2017 International Conference on Virtual Reality and Visualization (ICVRV)

Individual 3D Model Estimation for Realtime Human Motion Capture

Lianjun Liao*‡§, Le Su† and Shihong Xia*
*Institute of Computing Technology, CAS, Beijing, 100190
Email: LiaoLianjun@ict.ac.cn
†Civil Aviation University of China, Tianjin, 300300
‡University of Chinese Academy of Sciences, Beijing, 100049
§North China University of Technology, Beijing, 100144

***Abstract*—In this paper, we present a practicable method to estimate individual 3D human model in a low cost multi-view realtime 3D human motion capture system. The key idea is: using human geometric model database and human motion database to establish geometric priors and pose prior model; when given the geometric prior, pose prior and a standard template geometry model, the individual human body model and its embedded skeleton can be estimated from the 3D point cloud captured from multiple depth cameras. Because of the introduction of the global prior model of body pose and shapes into a unified nonlinear optimization problem, the accuracy of geometric model estimation is significantly improved. The experiments on the synthesized data set with noise or without noise and the real data set captured from multiple depth cameras show that estimation result of our method is more reasonable and accurate than the classical method, and our method is better noise-immunity. The proposed new individual 3D geometric model estimation method is suitable for online realtime human motion tracking system.**

***Keywords*-Human Model Estimation, Data Driven, Human Motion Capture**

I. Introduction

Human motion capture and tracking is a hot issue in computer vision and graphics. It mainly studies how to quickly reconstruct accurate human geometry model and human motion sequence from the input depth data stream. The motion capture technology has important application value in the movie stunt, animation games, sports training and other fields, for example, the captured human motion sequence can be used to guide sports training, or improve the sense of reality of game characters. However, up to now, the existing low-cost commercial RGB-D human capture system such as Kinect, suffered from ill-pose problem caused by limbs occlusions or self-occlusions, and cannot robustly reconstruct reasonable accurate 3D human motion sequence. In contrast with single-view-based system such as [1], the multi-view based methods such as [2], can achieve even more accuracy by minimizing the influence of ill-pose problem caused by limbs occlusion or self-occlusion. Realtime human motion capture systems are also reported. However, these methods need a pre-established human model, when human body size changed significantly or a pose is not in the database,it will fail.

To address this issue,we present a practicable method to estimate individual 3D human model in a low-cost multi-view realtime 3D human motion capture system. The key idea is: Based on human geometric model database and motion database, establish geometric priors and pose prior model; With geometric priors, pose prior and a standard template geometry model, the individual human body model and its embedded skeleton can be estimated from the captured 3D point clouds from multiple depth cameras. The main contributions in this work are as follows:

- We proposed a new individual 3D geometric model estimation method suitable for online realtime human motion tracking system.
- Successfully introduced the global prior model of human body pose and shapes into the nonlinear optimization problem of human geometry model estimation, and consequently the accuracy of geometric model estimation is significantly improved.

II. Related Work

Model-free methods such as [3]–[5], considering no prior information of human body, identify human pose in a frame through image or mesh feature point detection. The drawback is that it neglects the previous frames influence on the pose of the current frame, that is, ignoring the essence of human motion as a continuous process of spatial and temporal variation. Model-based methods need an 3D model scanned in advance, such as methods based on point cloud ICP [6]–[8], or methods based on multi view depth camera [9], [10]. The cost of the 3D scanner is high, and it is time-consuming to process the scanned data. It also suffers from error accumulation and tracking the long time movement.

Data driven methods, with the help of a 3D pose database constructed in advance from captured motion data, can usually achieve compelling results. Siddiqu et al. [11] and Baak et al. [12] estimate human pose in each frame by detecting feature points from depth image. Since the model in the database is the standard 3D human body model, it cannot always get reasonable results when the size of actors is much different from the standard model in database. Ye et al. [13] retrieve the optimal match between the 3D point cloud captured from multi depth camera and the 3D human pose in the database, and then estimate the full-body pose by deforming the retrieved pose back to the captured 3D point cloud through non-rigid

978-1-5386-2636-8/19/$31.00 ©2019 IEEE
DOI 10.1109/ICVRV.2017.00055

registration. The human pose database is composed of the 3D point clouds which are calculated from the depth images which are generated by projecting a standard 3D body model driven by an embedded skeleton. Zhang et al. [2] present an efficient physics-based motion reconstruction algorithm that, integrating the input depth data from 3 Kinect cameras, foot pressure data from wearable pressure sensors and detailed full-body geometry, can reconstruct offline full-body motion (i.e. kinematic data) and human dynamic data. When tracking the 3D skeletal poses, the global PCA for features dimension reduction process is done on the 3D body posture set in the CMU motion database, the reconstructed 3D pose prior is equivalent to imposing additional range constraint on human joint angle. Zhu et al. [14] succeeded in tracking human motion by combining the semantic feature detection of human limbs based on Bayesian estimation and inverse kinematical optimization calculation with constraints such as joint limit avoidance. However with the assumption that human head in the image is always located above its waist, therefore, it can not deal with the pose that does not meet this condition. Wei et al. [1] provide a fast, automatic method for capturing full-body motion data using a single depth camera. It described the real-time 3D human posture reconstruction from monocular depth image as formulating the registration problem in a Maximum A Posteriori (MAP) framework and iteratively registering a 3D articulated human body model with monocular depth image.

In general, from the above human motion capture system, we can see that the model-based method often is more better in accuracy and the human geometry model estimation plays an important role in these methods. Therefore, our motivation of this paper is how to reasonablely and accurately estimate the individual geometry model suitable for realtime human motion system. The most related work to ours in this paper is that proposed by Anguelov et al. [15] which introduce the SCAPE method, a data-driven method, for building a human shape model that spans variation in both subject shape and pose. Generally, the SCAPE method can reconstruct a reasonable and accurate 3D human geometric model and pose. However, the SCAPE method requires accurate point correspondence between non-rigid model and the target 3D point cloud to ensure the 3D human pose accuracy estimation and large-scale deformation, and when the target 3D human pose differs largely from the template model, the result of SCAPE is obviously unreasonable.

III. Overview

We propose an effective approach to online estimate individual human geometric model for a low-cost realtime 3D human motion capture system. Fig.1 gives the pipeline of our system. The input data is the 3D point cloud captured by multi-view caliberated depth camera and a standard template human geometric model. Then the individual 3D geometric models and embedded skeletons consistent with input point clouds are accurately constructed by making full use of human geometric model database and human pose database. We will demostrate that our method can quickly reconstruct reasonable and accurate individual 3D geometric models and its embedded skeletons, and it is suitable for realtime human motion capture system.

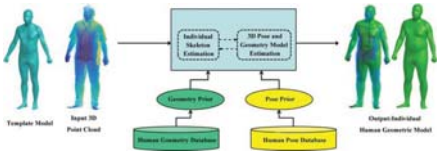


Fig. 1. System overview

IV. Data Acquisition

This section focuses on the data acquisition method from multiple depth cameras and the representation form of pose database.

A. Spatial and Temporal Alignment

There are four deep cameras (Microsoft Kinect v2.0) connected to one PC, which extended three PCI card to obtain three additional USB3.0 ports (only one USB3.0 hub on my mainboard). Since Microsoft's Kinect driver does not support multiple Kinect cameras connecting simultaneously, we use the open source device driver libfreenect2 [16] .

Temporal alignment: The frame rate of Kinect is 30fps, that is, the acquisition cycle is about 33ms. Attaching a timestamp for every depth frame when capturing, the synchronous group comprises of depth frames with time difference of less than half a period ($\leq 15ms$).

Camera parameters: There are four Kinect cameras located at the four corners of an approximate square capture scene, and all are positioned toward the center of the scene. The camera's intrinsic parameters can be read from device using libfreenect2 API. The extrinsic parameters can be easily calibrated with the help of a chessboard. We choose one camera's coordinate system as the reference coordinate system, and spatially align all the depth frames into the reference coordinate.

B. Pose Database

3D human pose representation. Just like [17], the 3D human body pose is defined as a DoF (Degree of Freedom) vector $\mathbf{q} \in \mathbf{R}^{36}$, including root (6 DoF), upperback (3 DoF), r/lclavicle (2 DoF), r/lhumerus (3 DoF), r/ lradius (1 DoF), neck (2 DoF) head (1, DoF), r/lfemur (3 DoF), r/ltibia (1 DoF) and r/lfoot (2 DoF).

3D human pose database. Same as [17], we select motion sequence database close to 2.5 hours in total time from the CMU human motion sequence database [18], and its movement types include: walking, running, boxing, kicking, jumping, dancing and waving, fitness and golf etc..

V. Geometry Model Database

3D human geometry model database. we use the CAESER human geometry model database [19] with details infomation, which contains 1517 male geometric models of A-pose and 1531 female geometric models of A-pose. The topology of all the mesh model in the database is consistent with

each other. The human geometry model is represented by a long vector $\mathbf{s}_i$ composited of the vertex set of the mesh model, and the database is represented by $\mathbf{S} = \{\mathbf{s}_i, i = 1, \cdots, N\}$.

3D human geometry model prior. With the human geometry model database $\mathbf{S}$ the global linear prior model of the human geometry model is established by principal component analysis (PCA) [20], and it is formulated as:

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{P}_{\boldsymbol{\beta},k} \cdot \boldsymbol{\beta} + \bar{\mathbf{s}} \tag{1}$$

where $\boldsymbol{\beta}$ is the low dimensional parameter vector of human geometry model, $\mathbf{P}_{\boldsymbol{\beta},k}$ is the matrix composed of vectors from the former k dimension of the principal component vectors, and $\bar{\mathbf{s}}$ is the average vector of human geometric models in the database. In our experiments, the principal component ratio is set to 95%, and the mean model is used as the template model. Moreover, the priors of the male and female human geometry models are constructed separately.

VI. Skeleton Estimation

Employing skeleton driven method to deform the human geometric model in pose dimension, we propose a novel method to automatically and accurately estimate the embedded skeleton when given a new human geometry model consistent with the geometric topology of the template model, a known geometric template model and its embedded skeleton.

Parameterization of embedded skeleton joint. Given the human geometric template model and its embedded skeleton, the coordinates $\mathbf{J}_i$ of the joint centers of the human skeleton can be expressed as weighted linear combinations of coordinates of their "nearest neighbors" vertices. It can be formalized as equation (2). Therefore, we can obtain the vertex weights when given the template geometric model and its embedded skeleton.

$$\mathbf{J}_i = \sum_{\mathbf{v}_{i,j} \in \mathbf{V}_i} w_{i,j} \cdot \mathbf{v}_{i,j} \tag{2}$$

where $\mathbf{V}_i$ is a set of "nearest neighbors" vertices of $i-th$ joint of the embedded skeleton, $\mathbf{v}_{i,j}$ is the coordinate of $j-th$ vertex in $\mathbf{V}_i$, $w_{i,j}$ is a weight corresponding to the vertex $\mathbf{v}_{i,j}$. $\mathbf{J}_i$ is the coordinate of the embedded skeleton joint i.

Solution of vertex weight w. Given the human template geometric model and its embedded skeleton, the geometric model vertices can be estimated and formalized as a constrained linear least squares problem:

$$\begin{aligned} &\arg\min_{\mathbf{w}_j} \| \sum_{\tilde{\mathbf{v}}_{i,j} \in \tilde{\mathbf{V}}_i} w_{i,j} \cdot \tilde{\mathbf{v}}_{i,j} - \tilde{\mathbf{J}}_i \|^2 \\ &\text{subject to } w_{i,j} \geq 0 \end{aligned} \tag{3}$$

where $\tilde{\mathbf{J}}_i$ is the coordinate of the embedded skeleton joint i of the template model, $\tilde{\mathbf{V}}_i$ is a set of "nearest neighbors" vertexs of ith joint of the embedded skeleton, $\tilde{\mathbf{v}}_{i,j}$ is the coordinate of jth vertex in $\tilde{\mathbf{V}}_i$, $w_{i,j}$ is a weight corresponding to the vertex $\tilde{\mathbf{v}}_{i,j}$. Those $\tilde{\mathbf{V}}_i$, $\tilde{\mathbf{v}}_{i,j}$, and $\tilde{\mathbf{J}}_i$ are known variables, and $w_{i,j}$ is the pending variable.

Estimation of embedded skeleton. The estimated individual geometry model has the same mesh topology as the template model, and all of them are A-pose. Therefore, according to the equation (2), when given the vertex coordinates of individual geometry model and vertex weights $w_{i,j}$, we can find joint coordinates $\mathbf{J}_i$ of embedded skeleton for the new human geometry model.

VII. Geometric Model Estimation

In this section, we will describe how to formulate the automatic estimation problem of the detailed individual geometric model as a nonlinear optimization problem and its iterative optimization method. Specifically, when given the 3D point cloud $\mathbf{P}$ captured from current frame, human geometric model database $\mathbf{S}$, pose database $\mathbf{Q}$, the individual human geometric model can be obtained (parameterized as human pose parameter vector $\tilde{\mathbf{q}}$ and human geometric model parameter vector $\widetilde{\boldsymbol{\beta}}$). it can be formalized as:

$$\begin{aligned} \tilde{\mathbf{q}}, \widetilde{\boldsymbol{\beta}} = \arg\min_{\mathbf{q},\boldsymbol{\beta}} \; & \lambda_1 E_{point} + \lambda_2 E_{plane} + \lambda_3 E_{\beta-prior} \\ & + \lambda_4 E_{bone-balance} + \lambda_5 E_{q-prior} \end{aligned} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are the weights of each energy item respectively. By introducting of $E_{q-prior}$ and $E_{\beta-prior}$ into the optimization objective function as the global prior of human body pose and shapes, the rationality and accuracy of individual 3D model can be significantly improved.

Usually the pose of a real actor differs from the standard A-pose in the model database, and it may be very small or large. In order to accurately estimate detailed individual mesh model, the template model should deform in the shape dimension,as well as in its pose dimension. In the iteration process, assuming that the topology of the embedded human skeleton remains the same, and the length of the skeleton will change naturally with the changes of the human geometry model.

Corresponding points term. It is to minimize the distance between the vertex set of the reconstructed human geometry model and the captured 3D point cloud. In order to improve the accuracy and rationality of nearest point matching, both the point-to-point and point-to-plane metric [21], [22] are used.

$$E_{point} = \sum_i \|\mathbf{v}_i(\mathbf{q},\boldsymbol{\beta}) - \mathbf{p}_i^*\|_F^p \tag{5}$$

$$E_{plane} = \sum_i \|\mathbf{n}_i^T(\mathbf{q},\boldsymbol{\beta}) \cdot (\mathbf{v}_i(\mathbf{q},\boldsymbol{\beta}) - \mathbf{p}_i^*)\|_F^p \tag{6}$$

where $\|\bullet\|_F^p$ is the $p-norm$. The point-to-point distance refers to the closest Euclidean distance between the human geometry model vertex $\mathbf{v}_i(\mathbf{q},\boldsymbol{\beta})$ and the captured 3D point $\mathbf{p}_i^*$. The point-to-plane distance refers to the distance from captured 3D point cloud to intersection point of the tangent plane of captured 3D point cloud $\mathbf{p}_i^*$ and the normals $\mathbf{n}_i^T(\mathbf{q},\boldsymbol{\beta})$ for vertex $\mathbf{v}_i(\mathbf{q},\boldsymbol{\beta})$ of human geometry model. Usually, increasing ratio of E_{plane} to E_{point} can guarantee good convergence of algorithm [22], and it is set to 1:3 in the experiment.

Symmetric skeleton length term. It encourages the symmetrical skeletal segments to be as the same in length as possible. In our experiments, we suppose only the four limbs are symmetry.

$$E_{bone-balance} = \sum_{\{(m,n)\}\in\{(M,N)\}} \|l_m - l_n\|^2 \tag{7}$$

where $\{(M,N)\}$ is symmetrical skeletal segments, l_m, l_n is respectively the left and right symmetry bone length.

Human geometry model prior term. It is a penalty to satisfaction degree of the probability distribution of the reconstructed individual geometry model and the probability distribution of the global space composed of the human geometry model in database. Assuming that the human geometry model database in the global space is subject to multivariate normal distribution, the the human geometric model prior is defined to maximize the following conditional probability (equation (8)). In general, the probabilistic maximization problem can be converted into the energy minimization problem (equation (9)):

$$Pr(\mathbf{s}(\boldsymbol{\beta})|\mathbf{s}_1,\cdots,\mathbf{s}_N) = \frac{\exp(-\frac{1}{2}(\mathbf{s}(\boldsymbol{\beta})-\bar{\mathbf{s}})^T\boldsymbol{\Lambda}_{\boldsymbol{\beta}}^{-1}(\mathbf{s}(\boldsymbol{\beta})-\bar{\mathbf{s}}))}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Lambda}_{\boldsymbol{\beta}}|^{\frac{1}{2}}} \tag{8}$$

$$E_{\beta-prior} = (\mathbf{s}(\boldsymbol{\beta})-\bar{\mathbf{s}})^T\boldsymbol{\Lambda}_{\boldsymbol{\beta}}^{-1}(\mathbf{s}(\boldsymbol{\beta})-\bar{\mathbf{s}}) = \boldsymbol{\beta}^T\mathbf{P}_{\boldsymbol{\beta},k}^T\boldsymbol{\Lambda}_{\boldsymbol{\beta}}^{-1}\mathbf{P}_{\boldsymbol{\beta},k}\boldsymbol{\beta} \tag{9}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$ is the matrix composed of the former k principal component vector from the covariance matrix human geometry model database, $\boldsymbol{\beta}$ is a pending low dimensional parameter vector of human body model, the $\mathbf{P}_{\boldsymbol{\beta},k}$ and $\bar{\mathbf{s}}$ have obtained respectively in section V, the matrix composed of the former k principal component vector from the human geometric model prior and the average vector.

Human pose prior term. The penalty is the satisfaction degree of the probability distribution of the reconstructed individual human pose and that of the global space composed of the human pose database. Given human pose databases $\mathbf{Q} = \{\mathbf{q}_i, i = 1,\cdots,H\}$, this section also employs the principal component analysis (PCA) [20] to establish a global linear prior model of human pose. It can be formalized as:

$$\mathbf{q} = \mathbf{P}_{\mathbf{q},b}\cdot\mathbf{w} + \bar{\mathbf{q}} \tag{10}$$

Where $\mathbf{w}$ is the low dimensional parameter vector of human body pose, $\mathbf{P}_{\mathbf{q},b}$ is the matrix composed of the former b principal component vector, and $\bar{\mathbf{q}}$ is the mean vector of human pose in the database. The principal component ratio is set to 95%. Similarly, assuming that the human pose database in the global space is subject to multivariate normal distribution, then the human pose prior can be defined to maximize the following conditional probability equation (11); It can also be converted into equation (12):

$$Pr(\mathbf{q}|\mathbf{q}_1,\cdots,\mathbf{q}_H) \propto \exp(-\frac{\|\mathbf{P}_{\mathbf{q},b}^T\cdot(\mathbf{P}_{\mathbf{q},b}\cdot(\mathbf{q}-\bar{\mathbf{q}}))+\bar{\mathbf{q}}-\mathbf{q}\|^2}{2\delta_{q-prior}^2}) \tag{11}$$

$$E_{q-prior} = \|\mathbf{P}_{\mathbf{q},b}^T\cdot(\mathbf{P}_{\mathbf{q},b}\cdot(\mathbf{q}-\bar{\mathbf{q}}))+\bar{\mathbf{q}}-\mathbf{q}\|^2 \tag{12}$$

where $\mathbf{q}$ is a pending human pose vector, The $\delta_{q-prior}$ is a soft constraint form of body pose prior.

Optimization method. By substituting equation (5), (6), (7), (9) and (12) into the equation (4), the final equation can be obtained. In our experiment, let $p = \frac{L_2}{L_1}$. The variables $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are the weights of each energy item, and set to 1, 3, 1E3, 0.03 and 10, respectively in our experiments. The nonlinear optimization problem, equation (4), about the human pose parameter vector $\tilde{\mathbf{q}}$ and human geometric model parameter vector $\widetilde{\boldsymbol{\beta}}$, is solved by make using of joint optimization strategy based on the Particle Swarm Optimization (PSO) [23]–[25]. When the optimization is done, the number of iterations is 10 times, and 2000 particles are generated randomly at each iteration step. Since the independence between the particles, GPU is used to accelerate the computation. In the experiment, the efficiency of the algorithm was tested 200 times. Averagely in 6-7 iterations, it will converge to the approximate optimal solution.

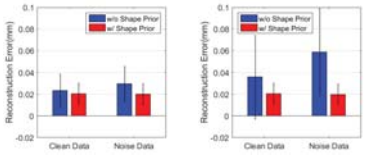
VIII. Results

In the following experiments, several virtual human models with large differences in stature and real actors were randomly selected in a variety of different pose (A-pose and other pose), and by comparing with the SCAPE [15] method which is a state-of-the-art data-driven method and is the most related work to ours in this paper, we will demonstrate the accuracy and rationality of our individual geometry model estimation method. The experimental environment is a PC machine with: a CPU of Intel Core i7-6850K 3.6GHz with 6 processor 12 thread, main memory size of 128GB, a CUDA enabled graphics card of NVIDIA GeForce GTX 1080 of 8GB memory based on Pascal architecture 1.733GHz.

Evaluation of pose and shape prior term The results from the experiment with or without $E_{q-prior}$ and $E_{\beta-prior}$ are shown in Fig.2 and Fig.3(a) respectively, and it proved that these two energe term can be significantly improved the rationality and accuracy of the reconstructed model.

Fig. 2. Evaluation of pose prior. Top row is the result of without $E_{q-prior}$, and the bottom row with $E_{q-prior}$.

Synthesized clean 3D point cloud In the experiment, the average model in CAESER database is used as the human template model. The SCAPE model is trained on the template model, which consists of 70 human template models in different pose (based on the multi-view pose human database supplied by SCAPE, obtained by the method [26]). The purpose of this experiment is to compare the modeling ability of our method with SCAPE [15].The input data are the 3D



(a) For shape prior. (b) Error Comparison

Fig. 3. Reconstruction Error

point cloud, which is synthesized from 45 male human body models of three different poses (one similar to the template A-pose, another two different obviously) randomly selected from the CAESER database.

Experiment on 45 test data of 3 kinds of poses, the average reconstruction error of human geometrical models was shown in Table I. Pose1 is of the pose in first row in Fig.4, Pose2 and Pose3 are second and third row respectively. The partial comparison results are shown in figure 5. It shows that our method is more accurate and robust than the classical SCAPE method. Show in Fig.4;

TABLE I
RECONSTRUCTION ERROR IN CLEAN DATA

	SCAPE's	Ours
Pose1	$1.79 \pm 0.88cm$	$1.74 \pm 0.85cm$
Pose2	$2.87 \pm 1.52cm$	$2.13 \pm 1.08cm$
Pose3	$4.4 \pm 5.6cm$	$2.1 \pm 1.02cm$

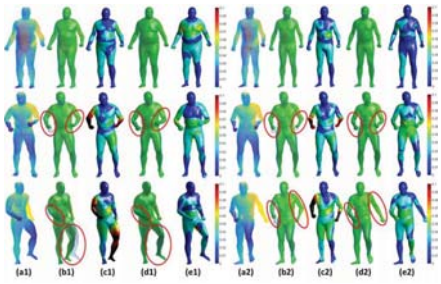


Fig. 4. Comparison test on synthetic clean data. Col. a1 and a2 are the input, b1 and b2 are the results of SCAPE [15], c1,c2 are its reconstruction errors, d1and d2 are the results of ours, e1 and e2 are our errors.

Synthesized noisy 3D point cloud This experiment is the same as previous one except that the input data are the 3D point cloud added with random Gaussian noise. From Table II and Fig.5, it also shows that our method is more accurate and robust than the classical SCAPE method.

TABLE II
RECONSTRUCTION ERROR OF NOISY DATA

	SCAPE's	Ours
Pose1	$2.69 \pm 1.21cm$	$1.76 \pm 0.87cm$
Pose2	$9.04 \pm 6.79cm$	$2.19 \pm 1.1cm$

Comparing the reconstruction results from noisy and clean synthetic data set, as shown in Fig.3(b), it is obvious that our method reconstruction accuracy achieves the same well, while the reconstruction error of SCAPE is much different suffering noisy data. Therefore, our method performs better than SCAPE in the anti-noise ability.

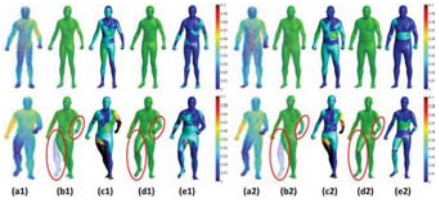


Fig. 5. Comparison test on synthetic noisy data. Col. a1 and a2 are the input, b1 and b2 are the results of SCAPE [15], c1,c2 are its reconstruction errors, d1and d2 are the results of ours, e1 and e2 are our errors.

Online test on real actors. In the previous experiment using synthetic 3D point cloud, the virtual human model is a scanning model with only underwear, but in actual scenarios, the actors are normally dressed. The online test on normally dressed actors, comparing the modeling ability of our method with SCAPE, also shows that our method is more accurate and robust than the classical SCAPE method.Shown as in Fig.5.

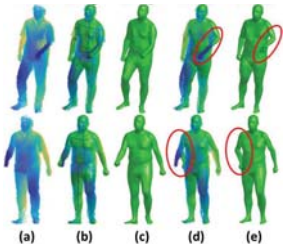
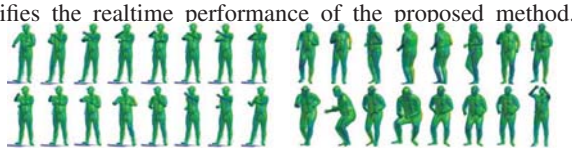


Fig. 6. Online test results. column a is the captured point cloud, b,c is the result of ours, d,e is SCAPE's

According to the above experimental results, both our method and SCAPE [15] can reconstruct a reasonable and accurate 3D geometric model and pose. However, when the human pose differs largely from the template model, our method can reasonably and accurately reconstruct the 3D body shape and pose, while the result of SCAPE is obviously unreasonable (shown as red circles highlight part in Fig.5 and Fig.6). The main reason is: firstly, the SCAPE requires accurate point correspondence between non-rigid model and the target to ensure the 3D human pose accuracy estimation and large-scale deformation. These accurate correspondences are usually required specified [15] by hand or like [27] given actor's height and weight information, to estimate an initial pose very similar to (or consistent) the pose of target point cloud. Secondly, in our method, the deformation of limb segments of 3D human pose is rigid deformation, and it can guarantee reasonable large scale deformation. However, in the SCAPE method, it is non-rigid deformation and can not guarantee the large scale pose deformation. In addition, the method of iterative optimization algorithm is the variable to solve in our iterative optimization algorithm composed of a shape and pose parameters of low dimensional vector (in the experiment is 51 dimensional), while in the SCAPE method is a transformation matrix of each patch (close to 100,000 dimension in the experiment), it is obvious that our efficiency is better than the SCAPE method. Therefore, our method is

more suitable for online application than the SCAPE method.

Application In this experiment, actors with different size were asked to complete a variety of motions (including: walking, walking and kicking, boxing, and etc.), Some test results are shown in Fig.7(a) and Fig.7(b), respectively. The experimental results show that our motion tracking method can be applied to different people of different size,suitable for a variety of daily action, and get accurate and reasonable motion estimation results. It verifies the effectiveness of the proposed method. In addition, the average frame rate of the above captured 3D human motion sequences is 20.25fps, which verifies the realtime performance of the proposed method.



(a) Alternate rolling hands. (b) Heterogeneous Motion

Fig. 7. Application Results

IX. Conclusion

This paper focuses on how to estimate the individual geometric model of human body quickly and reasonably based on the global human pose and shape prior model, from the 3D point cloud captured from multiple depth cameras. Our method successfully integrates the prior information of human geometric model and human pose into a unified optimization problem, making the estimated human model more accurate and reasonable. Due to the use of multi-view depth camera data, the ill-pose problem caused by occlusion or self-occlusion can be solved to some extent. The experiments show that based on our individual geometric model estimation method, it is easy to develop a low-cost, online realtime and accurate acquisition system for 3D human motion sequences.

However, the influence of muscle and skin on the deformation of human model geometry has not been considered in our method, and it will result in the lacking of local details of the reconstructed model. We will try to address this problem in the future research work.

Acknowledgment

This work was supported by the Knowledge Innovation Program of the Institute of Computing Technology of the Chinese Academy of Sciences under Grant No. ICT20166040, the Science and Technology Service Network Initiative of Chinese Academy of Sciences under Grant No. KFJ-STS-ZDTP-017, the National Natural Science Foundation of China under Grant No. 61772499.

References

[1] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–12, 2012.

[2] P. Zhang, K. Siu, J. Zhang, C. K. Liu, and J. Chai, "Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture," *ACM Trans. Graph.*, vol. 33, no. 6, p. 221, 2014.

[3] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 3108–3113.

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[5] R. Girshick, J. Shotton, P. Kohli, and A. Criminisi, "Efficient regression of general-activity human poses from depth images," in *International Conference on Computer Vision*, 2011, pp. 415–422.

[6] J. W. Daniel Grest and R. Koch, "Nonlinear body pose estimation from depth images," in *Joint Pattern Recognition Symposium*, 2005, pp. 285–292.

[7] S. Knoop, S. Vacek, and R. Dillmann, "Fusion of 2d and 3d sensor data for articulated body tracking," *Robotics and Autonomous Systems*, vol. 57, no. 3, pp. 321–329, 2009.

[8] D. Grest, V. Krger, and R. Koch, "Single view motion tracking by depth and silhouette information," in *Image Analysis, Scandinavian Conference, Scia 2007, Aalborg, Denmark, June 10-14, 2007, Proceedings*, 2007, pp. 719–729.

[9] B. Kai, R. Kai, Y. Schroeder, C. Bruemmer, A. Scholz, and M. A. Magnor, "Markerless motion capture using multiple color-depth sensors," in *Vision, Modeling, and Visualization Workshop 2011, Berlin, Germany, 4-6 October*, 2011, pp. 317–324.

[10] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, *Performance Capture of Interacting Characters with Handheld Kinects*. Springer Berlin Heidelberg, 2012.

[11] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real," *Dissertations and Theses-Gradworks*, vol. 32, no. 12, pp. 1–8, 2010.

[12] A. Baak, M. Mller, G. Bharaj, and H. P. Seidel, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *International Conference on Computer Vision*, 2011, pp. 1092–1099.

[13] M. Ye, R. Yang, and M. Pollefeys, "Accurate 3d pose estimation from a single depth image," in *IEEE International Conference on Computer Vision*, 2011, pp. 731–738.

[14] Y. Zhu, B. Dariush, and K. Fujimura, "Kinematic self retargeting: A framework for human pose estimation," *Computer Vision and Image Understanding*, vol. 114, no. 12, pp. 1362–1375, 2010.

[15] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *Acm Transactions on Graphics*, vol. 24, no. 3, pp. 408–416, 2005.

[16] "libfreenect2," 2017. [Online]. Available: https://github.com/OpenKinect/libfreenect2

[17] L. Su, J. X. Chai, and S. H. Xia, "Local pose prior based 3d human motion capture from depth camera," *Ruan Jian Xue Bao/Journal of Software(in Chinese)*, vol. 27, no. 2, pp. 172–183, 2016.

[18] "Mocap database," 2017. [Online]. Available: http://mocap.cs.cmu.edu/

[19] "Caeser human geometric model database," 2017. [Online]. Available: http://store.sae.org/caesar/

[20] K. P. F.R.S, "Liii. on lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[21] K. L. Low, "Linear least-squares optimization for point-to-plane icp surface registration," *Chapel Hill*, 2004.

[22] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM Trans. Graph.*, vol. 28, no. 5, p. 175, 2009.

[23] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, no. 0, 1995, pp. 1942–1948.

[24] F. Marini and B. Walczak, "Particle swarm optimization (pso). a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, no. B, pp. 153–165, 2015.

[25] J. C. Bansal, P. K. Singh, M. Saraswat, A. Verma, S. S. Jadon, and A. Abraham, "Inertia weight strategies in particle swarm optimization," in *Nature and Biologically Inspired Computing*, 2011, pp. 633–640.

[26] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 399–405, 2004.

[27] A. Weiss, D. Hirshberg, and M. Black, "Home 3d body scans from noisy image and range data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1951–1958.