A New Hierarchical Method for Markerless Human Pose Estimation

Yuan Lei, Huawei Pan*, Weixia Chen, and Chunming Gao

School of Information Science and Engineering, Hunan University, Changsha,410082, Hunan, China
38697014@qq.com

Abstract. We present a system for markerless human motion capture through a hierarchical method from multiple camera views. In the absence of markers, the task of recovering the human pose is challenging and requires strong image features and robust algorithm. We propose a solution which integrates the 2D posture information and the volumetric reconstruction. Firstly, the model's initia posture is obtained through the method of segmenting silhouette. After that, we track the human pose by using a hierarchical method, which is divided into three steps: head detection, torso prediction and limb matching. In order to gain the robust results, we discard the interior voxel data, use the middle voxel data for motion tracking, and use the surface voxel data for global optimization. The experiment results show that the method is valid and robust.

Keywords: hierarchical, markerless, tracking, Iterative Closest Point.

1 Introduction

Markerless human motion capture is an active topic in the areas of computer vision with many applications in animation, interactive games, motion analysis (sport, medical) and surveillance. One of the key technologies of motion capture is the pose estimation, which recovers 3D human body pose parameters from 2D videos. There has been significant work in recovering the full body pose from images and videos in the last ten-fifteen years. Most human pose estimation methods are divided into two categories: monocular [1, 2]and multi-view [3, 4]. Although the technology of 2D skeleton extraction is well enough to use for pose estimation, self-occlusion problem is still unresolved. Therefore, most researchers adopt based on multi-view method to estimate 3D human pose from voxels. The human pose estimation methods are divided into two: model-free[5]and model-based[6]. According to the experience, most of human pose estimation systems usually adopt the model-based method. Analysis method using synthesis technique is usually applied in the model-based method. Its basic principle is predicting the posture of the human model, fitting it to the feature extracted from images and updating the parameters of the human model.

* Corresponding author.

M. Chen, B. Leibe, and B. Neumann (Eds.): ICVS 2013, LNCS 7963, pp. 163–172, 2013.

In general, due to the inability of recovering the optimal parameters by global searching, we recover the human posture by searching around the initial values. And then, we can obtain the posture from each frame by the "predict-fit-update" processing. Thus, the initialization of the human body model (HBM) is an essential step in the human motion tracking process. So, the automatic initialization in the markerless motion capture system has become a problem that should be solved urgently. To solve this issue, this paper puts forward a method that can automatically acquire human posture based on the silhouette segmentation.

The problem of tracking the pose of human body is to estimate the position and configuration of the HBM from the video data and take them as the parameters of the tracked human body. Due to high dimensionality of the pose space, it is challenging to search the true body configurations for any search strategy. The existing literature for model-based tracking either approach the problem using a Bayesian filtering formulation[7–9], or as an pure optimization problem [10–12]. L. Sigal et al.[9] formulated the human pose and motion estimation by solving using a non-parametric belief propagation, which uses a variation of particle filtering that can be applied over a general graphical model with loops. Zheng Zhang et al.[10] estimated human posture by matching the voxels to the barrel model updated by Particle Swarm Optimization (PSO) algorithm.

2 System Flow

The flow chart of this motion capture system is shown in Fig.1. It mainly consists of initialization and tracking parts. The initialization part includes the system initialization, such as camera calibration using Zhang's method[13], and the model initialization, which is expounded in 3. The initial parameters of the human model are retrieved from labelled voxels, which is obtained by integrating the features extracted from a segment silhouette into reconstructed 3D volume data. The motion tracking can be complete through three steps. Firstly, the head information is searched by iteration of encircling, which is detailed in Sec.4.1. Then, the main vector of torso is predicted in Sec.4.2. At last, we introduce a method of tracking the limbs by matching them to the labelled voxels using ICP(Iterative Closest Point) algorithm described in Sec.4.3.

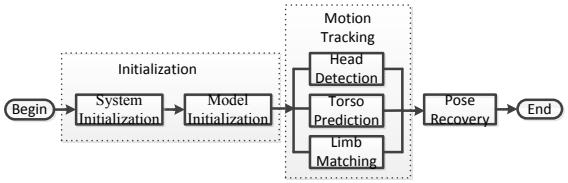

Fig. 1. The system flow chart

3 Model Initialization

For a model-based markerless motion capture method, an articulated body model is needed. The Cylinder model are flesh out by 10 cylinder, as shown in Fig.2,

and contains 15 joints: head, neck, root, shoulders, elbows, wrists, hips, knees, and ankles. P_{joint} denotes the joint coordinate in the world coordinate system.

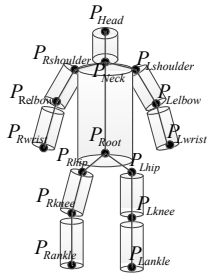

Fig. 2. 3D cylinder Model

According to the issue of pose initialization, we propose a method for model initialization, which consists of three steps: segmenting silhouette into different parts, labelling 3D voxels, and then extracting 3D human parameters, as shown in Fig.3. The core principle of our algorithm is recovering human 3D posture by combining the 3D information with the 2D features, which is extracted from one appropriate silhouette.

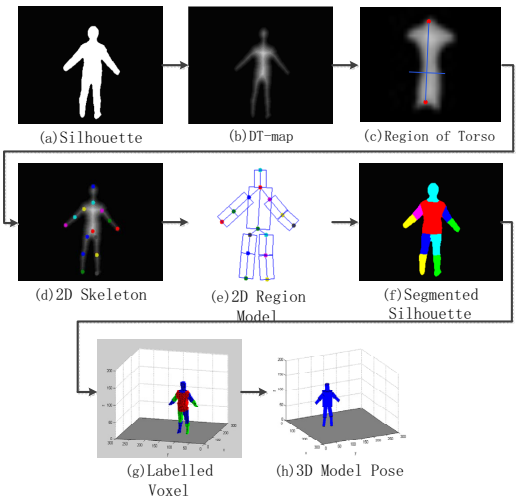

Fig. 3. Flow chart of pose initialization

3.1 Silhouette Segment

In order to obtain human pose features from image, we segment the silhouette to different parts by matching the pixels of the silhouette to a region model. The region model is built with a 2D skeleton which is extracted from 2D silhouette based on Distance Transformation and PCA(Principal Component Analysis) algorithm.

Step1.2D Skeleton Extraction. We present a robust method to extract 2D skeleton parameters from a silhouette, as shown in Fig.3(a). Firstly, we process the silhouette with the Fast Distance Transformation(Fast DT). The definition of the Fast DT can be described as a map whose value in each pixel p of the predicted contour, called Region of Interesting, satisfies the following condition:

$$D(p) = min\{d(p,q)|I(q) = 0\}. \tag{1}$$

Where I means a binary image, p is the foreground pixel, q is the background pixel. After then, we can obtain the DT map of the silhouette, as shown in Fig.3(b). Because of the torso is much broader than other parts of human bodies such as limbs and head, the DT value is higher. In our experiments, we also find that the DT value of the torso is very static, which attributes to the similar of the size between human beings mainly. Thus, the simple constant threshold can be used to separate the torso from the silhouette, as shown in Fig.3(c).

After that, PCA will be introduced to compute the principal normal vector of the torso and its orthogonal vector, represented by v_1 and v_2. We denote the point which has the maximum value in the DT map as P_{dt_max}. Therefore, with the length of the human vertebra, we can gain the general location of P_{root2d} by $P_{root2d} = P_{dt_max} + L_{torso} * v_1$, where the L_{torso} is the estimated length, P_{root2d} is estimated point. In order to gain the optimal location of P_{root2d}, we introduce a method based on the constraints:$\arg\max \sum_{i=1}^{n}\{i|i \in Rect(P_1, P_2, \delta)\}$,where $Rect(P_1, P_2, \delta)$ is the rectangle region built with the P_1 and P_2, its width is 2δ, as shown in Fig.4(a). Fig.4(b) shows the situation of satisfying constraints, and Fig.4(c) is the situation of non-satisfying constraints.

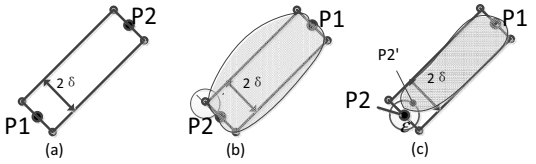

Fig. 4. Estimate articulation using energy constraint

Finally, we can extract the 2D joints from the silhouette, as shown in Fig.3(d), by combining the remaining pixels projection to the orientation of vector v_1 and v_2, with the human prior information.

Step2.Pixel Classification. We establish an adaptive human region model according to the 2D joint points. The human region model consists of 10 regions $I_n, (n = 1, \cdots, 10)$, as shown in Fig.3(e). Then, we match each pixel of the silhouette to the region model by calculating the minimum distance from the pixel to every part of the region model. The pixels can be labeled to the different values by the following function:

$$SL(p) = n, if D(p, I_n) = min\{D(p, I_1), \cdots, D(p, I_{10})\}. \tag{2}$$

where $n = 1, 2, \cdots, 10$,each region I_n is formed by four vertex t_1, t_2, t_3, t_4 , $D(p, I_n)$ is the minimum distance from the pixel p to the region, and can be computed by the algorithm of distance from point to polygon, as shown in the following equation.

$$D(p, I_n) = min\{d(p, L_k)\}. \tag{3}$$

Finally, each pixel is classified to 10 groups by the above method, and is labelled by different values, as shown in Fig.3(f), where different color regions represent different body parts.

3.2 Labelled Voxel Reconstruction

We gain 3D voxel by SFS(Shape-From-Silhouette) algorithm [14]based on a look-up table. In our paper, we establish a look-up table between voxels and their correspond projection pixels in each camera view. In general, the labelled silhouette is recorded as SL(Silhouette Labelled). According to the previous results, our look-up table of the SL is different from the traditional table. The pixel has different value when it is on different body part regions. If the voxel's projection in the SLwhich is belong to the head region, the voxel is marked as head. After that we can obtain human body part segmented and labelled voxels, as shown in Fig.3(g).

3.3 Initial Pose Extraction

According to the above results, 3D voxel data can be segmented into different parts, such as head, torso and limbs. Based on the topology of human, we figure out all junction voxels between two body part data that has the different $f(x, y, z)$. After that, we can get the rest point, such as the end of limbs and head, according to the information of the rest voxel data and the human skeleton prior length.

We can obtain the human 3D skeleton information by the previous method. Then, the skeleton parameter is modified by the body rigid constraint as the recovered human model parameters. Then, we can get the initial posture of the model, as shown in Fig.3(h).

4 Tracking

Motion tracking can be defined as pose estimation from image sequences with the temporal and volumic information . We introduce a hierarchical method for tracking human poses based on the human body model. Due to its unique shape and size, the head is easiest to find and is located firstly. After then, the torso vector can be predicted through the iteration of searching. Finally, we label the voxels of the frame t based on the Mahalanobis distance between the voxels and the predicted positions of the model parts. For each part of the model, we gain the rotate matrix and translation vector by using ICP algorithm. Finally the model parameters is recovered using the global optimization of the objective function.

According to the connectivity, we divided the voxels to three : surface, middle, interior. In order to reduce the computation, we subtract the interior voxels, and use the middle voxels for motion tracking, the surface voxels for global optimization.

4.1 Head Detection

We create a spherical crust template whose center is P_{head}^{t-1} and radius is r_{head}. P_{head}^{t-1} is the head position of frame t-1. The template initial center location of frame t that maximizes the number of the voxels, which are inside the crust, is represented by C_0^t. The sphere fitting algorithm is then applied by updating the center of the template (as shown in Fig.5). The updated center N_i^t is given as:

$$N_i^t = P_{head}^{t-1} + i(C_0^t - P_{head}^{t-1})/|C_0^t - P_{head}^{t-1}|. \tag{4}$$

where i is the step length. Then we recount the center C_i^t of the voxels which fall into this new sphere.

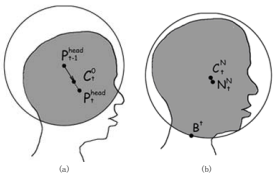

Fig. 5. Shpere fitting procedure Left: iteration begins; Right: iteration stops

The algorithm iterates until $\|C_n^t - N_n^t\| < Thresh$. The head position of the frame t is obtained by $P_{head}^t = N_n^t$. We obtained the junction voxels between body and head, defined by V_{neck}. Therefore, the $B^t = \frac{1}{N}\sum_{i=1}^{N} V_{neck}(i)$ is chosen as the neck point of frame t.

4.2 Torso Vector Prediction

Between the plane which is perpendicular to torso and passes neck point P_{neck}^{t-1} and the plane passes pelvis point P_{root}^{t-1} at t-1 time, we can build a spatial direction set as shown in Fig.6. For each point which distributes uniform in the neighborhood area with radiusδ in the neck and pelvis planes can train $(N+1)$ strips vectors. N is the number of sampling points in the neighborhood area of neck or root.

Each trained vector is used as a predictive torso direction at t time. For each vector, count up the t time voxel data number which falls into the section $[P_{neck}^t, P_{root}^{'t}(n)]$,$n = 1, \cdots, (N+1)$ and vertical distance is less than torso radius. The vector corresponds to the maximum number is as the optimal solution,which is direction of torso at the t time. And that, we can obtain P_{root}^t by the optimal torso direction.

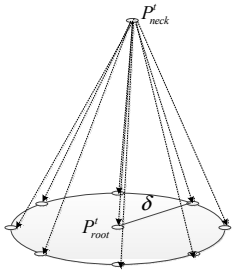

Fig. 6. Prediction vector set

4.3 Limb Matching

We first divided the voxels into two types: the surface and interior. The interior voxels are used for motion tracking, and surface voxels are used for global optimization. Acorrding to the previous sections, we got three points of skeleton, such as head, neck, and root. $X^t = \{x_1^t, x_2^t, \cdots, x_{12}^t\}$ notates the rest joint points. $V^t = \{v_1^t, v_2^t, \cdots, v_{12}^t\}$ and $M^t = \{m_1^t, m_2^t, \cdots, m_{12}^t\}$ denotes the voxel data and model sampled points of each body part respectively.

Step1.Update X^t and M^t. According to the previous section, the body parameter is assumed by $x_i^t = x_i^{t-1}, i = 1, \cdots, 12$. M^t is the points sampled from the HBM, which is established via the skeleton parameters X^t .

Step2.Label voxels. The voxels of frame t are labelled to different body parts, by matching it to the model M^t , according to the Mahalanobis distance, is as follows.

$$d(v, m_i^t) = (v - u_i)k_i(v - u_i)^T. \tag{5}$$

where v is the voxel, u_i is the mean value, k_i is the covariance matrix. The voxel is belonged to the part which has minimum distance value.

Step3.Bottom-up fitting. We employ a hierarchical method for fitting the voxel to the model points, using the ICP algorithm for the local optimization. As the HBM is a topological structure, we can figure out the skeleton point by the bottom-up method. The constraints is applied to the local optimization using ICP algorithm. As the right forearm for example(as shown in Fig.7), it includes two endpoints, right elbow and right wrist, which are represented as x_4^t, x_5^t at frame t. In our method, the x_4^t is computed first. So, the R and T is chosen as the local optima when it is satisfied the condition: $\arg\min \|(Rx_4^{t-1} + T) - x_4^t\|$. Then, the endpoint of wrist can be acquired by $x_5^t = Rx_5^{t-1} + T$. Finally, we can obtain the optical parameters of X^t gradually.

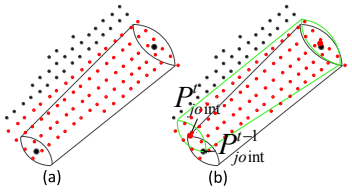

Fig. 7. Body Tracking

Step4.Global optimization. The fittness function is

$$f(R,T) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_s} \|(RP_i + T) - P_j^{'}\|^2. \tag{6}$$

where P is the point of the updated model, $P^{'}$ is the point of surface voxels. We use the Singular value decomposition (SVD) algorithm to calculate the optimal value of R, T. Finally, the human pose of each endpoint can be recovered by $P^t = RP^{t-1} + T$.

5 Experimental Results

We evaluate the performance of our hierarchical human pose estimation approach using the HumanEva-II dataset [15]. The dataset consists of synchronized video streams from 4 color cameras at 60 Hz along with ground truth 3D body poses obtained using a commercial motion capture system. We assess the accuracy of recovered poses using the evaluation metric proposed in Sigal et al. which measures the sum of the Euclidean distances to $K = 15$ virtual markers corresponding to the locations of the major joints. The error in the voerall estimated pose $X_{mrk}^{'}$to the ground truth poseX_{mrk} (in mm) is expressed as the average absolute distance between individual markers, $X_{mrk} = \{p_1, p_2, \cdots, p_K\}$.

$$Error(X_{mrk}, X_{mrk}^{'}) = \frac{1}{K} \sum_{k=1}^{K} \|p_k - p_k^{'}\| \tag{7}$$

When computing the result, we ignored 40 frames(298-337) for Subject 4 where the ground truth is not available. Fig.8 and Fig.9 gives the sample frames of the experimental results. Fig.10 shows the plots of the mean errors over the sequences. We conducted a quantitative comparison of our method against the method 1: L.Sigal et al.[15]provided a baseline algorithm that used a relatively standard Bayesian framework with optimization in the form of annealed particle filtering. Method 2: Poppe et al.[16]took an example-based approach to pose recovery. This approach is somewhat person specific and does not generalize well to unseen actions. The errors of the three algorithms performed on the HumanEva-II dataset. Compared with the results from Methods 1 and 2, our method provides more accurate results. This illustrates that our method can improve the performance of other human pose estimation methods. The Table.1 shows the time cost of the our method and the Method 1. The disadvantage of our method is large amount of calculation, due to the ICP algorithm we used.

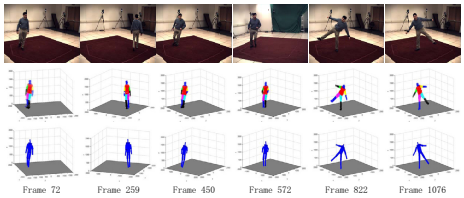

Fig. 8. The experimental results of human pose estimation on HumanEvaII with S2 subject

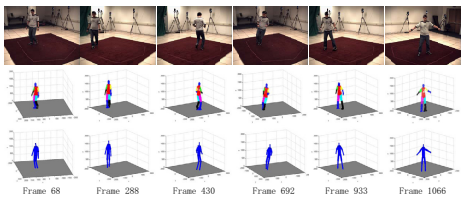

Fig. 9. The experimental results of human pose estimation on HumanEvaII with S4 subject

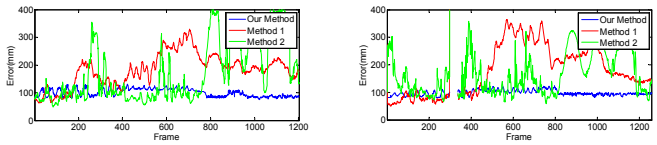

Fig. 10. Error analysis chart on S2 and S4 subject

Table 1. Running Time Comparison

Method	My method	Method 1
Time(second)	45.23	18.74

6 Conclusion

This paper proposes human motion initialization and motion tracking approach based on labelled voxels. Fisrtly we get the initial skeleton information by segmented silhouette, and then label the next frame body voxel data by tracking algorithm which includes head template fitting algorithm and body part tracking. Experimental results verify that the subsequent motion voxel data can be tracked easily and robustly precisely. Because the ICP algorithm is time costly, the deficiency of the method is that we cannot accelerate the running time, on the premise of the accuracy. And lacking of hardware equipment, we can not tackle the tracking procedure really in real time. We will focus the future work on parallel computing with our method, and make markless motion capture effectively in real time.

Acknowledgement. Project is supported by National Natural Science Foundations of China (No. 11201136). Project is supported by Guangdong Province's

Ministry of Education projects for Industry-Academy-Research cooperation (No 2011B090400002).

References

1. Mykhaylo, A., Stefan, R., Bernt, S.: Monocular 3D pose estimation and tracking by detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 623–630 (2010)
2. Ankur, A., Bill, T.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 44–58 (2006)
3. Atul, K., Niels, H., Graham, T., et al.: 3D human pose and shape estimation from multi-view imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 49–56 (2011)
4. Hofmann, M., Gavrila, D.M.: Multi-view 3D human pose estim-ation in complex environment. International Journal of Computer Vision 96, 103–124 (2012)
5. Smith, B.A.: Model Free Human Pose Estimation with Application to the Classification of Abnormal Human Movement and the Detection of Hidden Load. Virginia Polytechnic Institute and State University, Virginia (2010)
6. Caillette, F., Howard, T.: Real-Time Marklerless Human Body Tracking with Multi-View 3-D Voxel Reconstruction. In: Proc. BMVC, vol. 2, pp. 597–606 (2004)
7. Gall, J., Potthoff, J., Schnörr, C., Rosenhahn, B., Seidel, H.-P.: Interacting and annealing particle filters: Mathematics and a recipe for applications. Journal of Mathematical Imaging and Vision, 1–18 (2007)
8. Kaliamoorthi, P., Kakarala, R.: Human pose tracking by parametric annealing. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (2012)
9. Sigal, L., Isard, M., Haussecker, H., Black, M.J.: Loose-limbed People: Estimating 3D Human Pose and Motion Using Non-Parametric Belief Propagation. Internal Journal of Computer Vision 98, 15–48 (2012)
10. Zhang, Z., Seah, H.S., Quah, C.K.: Particle swarm optimization for markerless full body motion capture. In: Panigrahi, B.K., Shi, Y., Lim, M.-H. (eds.) Handbook of Swarm Intelligence. ALO, vol. 8, pp. 201–220. Springer, Heidelberg (2011)
11. Kehl, R., Van Gool, L.: Markerless tracking of complex human motions from multiple views. Computer Vision and Image Understanding 104, 190–209 (2006)
12. Shen, J.-F., Yang, W.-M., Liao, Q.-M.: Multiview human pose estimation with unconstrained motions. Pattern Recognition Letters 32, 2025–2035 (2011)
13. Zhang, Z.Y.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1330–1334 (1998)
14. Cheung, K.M., Kanade, T., Bouguet, J.-Y., Holler, M.: A Real-Time System for robust 3D voxel reconstruction of human motions. In: CVPR, vol. 2, pp. 714–720 (2000)
15. Sigal, L., Balan, A.O., Black, M.J.: 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)
16. Poppe, R.: Evaluating example-based pose estimation: Experiments on the HumanEva sets. In: CVPR EhuM2: 2nd Workshop on Evaluation of Articulated Human Moition and Pose Estimation (2007)