Pacific Graphics (2019)
C. Theobalt, J. Lee, and G. Wetzstein (Editors)

Short Paper

3D human body skeleton extraction from consecutive surfaces

Yong Zhang[†1], Fei Tan[1], Shaofan Wang[1], Dehui Kong[1], and Baocai Yin[1,,2]

[1]Beijing Key Laboratory of Multimedia & Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]Graz Faculty of Electronic Information and Electrical Engineering ,Dalian University of Technology ,Dalian 116024,China

Abstract

Extracting human body skeletons from consecutive surfaces is an important research topic in the fields of computer graphics and human computer interaction, especially in posture estimation and skeleton animation. Current approaches mainly suffer from following problems: insufficient time and space continuity, not robust to background, ambient noise, etc. Our approach is to improve against these shortcomings. This paper proposes a 3D human body skeleton extraction method from consecutive meshes. We extract the consistent skeletons from consecutive surfaces based on shape segmentation and construct skeleton sequences, then we use the continuous frame skeleton point optimization model we proposed to optimize the skeleton sequences, generating the final skeleton point sequences which are more accurate. Finally, we verify that our method can obtain more complete and accurate skeletons compared to other methods through many experiments.

1. Introduction

Extracting 3D human body skeletons from geometric surface from the grid sequence is an important research content in the fields of computer graphics, human-computer interaction, etc., and has important applications in pose estimation [SHRB11, PAG11], human body modeling [STG*97, BAS14] and skeleton manipulation [Fêd03, SSC03]. While many research work are devoted to human body skeleton extraction from static point clouds, existing methods cannot fully explore the spatial or temporal coherence of human poses and hence lead to low accuracy [LHW*13]. Whether in group skeleton multi-granular real-time extraction and tracking technology of two-dimensional, or in Kinect bone tracking data processing of 3D, Workers usually use the human body skeleton of 20 points which is the better reflection of actual human(the number of skeleton points on the limbs is 4, one skeleton point in the middle of the ankle, one skeleton point in the waist, one skeleton point in the middle of the shoulder, and one skeleton point in the head) in both group skeleton multi-granular real-time extraction and tracking technology of 2D and Kinect bone tracking data processing of 3D.

Existing skeleton extraction methods can be roughly divided into two categories: point clouds based methods [TZCO09,LHW*13, ZSW*18], and meshes based methods [TAOH12, CTO*10]. However, for the 3D human body, the number of skeleton points extracted by the above methods are inconsistent, incomplete, error branch or partial point position deviation and the original 3D human body cannot be better represented because of point cloud blocking and point cloud loss. Therefore, the human body skeleton extracted in this paper is necessary and has certain advantages in terms of integrity, correctness and standardness, and has certain practical value and significance for subsequent bone-based animation production and 3D human body operation. We propose a spartial-temporal consistency model (STC) for 3D human body skeleton extraction. Compared with traditional skeleton extraction methods, the contributions of STC are summarized as follows:

- The entire process of skeleton extraction is fully automated.
- The 3D human body skeleton extracted without manual intervention is a skeleton with 20 points which better represents the actual skeleton of the human body.
- The 3D human body skeletons we extracted are more suitable for applications such as post-skeleton animation.

2. Data preprocessing

We first give notations which shall be used in this paper. $\|\cdot\|_2, \|\cdot\|_0$ denote the ℓ_2, ℓ_0 norm of a vector or a matrix, respectively. $[\mathbf{A}]_{i,j}$ denotes the element of the ith row, jth column of a matrix $\mathbf{A}$, and $[\mathbf{A}]_j$ denotes jth column of a matrix $\mathbf{A}$.

The data preprocessing consists of three steps.

Multiview image collection: We collect multiview images of a moving human body of each action using the light field acquisition device (see Fig. 1), which contains 50 industrial cameras with a given frame rate.

Point cloud generation, normalization and alignment: We generate a 3D dense point cloud of human body using Patch based

[†] Chairman Eurographics Publications Board

DOI: 10.2312/pg.20191339

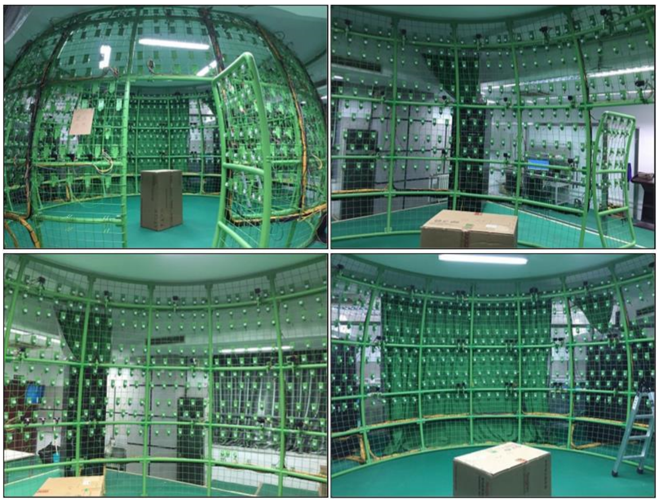

Figure 1: A light field acquisition system.

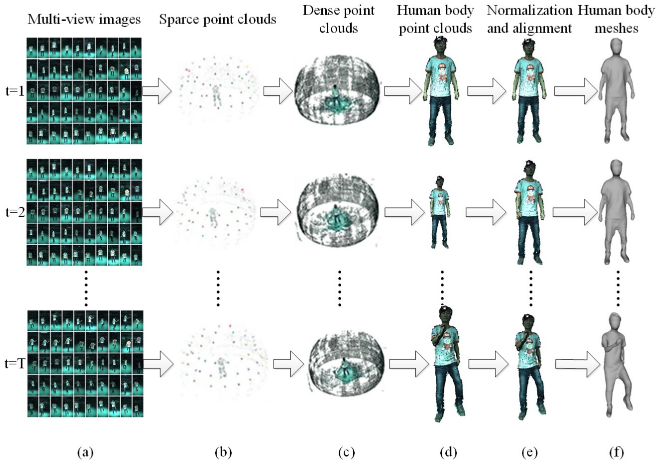

Figure 2: Triangular mesh reconstruction. Left to right: multiview image collection, sparse point cloud reconstruction, dense point cloud reconstruction, removing irrelevant points, and Poisson surface reconstruction.

Multiview Stereopsis (PMVS) based on camera parameters estimated by Structure from Motion [HZ08], and perform a normalization and alignment scheme aim to scale to unitBox and move to origin on the point cloud so that all point clouds of human body of an action sequence share similar sizes, geometric centers and orientations.

Triangular mesh reconstruction: To perform a semantic segmentation on human body, we require a mesh representation of human body besides the point cloud model. Thus we downsample the dense point cloud by merging multiple points within the same grid box into a single point, and then perform Poisson surface reconstruction to obtain a triangular mesh of human body (See Fig. 2).

3. Details of STC

We introduce the details of STC in this section. Fig. 3 shows a flowchart of STC, which mainly consists of three stages, each of which is detailed in the following subsections.

3.1. Initial skeleton extraction

The first stage of STC extracts initial skeletons from point clouds of each frame individually based on a semantic segmentation of triangular meshes of human body. Specially, this stage consists of four steps which are introduced as follows.

Pseudo-skeleton generation: We segment the mesh into several sematic patches using [KO19], and generate "pseudo-skeletons" using the centroid of each patch.

Determination of CShoulder and Waist: We connect each pair of pseudo-skeletons belonging to adjacent semantic patches with an edge, and CShoulder is recognized as the unique pseudo-skeleton which achieves the maximum degree. Similarly, Waist is recognized as the unique pseudo-skeleton which achieves degree three.

Determination of LShoulder and RShoulder: We set the patch corresponding to CShoulder as the target patch, and select the leftmost adjacent patch (i.e. left upper arm) and rightmost adjacent patch (i.e. right upper arm) of the target patch. Then we divide the points of target patch into three subpatches according to an equivdistant rule with respect to the leftmost and rightmost patches. Finally LShoulder and RShoulder are determined by the centroid of the leftmost and rightmost subpatches, respectively.

Standard skeleton completion: To fulfill an initial skeleton extraction with the same number and similar locations to standard skeletons, we divide the collection of all pseudo-skeletons and those four skeletons into six subsets corresponding to six components of human body: Torso, Head, LArm, RArm, LLeg, RLeg, according to their connectivity (see Fig. 4).Then we add or remove skeleton points to each component until the number of pseudo-skeletons reaches the standard number for current component.

3.2. Skeleton alignment

The second stage of STC is to match skeleton points between consecutive frames, i.e., to establish the correspondence between skeletons of different frames so that all the same skeletons are correctly matched. Since the blocks where Head and Neck locate and where Waist and center points of LShoulder and RShoulder locate can be judged based on the number of points on each branch and the connection with the center point of the Shoulder or Waist. Obviously, Head and Waist can be easily realize inter-frame match separately, and it is easy to distinguish two arms and two legs. To find a correspondence between two arms (and two legs) of pairwise adjacent frames, we denote $\mathbf{x}_{t,i} \in \mathbb{R}^3$ to be the coordinates of the ith skeleton of the tth frame; if

$$\sum_{i=2,3,4,5} \|\mathbf{x}_{t,i} - \mathbf{x}_{t+1,i}\|_2^2 < \sum_{i=2,3,4,5} \|\mathbf{x}_{t,i} - \mathbf{x}_{t+1,i+4}\|_2^2$$

holds, then the skeletons of two arms of the $(t+1)$th frame are correctly matched; otherwise we switch the skeletons of two arms

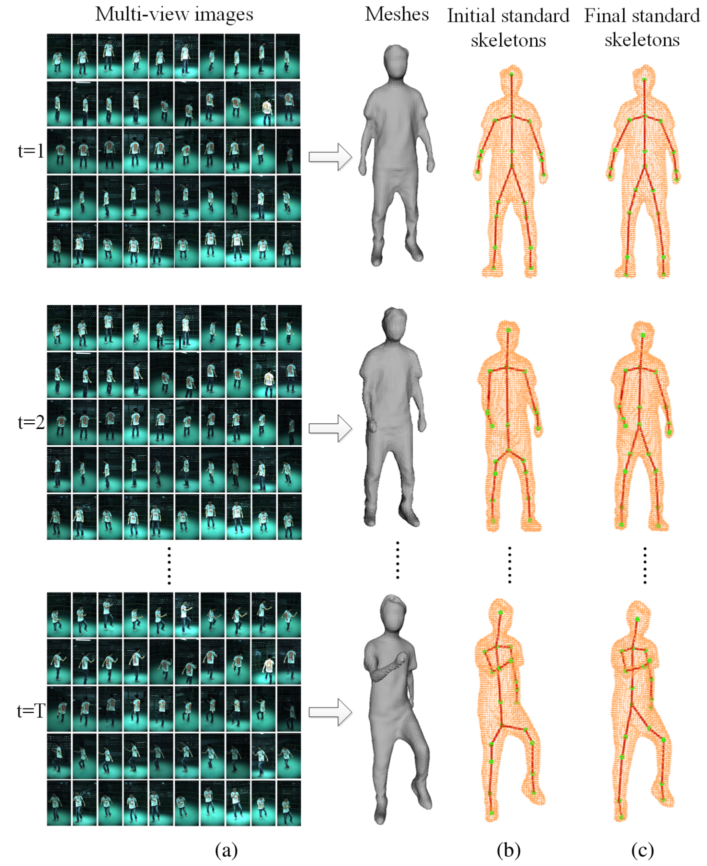

Figure 3: A flowchart of spatial-temporal consistency model. (a) Data preprocessing; (b) Initial skeleton extraction; (c) Skeleton adjustment.

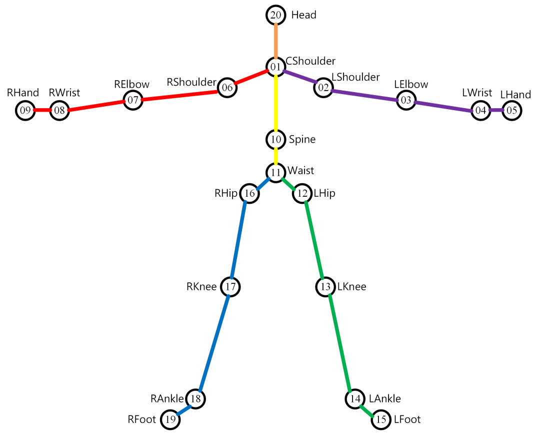

Figure 4: Standard human model of 20 body skeletons marked by black spheres and 6 body components marked by colored line segments: Torso, Head, LArm, RArm, LLeg, RLeg. The parent relationship of all nodes of the tree is given by (1).

of the $(t+1)$th frame from LArm to LArm. The correspondence between two legs is computed in a similar fashion.

The x,y,z coordinates of initial skeletons obtained in this section are denoted by $\mathbf{X}^{(1)}_{\text{init}},\mathbf{X}^{(2)}_{\text{init}},\mathbf{X}^{(3)}_{\text{init}}\in\mathbb{R}^{T\times 20}$, respectively, where $T,20$ are the frame number and skeleton number, respectively, and the tth row of $\mathbf{X}^{(k)}_{\text{init}}$ corresponds to the coordinates of initial skeletons at the tth frame, $k=1,2,3$, $t=1,\ldots,T$.

3.3. Skeleton adjustment

The third stage of STC adjusts the skeletons by using a spatial-temporal consistency adjustment model. As the position of each skeleton of a motion sequence exhibit continuous change, i.e., for almost all frames, the position of a skeleton can be given by the median value of the positions of the same skeleton of the front frame and latter frame; for another, for each frame, semantic segmentation produced by [KO19] is imprecise: most non-root skeletons locate far from the corresponding "parent skeletons" determined by

$$\begin{array}{llll}\text{parent}(02)=01 & \text{parent}(03)=02 & \text{parent}(04)=03 & \text{parent}(05)=04\\ \text{parent}(06)=01 & \text{parent}(07)=06 & \text{parent}(08)=07 & \text{parent}(09)=08\\ \text{parent}(10)=01 & \text{parent}(11)=10 & \text{parent}(12)=11 & \text{parent}(13)=12\\ \text{parent}(14)=13 & \text{parent}(15)=14 & \text{parent}(16)=11 & \text{parent}(17)=16\\ \text{parent}(18)=17 & \text{parent}(19)=18 & \text{parent}(16)=11, & \end{array} \tag{1}$$

except four ending skeletons (LHand, RHand, LFoot, RFoot) which locate close to their "parent skeletons". The reason is that each of those four skeletons locates at the end of a body component, and the segmented patch produced by [KO19] cannot distinguish that skeleton and its parent skeleton. Based on the argument, we propose the following spatial-temporal consistency adjustment model:

$$\begin{aligned}&\min_{\{\mathbf{X}^{(k)}\}_{k=1}^{3}}\sum_{k=1}^{3}\|\mathbf{D}\mathbf{X}^{(k)}\|_0+\alpha\sum_{k=1}^{3}\sum_{j=2}^{20}\left\|[\mathbf{X}^{(k)}]_j-(1-\varepsilon)[\mathbf{X}^{(k)}_{\text{init}}]_j-\beta_j\varepsilon[\mathbf{X}^{(k)}_{\text{init}}]_{\text{parent}(j)}\right\|_2^2\\ &\beta_j=\begin{cases}1 & \text{if } j=2,3,6,7,11,12,13,16,17,20\\ -1 & \text{if } j=4,5,8,9,14,15,18,19\end{cases},\ j=2,\ldots,20,\\ &[\mathbf{D}]_{i,j}=\begin{cases}-1 & \text{if } 2\le i\le T-1\wedge i=j\pm 1\\ 2 & \text{if } 2\le i\le T-1\wedge i=j\\ 0 & \text{otherwise}\end{cases},\ i,j,=1,\ldots,T,\end{aligned} \tag{2}$$

where the first term enforces the medium representation of skeletons of almost all frames, with $\mathbf{D}\in\mathbb{R}^{T\times T}$ representing the "median representation" matrix, and the second term enforces a framewise fine-tuning over all non-root skeletons for approaching or keeping away from the corresponding parent skeletons, with β_j being a pre-given parameter for determining whether each skeleton approach or keep away from its parent.

Model (2) can be solved by applying naive Lagrange multiplier

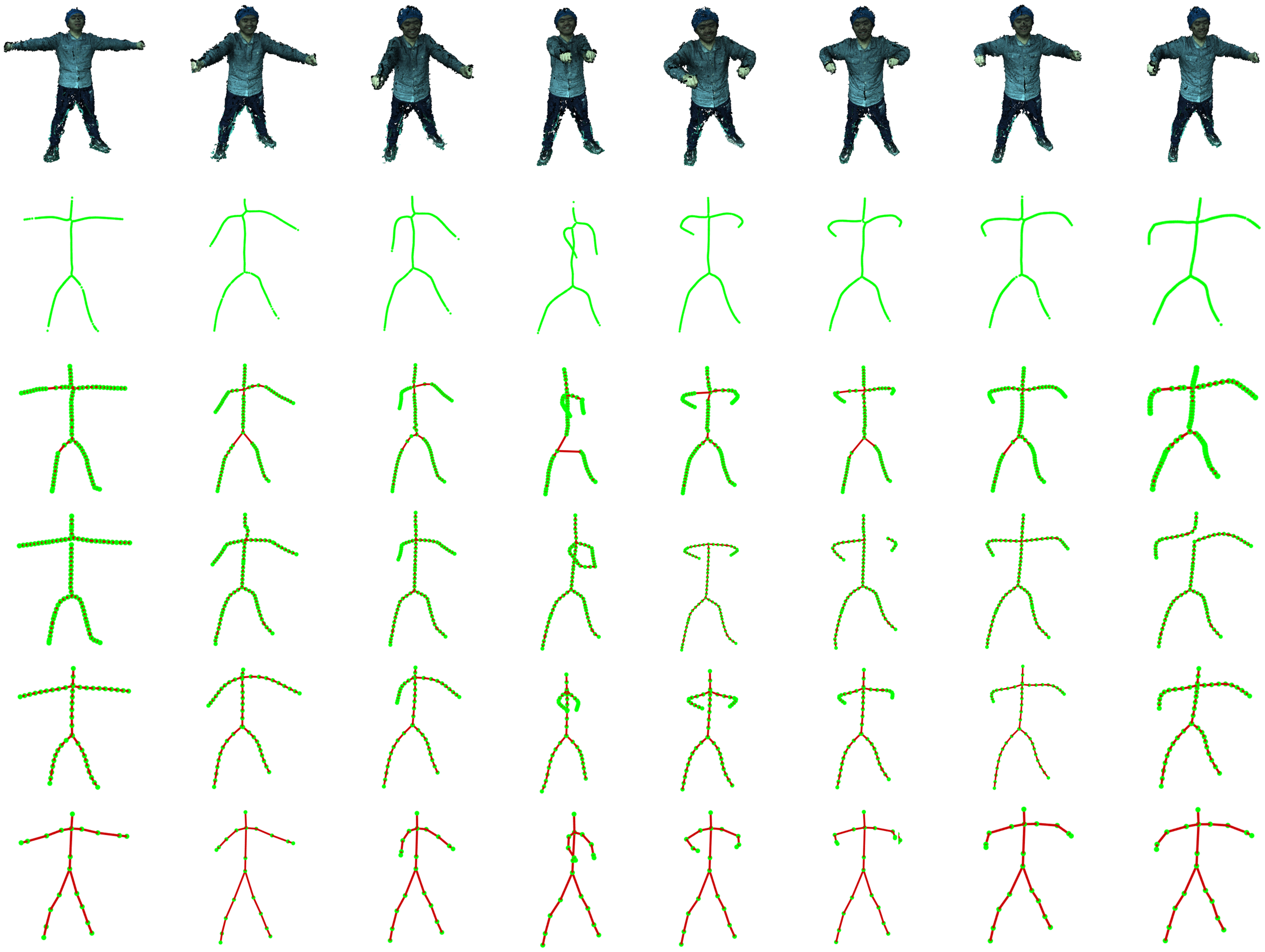

Figure 5: Qualitative results of Tagliasacchi et al. [TAOH12] (row 2), Cao et al. [CTO*10] (row 3), Huang et al. [LHW*13] (row 4), Zhang et al. [ZSW*18] (row 5) and STC (row 6) of 25th, 50th, 55th, 63rd, 76th, 84th, 88th and 90th frames of **Arm stretching**

.

method to the following Lagrangian function:

$$\min_{\mathbf{X}^{(k)},\mathbf{Y}^{(k)}} \sum_{k=1}^{3} \left(\|\mathbf{Y}^{(k)}\|_0 + \alpha \sum_{j=2}^{20} \left\| [\mathbf{X}^{(k)}]_j - \mathbf{z}_{jk} \right\|_2^2 + \lambda \|\mathbf{Y}^{(k)} - \mathbf{D}\mathbf{X}^{(k)}\|_2^2 \right)$$

$$\mathbf{z}_{jk} = (1-\varepsilon)[\mathbf{X}^{(k)}_{\text{init}}]_j + \beta_j \varepsilon [\mathbf{X}^{(k)}_{\text{init}}]_{\text{parent}(j)},\ k = 1,2,3 \qquad (3)$$

where $\mathbf{Y}^{(k)} \in \mathbb{R}^{T\times 20}$, $k = 1,2,3$ are auxiliary variable matrices for replacing $\mathbf{D}\mathbf{X}^{(k)}$, and $\lambda \in \mathbb{R}^{+}$ is the penalty parameter. We solve (3) by alternating solving two subproblems regarding $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$.

4. Experimental results

In this section, we evaluate the effectiveness of STC by comparing it with state-of-the-art methods. We collect multiview intensity images of the **Arm stretching** action by using 50 industrial cameras with 2.2 million pixels through the light field acquisition system (see Fig. 1), and the resolution of captured images is 2048*1088. All the experiments are conducted on an Intel(R) Core(TM) i5-8250U CPU 1.8 GHZ CPU with 8GB RAM using MATLAB R2016.

We select four state-of-the-art methods for comparative experiments: Tagliasacchi et al. [TAOH12], Cao et al. [CTO*10], Huang et al. [LHW*13], and Zhang et al. [ZSW*18], and show qualitative results for **Arm stretching** in Fig. 5. We summarize the main shortcomings of comparative methods as follows.

Cao et al. [CTO*10] suffer from missing of skeletons, especially on the junction of LArm and torso, the junction of RArm and torso the junction of LLeg and torso, the junction of RLeg and torso (see the 25th, 50th, 55th, 63rd, 76th, 84th, 88th and 90th frames of Fig. 5) and great prediction errors on the junction of LArm and torso, the junction of LKnee and torso (see 63rd frame of Fig. 5).

Huang et al. [LHW*13] suffer from obvious problems such as missing of skeleton points (the 76th, 84th, 90th frames of Fig. 5), missing of branches (the 76th of Fig. 5), incorrectness of connection between branches (see the 63rd, 90th of Fig. 5).

Zhang et al. [ZSW*18] occasionally produce incomplete skeletons on Head (see the 50th, 55th frames of Fig. 5).

In contrast, STC produces more accurate skeletons generally, without the appearance of wrong branches, and more complete than above skeletons, and are consistent, response to human posture better. Because initial standard skeleton extraction algorithm based on shape segmentation can extract the 3D human body skeleton with 20 points. The temporal consistency preserving skeleton optimization algorithm has the position constraints of the intra-frame skeleton points and the position constraints of inter-frame skeleton points. Our optimization model make the final standard skeletons are more accurate, more tidy, and more conformable to the original input surfaces, more in line with the actual human body skeleton points distribution. Therefore, the method proposed in this paper is better than the traditional skeleton extraction method, and is more convenient to be used by subsequent posture estimation, human body modeling and operation.

5. Conclusion

We propose a sort of 3D human body standard skeleton extraction method from consecutive surfaces, which can generate more complete, tidier, more accurate 3D human body standard skeletons. Our method can be applied to 3D human body standard skeletons extraction from meshes which are reconstructed by multiview images of moving body or 3D human motion surfaces which are scanned, while requiring without manual intervention. However, because our initial skeleton extraction is based on shape segmentation, so whether our skeleton extraction is ideal or not depends on the normalization and the success of shape segmentation. Despite of this, our 3D human body standard skeletons from continuous frame meshes are more standardized, more effective and are more conducive to be used by subsequent posture estimation, body modeling and operation.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U1811463, 61602486, 61876012, in part by the Beijing Municipal Science and Technology Project under Grant Z171100004417023,in part by Beijing Educational Committee under Grant KM201510005024.

References

[BAS14] BÆRENTZEN J. A., ABDRASHITOV R., SINGH K.: Interactive shape modeling using a skeleton-mesh co-representation. *ACM Transactions on Graphics 33*, 4 (2014), 132. 1

[CTO*10] CAO J., TAGLIASACCHI A., OLSON M., ZHANG H., SU Z.: Point cloud skeletons via laplacian based contraction. In *Shape Modeling International Conference* (2010), IEEE, pp. 187–197. 1, 4

[Fêd03] FÊDOR M.: Application of inverse kinematics for skeleton manipulation in real-time. In *Proceedings of the 19th Spring Conference on Computer Graphics* (2003), ACM, pp. 203–212. 1

[HZ08] HARTLEY R., ZISSERMAN A.: Multiple view geometry in computer vision. *Kybernetes 30*, 9/10 (2008), 1865–1872. 2

[KO19] KLEIMAN Y., OVSJANIKOV M.: Robust structure-based shape correspondence. *Computer Graphics Forum 38*, 1 (2019), 7–20. 2, 3

[LHW*13] LI G., HUI H., WU S., COHEN-OR D., GONG M., HAO Z.: L1-medial skeleton of point cloud. *ACM Transactions on Graphics 32*, 4 (2013), 1–8. 1, 4, 5

[PAG11] PAN H. W., AI C., GAO C. M.: A new approach for body pose recovery. In *International Conference on Virtual Reality Continuum and Its Applications in Industry* (2011), pp. 243–248. 1

[SHRB11] STRAKA M., HAUSWIESNER S., RÜTHER M., BISCHOF H.: Skeletal graph based human pose estimation in real-time. In *The British Machine Vision Conference* (2011), pp. 1884–2020. 1

[SSC03] SINGH V., SILVER D., CORNEA N.: Real-time volume manipulation. In *Proceedings of the 2003 Eurographics/IEEE TVCG Workshop on Volume Graphics* (2003), ACM, pp. 45–51. 1

[STG*97] STORTI D. W., TURKIYYAH G. M., GANTER M. A., LIM C. T., STAL D. M.: Skeleton-based modeling operations on solids. In *Proceedings of the Fourth ACM Symposium on Solid Modeling and Applications* (1997), ACM, pp. 141–154. 1

[TAOH12] TAGLIASACCHI A., ALHASHIM I., OLSON M., HAO Z.: Mean curvature skeletons. *Computer Graphics Forum 31*, 5 (2012), 1735–1744. 1, 4

[TZCO09] TAGLIASACCHI A., ZHANG H., COHEN-OR D.: Curve skeleton extraction from incomplete point cloud. *ACM Transactions on Graphics 28*, 3 (2009), 1–9. 1

[ZSW*18] ZHANG Y., SHEN B., WANG S., KONG D., YIN B.: L0-regularization-based skeleton optimization from consecutive point sets of kinetic human body. *ISPRS Journal of Photogrammetry and Remote Sensing 143* (2018), 124–133. 1, 4, 5