A New Hierarchical Particle Filtering for Markerless Human Motion Capture

Yuanqiang Dong, Guilherme N. DeSouza
Electrical and Computer Engineering Department
University of Missouri,
Columbia, MO, USA

Abstract

Particle filtering (also known as the condensation algorithm) has been widely applied to model-based human motion capture. However, the number of particles required for the algorithm to work increases exponentially with the dimensionality of the model. In order to alleviate this computational explosion, we propose a two-level hierarchical framework. At the coarse level, the configuration space is discretized into large partitions and a suboptimal estimation is calculated. At the fine level, new particles in the vicinity of the suboptimal estimation are created using a more likely and narrow configuration space, allowing the original coarse estimate to be refined more efficiently. Our preliminary results demonstrates that this hierarchical framework achieves accurate estimation of the human posture with significantly reduction in the number of particles.

Keywords: coarse-to-fine, bottom-up aggregation of state estimations.

1 Introduction

Model-based analysis-by-synthesis is the dominating methodology in markerless human motion capture. Three general approaches exist for estimating the human posture by generating possible pose configurations from a 3D articulated human model. **Continuous** methods use optimized algorithms to search for a global [1] or local [2, 3, 4, 5] estimation of the pose configuration. **Stochastic** methods such as particle filtering [6, 7], Markov Chain Monte Carlo [8] and Belief Propagation [9, 10, 11] adopt the sampling technique to search the configuration space of human postures. **Hybrid** methods [12] combines both the optimization and stochastic sampling to search for a global estimates of the human posture. In our work, we consider a model-based 3D human motion capture using particle filters.

Particle filtering has proved to be an effective and robust technique for contour tracking[13, 14, 15]. However, the application of particle filtering to markerless human motion capture, suffers from the so called *curse of dimensionality*. That is, in particle filtering, the number of particles required to efficiently represent the state density increases exponentially with the number of degrees of freedom of the configuration space. One effective way to reduce the number of particles is based on the idea of Simulated Annealing [16, 17]. Other methods include: 1) adding constrains to the kinematics and to the human movements, as done in [18]; 2) imposing strong motion priors to constrain the search space by learning the motion model [19, 20]; and 3) hierarchically constrain possible configuration states using the observed poses of specific body parts, such as hands, face, torso, etc. [21, 22, 23].

In this work, we address the problem of dimensionality and the associated computational complexity by proposing an efficient coarse-to-fine framework using monocular image sequences. The rest of this paper

978-1-4244-2775-8/09/$25.00 ©2009 IEEE

is organized as follows: Section 2 presents our hierarchical particle filtering. Next, we provide preliminary results using monocular image sequence in Section 3. Future work and final conclusion are given in Section 4 and 5 respectively.

2 Two Level Hierarchical Particle Filter

As we mentioned above, the major drawback in particle filtering when applied to human motion capture resides in the enormous amount of particles required to achieve good results. In order to reduce the number of particles, we proposed a hierarchical framework characterized by two distinct levels: coarse and fine. At the coarse level, we discretize the configuration space using a small and finite number of partitions. That choice guarantees right from the beginning a small search space. Also, particles derived from this discrete configuration space are assigned weights based on a commonly used likelihood function. The algorithm then selects the suboptimal estimate as the one more likely to be the target configuration state. Next, at the fine level, we refine the search for particles strictly in the vicinity of the suboptimal estimate from the coarse level. In other words, at the first level of filtering, the algorithm localizes a narrow discrete configuration space for the target configuration state, while at the second level consists of a refinement of the above localization. The final configuration state is estimated by the aggregation of individual estimates for the pose of each body part in the configuration space. The summary of the proposed algorithm is listed in the pseudo code.

2.1 Coarse Level

At this level, the algorithm calculates an initial estimate for the configuration state. In order to do that, we must execute three major steps: 1) partitioning of the configuration space; 2) assigning likelihood to configuration states (particles); and 3) selecting the suboptimal estimate. In the next three subsections we will explain each of these steps.

Algorithm 1 Coarse to Fine Particle Filtering

N: number of particles at coarse level
M: number of particles at fine level
t: time stamp

Coarse Level

$$\left\{\left\lceil \frac{\theta_H^i - \theta_L^i}{\delta^i} \right\rceil\right\}_{i=1}^{h} = QuantizeStateSpace(\left\{\theta_L^i, \theta_H^i\right\}_{i=1}^{h})$$

$$\{s_j(t)\}_{j=1}^{N} = ResampleState(\{s_j(t-1), \pi_j(t-1)\}_{j=1}^{N})$$

$$\{s_j(t+1)\}_{j=1}^{N} = PredictState(\{s_j(t)\}_{j=1}^{N}, \widehat{X}(t), \widehat{X}(t-1))$$

$$\{\pi_j(t)\}_{j=1}^{N} = CalcWeight(\{s_j(t+1)\}_{j=1}^{N})$$

$$\widehat{X}_c(t+1) = SubEstState(\{\pi_j(t+1)\}_{j=1}^{N})$$

Fine Level

$$\{y_j(t+1)\}_{j=1}^{M} = GenRefinementParticle(\widehat{X}_c(t+1))$$

$$\{\lambda_j(t+1)\}_{j=1}^{M} = CalcBodyWeight(\{y_j(t+1)\}_{j=1}^{M})$$

$$\widehat{X}(t+1) = EstState(\{\lambda_j(t+1)\}_{j=1}^{M})$$

2.1.1 Quantization of the Configuration Space

Let the number of dimensions of the configuration state X be h, and δ^i denote the quantization step for each dimension $i = 1, ...h$. Then, the range $\left[\theta_L^i, \theta_H^i\right]$ of the joint angle θ^i is divided into $\left\lceil \frac{\theta_H^i - \theta_L^i}{\delta^i} \right\rceil$ equal partitions. This indicates that possible configuration states will be greatly reduced. In our work, the step size δ^i was determined using the statistics of the human motion from a training sequence – that is, the variance in angular values between consecutive frames.

At the coarse level, the target configuration state X is estimated at each time t from a set of N particles

$$\boldsymbol{S}_N(t) = \{(s_j(t), \pi_j(t))\}_{j=1}^{N} \quad (1)$$

where s_j denotes a possible configuration state and π_j the associated likelihood. The set of particles $\boldsymbol{S}_N(t)$ is chosen using proportionality of weights by binary subdivision [13]. The new set of particles,

$\boldsymbol{S}_N(t+1)$, is predicted using the following first order dynamic model [24]:

$$s_j(t+1) = s_j(t) + A\left(\widehat{X}(t) - \widehat{X}(t-1)\right) + B\omega \quad (2)$$

where $\widehat{X}(t)$ is the optimal estimation of the configuration state X at time t. Each element in the diagonal matrix A is set experimentally to a constant value of 0.2. Each element in the second diagonal matrix B is set to the quantization step δ^i. Finally, ω is a vector of independent random numbers drawn from a standard normal distribution.

2.1.2 Likelihood Function

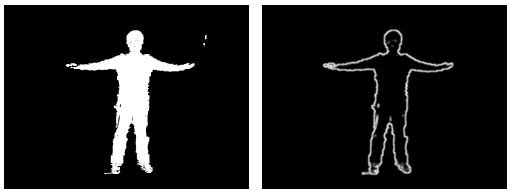
For each new particle generated by equation (2), a likelihood function is required in order to measure the similarity between the associated configuration state and the human pose as observed in the image. As described in [16], one could combine edge and region features to form such likelihood function, which starts with a silhouette extraction algorithm we developed in [25]. That is, given a silhouette image – Figure 1(a) – where 1 denotes the foreground object and 0 denotes the background, a gradient map I_g of the silhouette image is created by convolving a 5×5 Gaussian kernel with the original silhouette image [21]. Figure 1(b) depicts such gradient map, where the value of the gradient map $I_g \in [0,1]$ represents the proximity to the edges.

This likelihood function consists of two terms: one derived from edge and the other from region features. That is:

$$p_{g,r}(Z \mid X) \propto$$

$$exp\left\{-\frac{1}{2}\left(\frac{\sum_{m=1}^{M_g}(1-g_m)^2}{\sigma_g^2 M_g} + \frac{\sum_{n=1}^{N_r}(1-r_n)^2}{\sigma_r^2 N_r}\right)\right\} \quad (3)$$

where g_m denotes the value of the gradient map at the m^{th} sample point corresponding to a point on the projection the human contour, and r_n denotes the value of the foreground pixel at the n^{th} sample point corresponding to a point inside the same projected human contour. More details on the derivation of the 2D projection of the human contour can be found in [24].



(a) Silhouette Image (b) Gradient Map

Figure 1: Create gradient map from extracted silhouette image

2.1.3 Suboptimal Estimation of Configuration States

The term "suboptimal" means that the estimation of the target configuration state at the coarse level is only a rough one. However, we expect this suboptimal estimate to be sufficiently accurate to initiate the search at the fine level. Moreover, this suboptimal estimate could constrain the future search in the most likely discrete configuration partitions. In a consequence, the algorithm could avoid computations on those unlikely partitions. In our work, we assume the suboptimal estimate $\widehat{X}_c$ as the configuration state with the maximum weight. That is

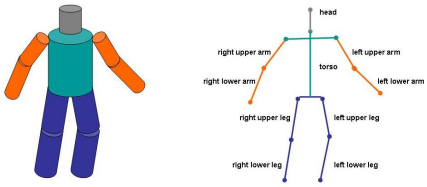
$$\widehat{X}_c(t+1) = s_j(t+1) \quad (4)$$

$$j = argmax\ \pi_j(t+1),\ j = 1, 2, ..., N$$

where $s_j(t+1)$ is the predicted configuration state and $\pi_j(t+1)$ is the associated weight using likelihood function (3)

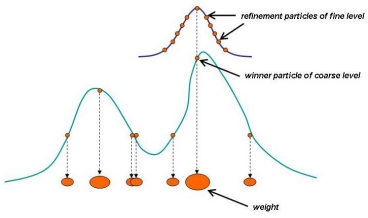
2.2 Fine Level

Given the suboptimal estimation $\widehat{X}_c$, the second level of our algorithm, the fine level, samples a second set of particles in the vicinity of $\widehat{X}_c$. Figure 3 illustrates this concept by depicting the refinement produced by the fine level on top of the coarse estimate. The



(a) Cylinder Segment (b) Kinematic Tree

Figure 2: 3D human model



(a) 3D Human Model

Figure 3: Refinement particles are generated in the vicinity of suboptimal estimate of coarse level

new set of configuration states is obtained using the following simplified model:

$$y_j(t+1) = \widehat{X}_c(t+1) + D\omega,\ j = 1, ..., M \qquad (5)$$

where ω is a vector of independent standard random variables similarly to the one mentioned for the coarse level. Also similar to the coarse level, D is a diagonal matrix with elements equal to the quantization step. As for the estimate of the configuration state, the fine level weights the new set of particles using the likelihood function (3). Finally, the optimal estimation for the target configuration state X is obtained by a bottom-up aggregation as will be described in the following section 2.3.

2.3 Bottom-up Aggregation of State Estimations

For the purpose of simplicity, we will discard the time step from equations in this section. Instead of using eq (4) to generate the final optimal estimate, we propose a bottom-up scheme to aggregate the individual estimates for the configuration state of each body part. The reason is independent search for the optimal estimates of each body part makes the computation in parallel, thus implicitly reducing the number of particles at the fine level. Similar to [26], we first separate configuration state of the whole human body into several decoupled configuration states of body parts. Then the likelihood function (3) is used to measure the similarity between the observation and the configuration state for each body part. Finally the optimal estimate will be obtained by aggregating the optimal estimates of each body part. The mathematical details will be presented in the following:

We break the h dimensional configuration state y_j of the whole human model into n decoupled body parts

$$y_j = (p_{j,1}^T, ..., p_{j,k}^T, ... p_{j,n}^T)^T \qquad (6)$$

$$p_{j,k} = (y_{j1}, ..., y_{jL(k)})^T,\ \sum_{k=1}^{n} L(k) = h$$

where $p_{j,k}$ corresponds to one body part for j^{th} configuration state. $L(k)$ denotes the associated number of joints for k^{th} body part . We also build the likelihood for j^{th} configuration state of body part

$$\lambda_j = (\lambda_{j1}, ..., \lambda_{jn})^T \qquad (7)$$

Therefore, the optimal estimation $\widehat{X}$ of the target configuration state X will be the aggregation of individual estimates from n body parts

$$\widehat{X} = (p_{j,1}^T, ... p_{o,k}^T, ... p_{m,n}^T)^T \qquad (8)$$

$$o = argmax\ \lambda_{qk},\ q = 1, ..., N$$

3 Experiments

The software was implemented in Matlab. Since the estimation of human posture from single view is restrictive to specific class of human motion, we used 330 frames of ***Combo_ 2*** sequences in **HumanEva-I** Dataset [27]. The pose estimation of selective image frames are shown in Figure 5.

# of Particles		Precision	
coarse	fine	coarse	coarse + fine
50	200	62.3%	73.8%
300	200	67.1%	74.9%
300	500	67.2%	76.3%

(a)

# of ParticlesRecall			
coarse	fine	coarse	coarse + fine
50	200	74.1%	87.6%
300	200	79.8%	89.3%
300	500	80.2%	91.3%

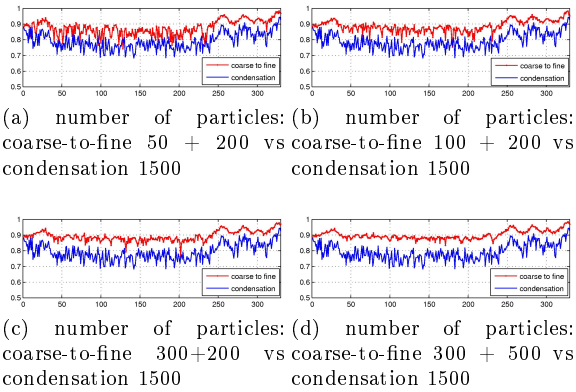
(b)

Table 1: Precison and Recall for Hierarchical PF for different numbers of particles

We use *precision* and *recall* as the measurement metric. In the first experiment, we select different number of particles at both coarse level and fine level. The averages of *precision* and *recall* over 330 frames are listed in Table 1.

In the second experiment, we compare the averages of *precision* and *recall* using proposed algorithm and Condensation algorithm. The result is shown in Table 2.

We also compare likelihood of the optimal estimates between our proposed algorithm and Condensation algorithm. We use different number of particles for our proposed algorithm to compare the likelihood over 330 frames with Condensation algorithm using 1500 particles. Figure 4 demonstrates that our proposed algorithm generates higher likelihood than the condensation algorithm even if we only use 250 particles in the proposed algorithm. We noticed that the proposed algorithm generates more smooth state estimation as the number of particles at both coarse and fine levels increases.



(a) number of particles: coarse-to-fine 50 + 200 vs condensation 1500

(b) number of particles: coarse-to-fine 100 + 200 vs condensation 1500

(c) number of particles: coarse-to-fine 300+200 vs condensation 1500

(d) number of particles: coarse-to-fine 300 + 500 vs condensation 1500

Figure 4: Likelihood for Condensation Algorithm versus Hierarchical PF

4 Future Work

The likelihood function in our work combines both the edge and region information during the estimation, but it only captures the degree of the matching between the observed configuration states and the expected ones. That is, the projection of an incorrect configuration of the human model onto the image plane for a single view may appear to fit the observation better than the projection of a more proper configuration. That is because the likelihood functions used in this work and other works only consider the overlaping pixels (matching) as a measurement of the likeness of the two images, but does not penalize the images for the non overlaping pixels (mismatching). Therefore, when using only single view, this drawback may cause the false selection of the suboptimal estimation. We will further investigate the use of new likelihood functions to capture both the degree of matching and mismatching. We also hope to present more performance evaluation of accuracy against groundtruth.

5 Conclusion

Our work resolved the computational complexity of the particle filtering with its application to human motion capture. The proposed hierarchical framework separate the state estimation into two different levels at each time step. The algorithm reduces

# of Particles		Precision	
coarse + fine	condens.	coarse + fine	condens.
50 + 200	600	73.8%	72.1%
300 + 200	1000	74.9%	73.2%
300 + 500	1500	76.3%	76.1%

(a)

# of Particles		Recall	
coarse + fine	condens.	coarse + fine	condens.
50 + 200	600	87.6%	76.3%
300 + 200	1000	89.3%	78.5%
300 + 500	1500	91.3%	79.1%

(b)

Table 2: Precison and Recall for Condensation Algorithm versus Hierarchical PF

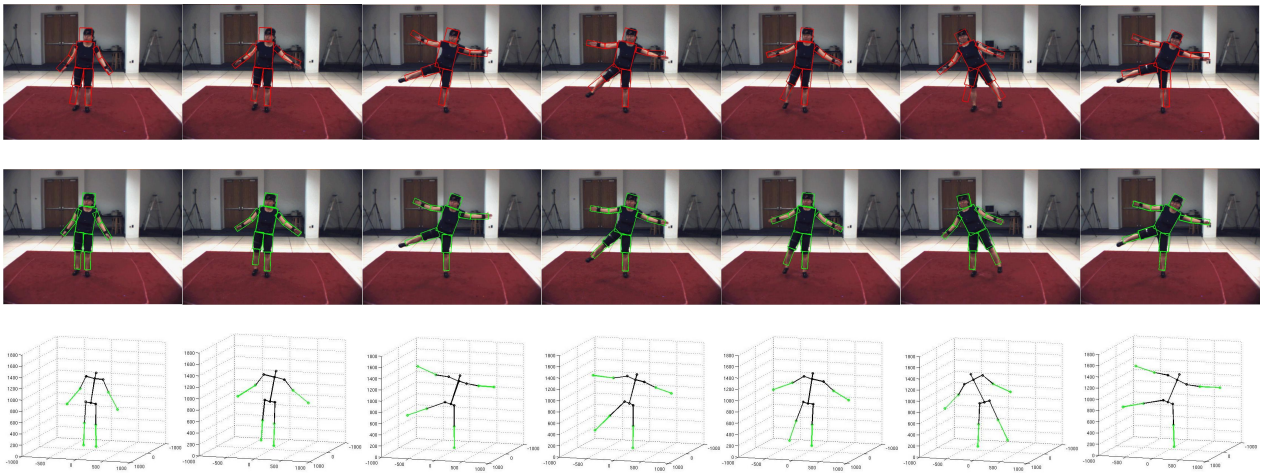


Figure 5: Pose Estimation using the Proposed Hierarchical PF for frame 2517, 2524, 2553, 2584, 2593,2640,2799 separately.The first row shows the suboptimal estimate of coarse level; The second row shows the optimal estimate estimate of fine level; The third row shows the estimated 3D Human Model.

the number of particles required to effectively represent the state density function at both levels. At the coarse level, the configuration space is quantitized to a finite number of large partitions. At the fine level, the optimal estimation of the state configuration is obtained by generating refinied particles in the vicinity of the suboptimal estimate of the coarse level. The optimal estimate is determined by a bottom-up aggregation of individual estimates of each body part. We showed that our work required less number of particles to achieve higher accuracy than the original condensation algorithm.

References

[1] J. Gall, T. Brox, B. Rosenhahn, and H. P. Seidel, "Global stochastic optimization for robust and accurate human motion capture," *Techical Report*, 2007.

[2] D. Gavrila and L. Davis, "3-d model-based tracking of humans in action: a multi-view approach," *in Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 73–80, 1996.

[3] J. Gall, B. Rosenhahn, and H. P. Seidel, "Drift-free tracking of rigid and articulated objects," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[4] L. Mundermann, S. Corazza, and T. Andriacchi, "Accurately measuring human movement using articulated icp with soft joint constraints and a repository of articulated models," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[5] R. Plankers and P. Fua, "Articulated soft objects for multiview shape and motion capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sep 2003.

[6] P. Wang and J. Rehg, "A modular approach to the analysis and evaluation of particle filters for figure tracking," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[7] H. Sidenbladh, M. Black, and D. Fleet, "Stochastic tracking of 3d human figures using 3d image motion," in *European Conference on Computer Vision*, 2000, pp. 702–718.

[8] M. W. Lee and I. Cohen, "Proposal maps driven mcmc for estimating human body pose in static images," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[9] R. Wang, W. K. Leow, and H. W. Leong, "3d-2d spatiotemporal registration for sports motion analysis," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[10] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[11] M. Lee and R. Nevatia, "Human pose tracking using multi-level structured models," in *European Conference on Computer Vision*, 2006, pp. 368–381.

[12] J. Gall, B. Rosenhahn, T. Brox, and H. P. Seidel, "Optimization and filtering for human motion capture," *International Journal of Computer Vision*, 2008.

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

[14] A. Blake and M. Isard, "Active contours," 1998.

[15] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," *in Proceedings of European Conference Computer Vision (ECCV)*, pp. 3–19, 2000.

[16] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2000, pp. 126–133.

[17] J. Deutscher, A. Davison, and I. Reid, "Automatic partitioning of high dimensional search spaces associated with articulated body motion capture," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2001, pp. 669–676, kauai, USA.

[18] M. A. Brubaker and D. Fleet, "The kneed walker for human pose tracking," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[19] H. Sidenbladh and M. Black, "Learning image statistics of people in images and video," *International Journal of Computer Vision*, March 2003.

[20] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H. P. Seidel, "Markerless motion capture of man-machine interaction," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[21] P. Azad, A. Ude, T. Asfour, G. Cheng, and R. Dillmann, "Image-based markerless 3d human motion capture using multiple cues," in *International Workshop on Vision Based Human-Robot Interaction*, 2006, italy.

[22] P. Azad, A. Ude, T. Asfour, and R. Dillmann, "Stereo-based markerless human motion capture for humanoid robot systems," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2007, roma, Italy.

[23] M. W. Lee, I. Cohen, and S. K. Jung, "Particle filtering with analytical inference for human body tracking," in *IEEE Workshop on Motion and Video Computing*, 2002.

[24] P. Azad, A. Ude, R. Dillmann, and G. Cheng, "A full body human motion capture system using particle filtering and on-the-fly edge detection," in *Proceedings of IEEE International Conference on Humanoid Robots*, 2004, santa Monica, USA.

[25] Y. Dong, T. X. Han, and G. N. DeSouza, "Illumination invariant foreground detection using multi-subspace learning," *submitted to Computer Vision and Image Understanding*, 2008.

[26] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, February 2005.

[27] L. Sigal and M. J. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," *Technical Report CS-06-08*, 2006.