exists an inconsistency between the observation model (image likelihood) and the true model (ground truth). The reason is that the image likelihood depends on the edge and silhouette information, which are inevitably contaminated by noises. This is a rather challenging problem for articulated 3D human body tracking. For example, if the search area is too small, it would be trapped into local optima and cannot acquire the true pose estimation. On the contrary, if the search step is too large, it would find many pseudo poses, wasting lots of computation. If there are too many bad particles which correspond to pseudo poses, the number of valid particles is reduced, and as a result, in many cases, leads to large tracking errors. Therefore, we need a more effective searching mechanism to make a good trade-off between global optima and local optima.

3. Our Algorithm

As we know, the annealed particle filter (APF) [6] is one of most successful algorithms in 3D articulated human body tracking. Before introducing our algorithm, we first analyze of APF, and then give a detail description of our algorithm.

By analyzing the APF algorithm, we know that on each layer m, APF samples particles according to a sampling covariance P_m. The initial value of the sampling covariance P_0 is learned from training data and then decreases annealingly as m increases. Furthermore, an annealing factor β_m is introduced to the image likelihood model so that the weights of particles also decrease step by step with layers. APF achieves a better performance than the regular PF, because the annealing factor gradually confines the search space to a relatively small region.

Recently PSO (particle swarm optimization) [15, 16, 17, 18], a new population based stochastic optimization technique, has received more and more attention because of its great searching power. Unlike other particle based stochastic optimization techniques such as genetic algorithms, the particles in PSO interact locally with one another and with their environment in analogy with the *'cognitive'* and *'social'* aspects of animal populations, found in fish schooling, birds flocking, and insects swarming. Starting from a diffuse population, now called a swarm, individuals, now

termed particles, move about the search space and eventually cluster in the regions where the optima are located.

To combine the merits of APF and the searching power of PSO, we introduce the sampling covariance item and the annealing factor item into the velocity updating equation of PSO iteration, and propose the annealed PSO based particle filter (APSOPF) algorithm, where the particle $x^{i,n+1}$ and its velocity $v^{i,n+1}$ are updated in the following way,

$$v^{i,n+1} = r_1 P_n + \beta_1 |r_2|(p^i - x^{i,n}) + \beta_2 |r_3|(g - x^{i,n}) \quad (4)$$

$$x^{i,n+1} = x^{i,n} + v^{i,n+1} \quad (5)$$

where p^i is a function of the best state found by the ith particle (called pbest), and g is the best state found so far among all particles (called gbest). r_1, r_2, r_3 are random numbers sampled from the Gaussian probability distribution $\mathcal{N}(0,1)$. Given a sampling factor $\alpha_0 < 1$, the diffusion covariance matrix p_n is evolved as follows.

$$P_n = \alpha_0 * P_{n-1} \quad (6)$$

The diagonal elements in the initialized covariance matrix P_0 are allocated to the maximum expected movement of the corresponding model configuration parameters over one time step, which is learned from training data as motion prior. β_1 and β_2 are annealing items. Given an initial β_0, they can be set as:

$$\beta_1 = \beta_2 = \beta_0 exp(1 - \frac{n}{M}) \quad (7)$$

where n is the current number of the PSO iterations and M is the maximal number of iterations.

In the canonical PSO [15], the first part of the velocity updating equation in the PSO iteration is the inertia velocity. While in our algorithm, the inertial velocity part is not used, because we find this part produces many pseudo particles or bad particles which correspond to impossible human poses. The reason is that the motion of each human body part is different and complex, and the 'history memory' (inertial velocity) of particles may be not reliable. So, we replace the inertial velocity with a sampling covariance. In this way the motion prior is introduced into the velocity updating equation. Meanwhile, this sampling covariance gradually decreases with iterations according to Eq.(6). Thus during the PSO search, the initial searching area is larger and it decreases with iterations, which confines the particles to a reasonable region.

Furthermore, the canonical PSO in [15] sets β_1, β_2 to a constant. However, this is not reasonable in practice. For example, due to the inconsistency between the likelihood model and the ground truth, the image likelihood may be unreliable, and thus the pbest and gbest may not be reliable guides. In our work, these two parameters are changed in an adaptive simulated annealing way as in Eq. (6). In this way, on the one hand the pbest and gbest still play the role of guiding the particles; on the other hand, their guiding abilities are gradually weakened with PSO iterations. So at the final PSO stage each particle still keeps their own self-searching capability. This effectively avoids completely trusting the pbest and gbest, and consequently tackles the problem of inconsistency between the likelihood model and the true model in human motion tracking.

The detail of our algorithm is presented in Algorithm 1.

Algorithm 1 Annealed PSO based Particle Filter

$[\{p_t^i, \pi_t^i\}_{i=1}^N, \overline{x}_t] = \textbf{APSOPF}[\{p_{t-1}^i\}_{i=1}^N, M, \alpha_0, \beta_0]$

Input: Individual best particles $\{p_{t-1}^i\}_{i=1}^N$ at time $t-1$, the maximal iteration number M, the sampling covariance factor α_0 and the annealing factor β_0, where $0 < \alpha_0, \beta_0 < 1$

Output: $\{p_t^i, \pi_t^i\}_{i=1}^N$ and the mean state $\overline{x}_t$ at time t

1. Randomly propagate: $x_t^{i,0} \sim \mathcal{N}(p_{t-1}^i, P_0)$

1. **for** $n = 0, 1, 2, \ldots, M$ **do**

2. Evaluate the image likelihood: $f(x_t^{i,n}) = w(x_t^{i,n}, Y)$

3. Update the individual best $\{p_t^i\}_{i=1}^N$ and the global best g of particles

$$p_t^i = \begin{cases} x_t^{i,n}, & \textbf{if } f(x_t^{i,n}) > f(p_t^i) \\ p_t^i, & \textbf{else} \end{cases}, \quad g_t = \arg\max_{p_t^i} f(p_t^i)$$

4. Carry out the PSO iteration based on Eq. (4) and (5)

5. Update P_n, β_1, β_2;

6. **end for**

$\pi_t^i = f(p_t^i) / \sum_{i=1}^N f(p_t^i)$, $\overline{x}_t = \sum_{i=1}^N \pi_t^i p_t^i$

4. Experimental results

4.1. Dataset

In this paper, we use the Brown dataset [9] to test the algorithms. The Brown data are collected for quantitatively evaluating and comparing different human tracking algorithms. The data are taken with four synchronized cameras, including training data and testing data, and associated ground truth body poses. The image size is 644×488 and the frame rate is 60fps.

4.2. Error Measure

We use the error measure proposed in [9]. First, we compute a 3D error for each particle in each frame based on 15 virtual markers that correspond to the locations of the joints and the limb endpoints. For each particle p_t^i, the full pose error $\delta(p_t^i, \gamma_n)$ is computed as the average distance of all

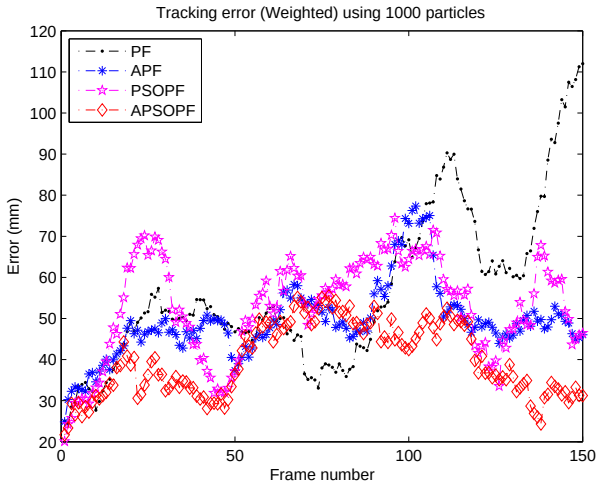

Figure 3. Tracking errors obtained by APSOPF, PSOPF, APF, PF.

virtual markers $m \in \Gamma$ with respect to the true pose γ_n:

$$\delta(p_t^i, \gamma_n) = \frac{\Sigma_{m\in\Gamma}||m(p_t^i) - m(\gamma_n)||}{|\Gamma|} \tag{8}$$

where $m(x)$ returns the 3D location of markers m for the body model x. Then we use the weighted average error measure for each frame:

$$\Delta_w = \sum_{i=1}^{N} \pi_t^i \delta(p_t^i, \gamma_n) \tag{9}$$

4.3. Tracking Results

We compare our method with PF [2], APF [6], and PSOPF [17]. For fairness, we use the same configurations: 1000 particles (for APF, it is equal to 5 layers for the annealing with 200 particles per layer. For APSOPF and PSOPF, 50 particles are used with 20 maximal iterations), a likelihood based on edges and silhouette, first-order temporal dynamics and 4 camera views.

For PF and APF, we use a hard prior that discards particles corresponding to impracticable body poses, for example, body self-intersection. But for APSOPF and PSOPF, we can not discard these bad particles, so we randomly select good particles to replace the bad ones. Moreover, we restrict the velocity of particles to be less than 1/3 of the maximum absolute inter-frame angular difference, enforcing the efficiency of the particles.

As shown in Fig.3 and Fig.4, it is obvious that the average error obtained by our method (40.0069 mm) is lower than that of APF (49.3169 mm), PSOPF (52.5961 mm) and PF (56.6869 mm). In more detail, from Fig.3, we can see that in the neighborhood of the 100th frame, APF has an error peak (up to 80 mm). This is because that around

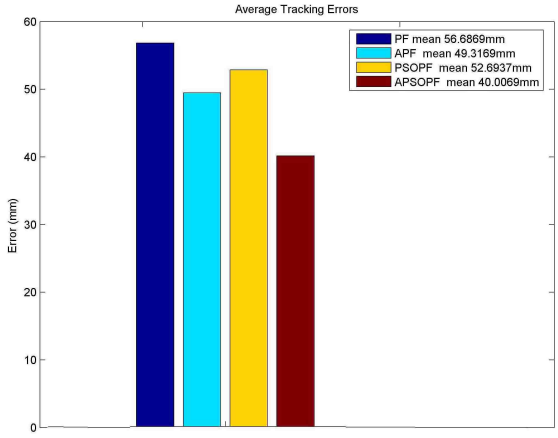

Figure 4. Average tracking error.

the 100th frame the person turns around leading to self-occlusion. The corresponding images are ambiguous and APF yields large tracking errors. However, because of the introduction of the annealing strategy into the velocity equation, our APSOPF can tackle the inconsistency between the observed model and the true model. As a result, APSOPF can still track with rather lower errors (less than 55 mm) around the 100th frame. For APSOPF, particles communicate and exchange information with each other, and this mechanism makes our algorithm more efficient than APF and PF. Meanwhile, the average number of PSO iterations for obtaining our tracking results is about 11, meaning that our algorithm can track human body with rather few particles.

Fig.5 presents the sample images around the 100th frame with the actual pose estimation for this walking sequence. The poses are projected to the second camera view. Compared with APF, APSOPF produces a better tracking result with minimal tracking errors, and it is robust to ambiguous image data such as self-occlusion (see frames 99, 103, 107 in Fig.5).

5. Conclusions

In this paper we propose a new tracking approach, APSOPF, which introduces the annealing strategy (sampling covariance and annealing factor) into the velocity updating equation of the PSO iterations. The sampling covariance makes the PSO searching more effective, and the annealing factor imposed on the individual best particles and the global best particles effectively alleviates the problem of the inconsistency between the observation model and the true model. APSOPF can tackle this problem more successfully than the competing particle filters PF, APF and PSOPF. Even in the case of image noise and body self-occlusion, APSOPF obtains better tracking results with minimal tracking errors.

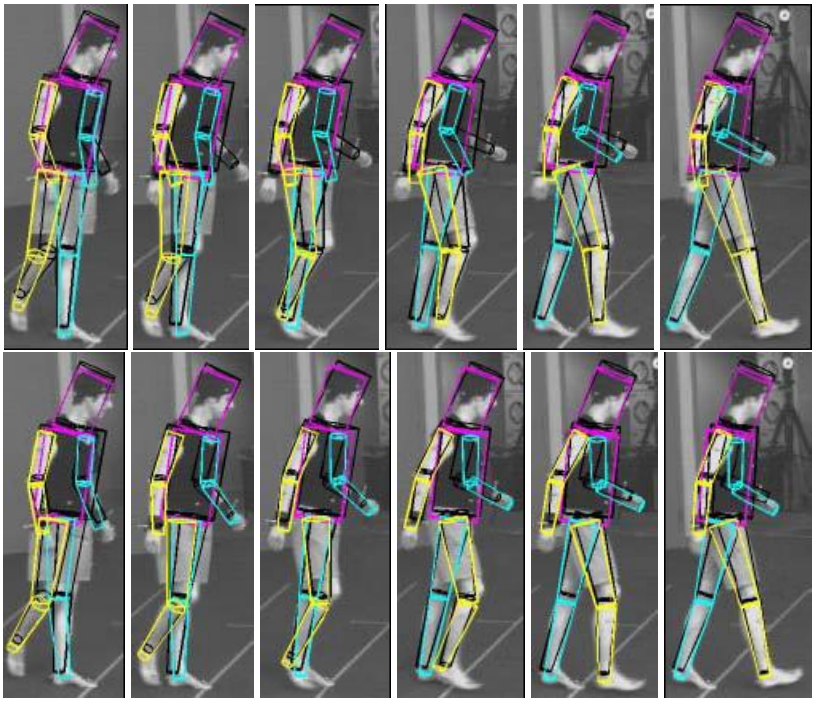

Figure 5. Tracking results for frame # 95,99,103,107,111,115 (first row: APF, second column: APSOPF). Colorized line: tracking results, black line: groundtruth.

References

[1] N. R. Howe, M. E. Leventon, and W.T. Freeman, "Bayesian Reconstruction of 3D human motion from single-camera video", *In: NIPS*, pp. 820-826, 2000. 1

[2] M. Isard and A. Blake, "Condensation: conditional density propagation for visual tracking", *IJCV.*, 29(1): 5-28, 1998. 1, 5

[3] M. W. Lee and I. Cohen, "A Model-Based Approach for Estimating Human 3D Poses in Static Images," *IEEE Trans. on PAMI.*, 28(6): 905-916, 2006. 1

[4] A. O. Balan and M. J. Black, "An Adaptive Appearance Model Approach For Model-based Articulated Object Tracking", *In: CVPR*, pp. 758-765, 2006. 1

[5] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking", *IEEE Trans. on PAMI.*, 25(10): 1296-1311, 2003. 2

[6] J. Deutscher, and I. Reid, "Articulated Body Motion Capture by Stochastic Search," *IJCV*, 61(2): 185-205, 2005. 2, 3, 5

[7] X. Xu, and B. Li, "Learning Motion Correlation for Tracking Articulated Human Body with a Rao-Blackwellised Particle Filter", In *In: ICCV*, 2007. 2

[8] J. Darby, B. Li and N. Costen, "Behaviour based particle filtering for human articulated motion tracking", *In: ICPR*, 2008. 2

[9] A. Balan, L. Sigal and M. Black, "A Quantitative Evaluation of Video-based 3D Person Tracking", *IEEE Workshop on VS-PETS*, pp. 349-356, 2005. 2, 3, 4

[10] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models", *Journal of Machine Learning Research*, vol. 6, pp. 1783-1816, 2005. 2

[11] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley, " Real-time body tracking using a gaussian process latent variable model", *In: ICCV*, 2007. 2

[12] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion", *IEEE Trans. on PAMI.*, 30(2):283C298, 2008 2

[13] R. Urtasun, D. Fleet, and P. Fua, "3D People Tracking with Gaussian Process Dynamical Models", *In: CVPR*, pp. 238-245, 2006. 2

[14] L. Raskin, M. Rudzsky, and E. Rivlin, "Tracking and Classifying of Human Motions with Gaussian Process Annealed Particle Filter", *In: ACCV*, pp. 442-451, 2007. 2

[15] J. Kennedy, and R. C. Eberhart, "Particle swarm optimization", In *Proc. IEEE Int'l Conf. on Neural Networks*, pp. 1942-1948, 1995. 3, 4

[16] M. Clerc, and J. Kennedy, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space", *IEEE Trans. on Evolutionary Computation*, 6(1): 58-73, 2002. 3

[17] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu, "Sequential Particle Swarm Optimization for Visual Tracking", *In: CVPR*, 2008. 3, 5

[18] X. Zhang, W. Hu, Wei Li, W.Qu, S.Maybank, "Multi-Object Tracking via Species Based Particle Swarm Optimization", *IEEE International Workshop on Visual Surveillance*, 2009. 3

[19] L. Sigal, "Continuous-state GraphicalModels for Object Localization, Pose Estimation and Tracking", *PhD Thesis*, Brown University, 2008. 1