3D POSE ESTIMATION IN HIGH DIMENSIONAL SEARCH SPACES WITH LOCAL
MEMORIZATION

Weilan LUO  Toshihiko YAMASAKI  Kiyoharu AIZAWA

Dept. of Information and Communication Engineering
The University of Tokyo

ABSTRACT
In this paper, a stochastic approach for extracting the articulated 3D human postures by synchronized multiple cameras in the high-dimensional configuration spaces is presented. Annealed Particle Filtering (APF) [1] seeks for the globally optimal solution of the likelihood. We improve and extend the APF with local memorization to estimate the suited kinematic postures for a volume sequence directly instead of projecting a rough simplified body model to 2D images. Our method guides the particles to the global optimization on the basis of local constraints. A segmentation algorithm is performed on the volumetric models and the process is repeated. We assign the articulated models 42 degrees of freedom. The matching error is about 6% on average while tracking the posture between two neighboring frames.

Index Terms— Visual hull, twist, annealing simulation, particle filter, tracking

1. INTRODUCTION
Kinematic body motion capture or 3D spatio-temporal surfaces reconstruction from synchronous multi-camera or multi-view video sequences is still a challenging and fundamental problem for many applications, including 3D animation movies and games, medical diagnostics motion analysis, or robot motion simulation. Marker-based motion capture system is capable to provide motion capture data of high accuracy quickly. However, it requires people to wear skin-tight clothing with markers and special capture hardware is needed. It is hard to capture motion robustly while human wears loose apparels. In the past years marker-less motion capture has received more and more attention and been used in commercial fields varying from surveillance to character animation and 3D movie display.

APF is shown to be effective in visual motion tracking of articulated body simplified by cones with elliptical cross-section with 29 degrees of freedom [1]. However, the poses of the hands and feet are ignored and the oversimplified models are too crude for recovering complicated shape and motion precisely. Therefore, we employ the volumetric models generated from multi-view images directly for 3D pose estimation. We select 5% of volume data for motion tracking to decrease the computational cost as the number of voxels in each model is beyond 100 thousands. We develop the stochastic search approach with the memorization of local optimization in order to avoid the local misaligned problem. Furthermore, our proposed modified APF is also able to track the motion clip of quick movement or human with general apparel in high dimensional spaces.

We assume the volumetric model of the first frame as the template body and segment it into 15 body segments. The corresponding skeleton model is also given. In our work, we first reconstruct the mesh surface by marching cubes [2] and extract the articulated skeleton model from it. Then we segment the mesh surface into 15 parts based on the skeleton and geodesic distances. The labeled model with an underlying skeleton is employed for motion tracking as a template.

2. RELATED WORK
Marker-less human motion tracking has been a challenging problem in fields of computer graphics and computer vision for years. It is intuitive to represent kinematic postures by articulated skeleton models. Therefore, several simply general geometric representation methods are used to replace the body segments of a human.

Deutscher et al. [1] reconstructed the subject body shape by cones with elliptical cross-sections with 29 degrees of freedom. They modified conventional particle filters by layering the search space based on annealing to estimate the articulated body motion in the high dimensional spaces. Plankers et al. [3] developed a method to reconstruct articulated objects by metaballs and estimate human motion using the Levenberg-Marquart least squares estimator. However, in general such models are too simple to recover shape and motion accurately.
In contrast to these methods, Corazza et al. [4] proposed an automatic generation method of a subject-specific model with joint center locations instead of using simplified models. A marker-less motion tracking taking advantage of visual hulls, subject specific model and using different number of cameras and dataset was present in [5]. However, the articulated ICP method does not work well for fast motion clips and the tracking results rely on the qualities of the visual hull.

The kinematic chains were extracted directly from time-varying meshes which had the time-varying topologies (vertices, faces and colors) in [6, 7]. The motion extraction method was based on Reeb graph, and a geodesic function utilizing principal component analysis was given for pose estimation in [6]. Lee et al. [7] utilized the extracted skeleton chains, segmented the time-varying mesh surfaces based on skeleton modes by distance calculations, and refined the 3D pose using the decomposition results.

We present a model-based tracking algorithm that follows the idea of crosswise combining the particles with local maximum and annealing to generate the particles to the globally optimal solution. Stochastic approaches are widely used in the field for motion tracking with global optimization. 3D representation by volume data can be generated easily from multi-view images. So we estimate the articulated postures using visual hulls instead of silhouettes and rim features. Sample data from the template are chosen and deformed stochastically to match the volumetric model to be tracked. The weights for all body segments and the whole body are calculated and sorted. Half of the particles are reconstructed by combining the 15 selected particles according to sorted indexes for all body segments. New particles for the next layer are generated using the global weights. The process is repeated until layer 1. The solution converges to the global maximum while local optimizations are also enforced.

3. MODEL-BASED POSE ESTIMATION

In this section, a developed articulated APF approach is used for rigid transformation in high dimensional configuration space. Assume that we have a segmented volumetric model and try to estimate the 3D posture for the next frame. As the number of the voxels is about 100 thousands while the size of the voxel is set to be 10mm, we select five percent of the data randomly from the previous model for calculating the motion quickly. In addition, we let the number of the sample data of each body segment be same to ensure the same significance of each limbs while tracking.

Then the current model is deformed to match the visual hull in the next frame with 42 degrees of freedom. Degrees of freedom of the global translation and rotation are treated as six. Wrist, knee and ankle joints are defined with two degrees of freedom. Shoulder, hip, neck and upper body joints are given three degrees of freedom. A deformed model (particle) can be represented by \( \chi = (x_1, x_2, \ldots, x_{42}) \). Twists representation and exponential coordinates are employed for rigid motion estimation. The right posture is extracted while the deformed sample model is fitted to the target. A difference function \( D_f(X, Z) \) between two volumetric models is given to measure the result of human motion tracking by comparing the voxels in all body segments.

\[
D_f(X, Z) = \frac{1}{15} \sum_{i=1}^{15} \left( 1 - \frac{p_i(X, Z)}{N_i} \right)
\]  

(1)

where \( X \) is the segmented volumetric model and \( N_i \) is the number of volume data labeled as \( i \) and \( Z \) is the model to be matched. \( p_i(X, Z) \) is the number of the volume data in the body segment \( i \) which also belongs to the model \( X \). In addition, we measure the “fitness” between two models to generate new particles by the weighting function as shown in the following

\[
\omega(X, Z) = \exp \left( -\frac{1}{15} \sum_{i=1}^{15} \frac{p_i(X, Z)}{N_i} \right)
\]  

(2)

Our method tends to seek for the optimal solution by annealing to produce new particles according to the weights. Unfortunately, the extracted 3D posture with global optimization for the human body may not fit to each body segment so that it is difficult to avoid the local misaligned problem. Therefore, the local “fitness” of each body segment is also taken into account in our algorithm. Deutscher [1] developed the APF method by using a crossover operator imitating Genetic Algorithm (GA) and hierarchical partitioning. Our proposed method also simulates the breeding of individuals as the conventional genetic algorithms by taking local errors minimization into consideration. The global tracking algorithm with local memorization while breeding new particles is conducted according to the following steps:

1) Start an annealing run at the layer \( M \).

2) Deform the sample to construct \( N \) un-weighted particles.

3) Assign a normalized global weight to each particle as described in [1]. The local weights for all body segments are also calculated corresponding.

![Fig. 1. Two volumetric models. (a) The model and the corresponding bounding box. (b) The labeled model.](image)
4) Sort the particles by local weights and combine them. Assume 15 particles \( s_i = (x_{i1}, x_{i2}, \ldots, x_{i42}), \) \( i = 1, \ldots, 15 \) fitting well to the body segment \( i \) are selected. We construct the new particles \( s^* \) from \( s_i \) by choosing the values affecting the transformation of the segment \( i \) and set others 0. We set \( n_k = 0, k = 1, 2, \ldots, 42 \). If the \( k \)th data in \( s_i^* \) is nonzero, \( n_k = n_{k+1}. \) Define \( s^* = \sum s_i^* = (x^*_1, x^*_2, \ldots, x^*_{42}) \). Then the 15 selected particles are combined to form the new particle \( s^{\text{new}} \) where

\[
s^{\text{new}} = (x^*_1/n_1, x^*_2/n_2, \ldots, x^*_{42}/n_{42}) \tag{3}
\]

5) Recalculate the normalized global weight for each particle and draw \( N \) new particles randomly according to the weights.

6) The selected particles are used to initialize layer \( m-1 \). The process is repeated until we arrive at the layer 1.

7) The optimal solution is estimated by combining the particles of layer 1 according to the normalized weights.

In our experiments, it was found that setting the layer number \( M = 10 \) with particle number \( N = 300 \) worked well for human motion tracking. The previous frame of the model needs to be separated into 15 parts while estimating the next posture of human body as shown in Figs. 1, 2. The model segmentation method will be introduced in section 4.

4. SEGMENTATION

It is intuitive that the registration algorithm will be time-consuming and hard to achieve accurate results if we register part of the human body such as the right hand to the whole model. So we intend to divide the model into several parts that the registration algorithm will be applied to match part to part for each body segment.

As our modified APF approach is able to estimate the articulated pose with global optimization, the observed volumetric model is partitioned into several limbs in accordance with the minimum Euclidean distances to the deformed template visual hull. A rapid and efficient minimum distance calculation method is provided. The bounding box of the segmented model is given. Then this volumetric model is able to be represented by the boolean values 0 or 1. We detect the neighboring voxels of each data in the visual hull to be segmented in turn.

5. EXPERIMENTAL RESULTS

We use the public datasets provided by Gall [8] to test our algorithm at first. The purpose of our approach is to extract 3D articulated kinematic chains directly from a time-varying volume sequence. In our program, the property of binary representation of volumetric model makes it easy for comparing models or distance calculation.

As seen in Fig. 2, the extracted articulated skeleton by our proposed method fits well to the volumetric model as shown in the top left plot (a) than the APF method. The indexes in Fig. 2(c) represent the body segments. The last column in Fig. 3 depicts the global error using APF and our method. The deformed model by APF did not track body segments well such as the head, the right lower arm, the left foot and the right hand. The correspondent indexes shown in Fig. 3 are 1, 2, 4 and 6. For instance, our method decreased the mismatching rate of the left foot from 35% to 7% as seen in Fig. 3. Also it is obvious as shown in Fig. 2 that our method has located the right foot more accurately than the APF method.

In Fig. 4, the dotted line means the difference between the visual hull to be tracked and the previous one by comparing the volume data. The mean mismatching rate is about 6 percent after using our modified APF method. The difference between the deformed template and the current model (the black dotted curve) increases with \( t \). One of the reasons is that only

Fig. 2. Tracking results by APF and our method. (a) The model to be tracked. (b) (c) The deformed models by APF and our method respectively. (d) (e) Motion capture capture data.
rigid transformations are taken into consideration while non-rigid deformations are also very important for motion tracking. The errors will be accumulated if we just deform the template model according to the rigid transformations. The other one is that errors also exist in models segmentation. The results of model segmentation rely on the accuracy of motion capture data and the volumetric models.

The noisy data existing in the visual hulls cause errors for motion tracking. Then the mislabeled volume data as shown in Fig. 5 will cause problems while tracking. For \( t = 23 \), the position of the upper left leg is moved toward the right part, so some volume data in the right leg are labeled as the left part. It is necessary to do some refinement to erase the noisy data.

6. CONCLUSION

We proposed a model-based markerless motion tracking method capable of extracting robust 3D articulated postures with 42 degrees of freedom through a sequence of visual hulls. Although the accuracy relies on the quality of visual hulls, our tracking method shows efficiency and robustness for human tracking in high dimensional configuration spaces comparing with other methods such as the APF algorithm. A sequence of subject-specific models is necessary for articulated ICP algorithm while we utilize the motion tracking data to segment the volumetric models. In addition, time-varying mesh surfaces can be extracted easily. The proposed refinement method in this paper improves the accuracy of the pose estimation only slightly. We prefer to deform a template mesh surface by the motion tracking data and project all vertices to the images to avoid the noise exist in visual hulls.

7. ACKNOWLEDGMENT

This work is partially supported by National Institute of Information and Communication Technology, Japan.

8. REFERENCES