



3D human body skeleton extraction from consecutive surfaces using a spatial–temporal consistency model

Yong Zhang¹ · Fei Tan¹ · Shaofan Wang¹ · Baocai Yin¹

Published online: 18 May 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Current approaches of human body skeleton extraction mainly suffer from following problems: insufficient temporal and spatial continuity, unrobust to background, ambient noise, etc. This paper proposes a three-dimensional human body skeleton extraction method from consecutive meshes. We extract the consistent skeletons from consecutive surfaces based on shape segmentation and skeleton sequences; then, we present a spatiotemporal skeleton optimization model to adjust the skeleton sequences. Experiments on multiview images captured from a light field device demonstrate that our method captures more complete and accurate skeletons compared to state-of-the-art methods.

Keywords Human body skeletons · Consecutive surfaces · Spatio-temporal consistency model

1 Introduction

Extracting three-dimensional (3D) human body skeletons from geometric surfaces is an important research topic in the fields of computer graphics, pattern recognition, and human–computer interaction. It has wide applications in pose estimation [24,28], human body modeling [2,31], and skeleton manipulation [10,30]. In general, skeleton extraction methods can be divided into two categories according to different inputs: point cloud-based methods [13,33,35] and mesh-based methods [7,32]. While many research work are devoted to human body skeleton extraction from static point clouds (e.g., [13]), the results are unsatisfactory due to many factors, e.g., self-occlusion, environmental noise.

Although a few skeletonization technologies from point clouds have been proposed and achieved great successes [7, 13,32], those approaches suffer from issues when applying

to extract body skeletons from consecutive point clouds: the number of body skeletons extracted from different frames may differ; many body skeletons locate with great errors due to the ignorance of prior structure of human body. The first issue prevents those technologies from directly applying to skeleton-based animation as well as 3D human body operation, while the second issue reduces the effect of human–computer interaction. In essence, spatial and temporal coherence of human poses is far from being sufficiently explored.

The recent development of light field acquisition devices provides an opportunity for solving those issues. Compared with traditional multiview devices, light field devices can automatically capture a sequence of multiview images of human body that performs continuous motion based on the lighting and frequency setting, while traditional multiview devices can only collect single-frame multiview image sets under specific actions. Based on a light field device, we propose a spatial–temporal consistency model (STC) for extracting human body skeletons from consecutive point clouds. The key of STC is a spatial–temporal consistency adjustment model, which fine-tunes the location of skeletons by exploiting both inter-frame and intra-frame consistency of skeletons. Compared with traditional skeleton extraction methods, STC is unsupervised, completely automated, and requires no training stages. Experimental results on light field acquisition data indicate that 3D human body skeletons pro-

This work was supported by the National Natural Science Foundation of China under Grants 61772049, 61632006, 61876012, U19B2039, 61906011, the Beijing Natural Science Foundation under Grant 4202003, and the Beijing Municipal Science and Technology Project under Grant Z171100004417023.

✉ Shaofan Wang
wangshaofan@bjut.edu.cn

¹ Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

duced by STC are precise and suitable for applications such as skeleton-based animation.

This paper is organized as follows. Section 2 reviews related work on body extracting skeletons of human body and skeletonization of objects. Section 3 proposes the methodology of STC. Section 4 gives the experimental results. Section 5 concludes this paper.

2 Related work

We review related work on extraction of human body skeletons and object skeletonization based on static surfaces and consecutive surfaces.

2.1 Skeleton extraction from static surfaces

Static surfaces can be divided into static mesh surfaces and static point clouds.

2.1.1 Skeleton extraction from mesh surfaces

Curvature flow methods Tagliasacchi et al. proposed a curve skeletonization method based on the property of area minimization of mean curvature flows (MCF) by assigning the curvature flow with extreme values [33]. Chuang et al. provided an effective flow curve method for mesh evolution, which requires no initialization at the beginning of each step due to the proposed finite-elements hierarchy [6].

Contraction methods Au et al. generated curve skeletonization by contracting the mesh geometry into zero volume skeleton shape using implicit Laplacian smoothing and global position constraints; the skeletonization method can retain the shape and topology relationship of the original mesh [1]. Cao et al. developed a curve skeletonization technology based on local Delaunay triangulation and topology refinement [7]. A contraction operation was proposed to repair collected skeletons from meshes with missing data and applied to surfaces with boundary. Jiang et al. proposed an algorithm to extract curve skeletonization from triangle meshes [15]. In this work, the initial skeletonization map is constructed by copying the connectivity and geometric information of the input mesh, and then, the nodes of the skeletonization map are iteratively generated by using the coupling process of graph contraction and surface clustering.

Mesh decomposition methods Katz et al. proposed a skeleton extraction method by dividing complex objects into simple sub-objects using a hierarchical mesh decomposition algorithm [16]. Li et al. decomposed meshes into semantic segmentation based on the idea of edge contraction and space scanning [20]. Sharf et al. proposed a real-time skeleton

extraction algorithm from both point clouds and polygonal meshes [29]. The algorithm is based on a deformable model evolution process, which captures the volume and shape of objects. The deformable model, which consists of multiple competing fronts, evolves in the interior of objects in a rough-to-fine fashion, tracks the center of these surfaces, and then merges and filters the generated arcs to obtain curve skeletons of objects. Chuang et al. proposed a 3D object shape description based on generalized cylinders [4]. The derived generalized cylinder representation is better than the object shapes based on simple generalized cylinder subclasses.

2.1.2 Skeleton extraction from point clouds

Skeleton extraction from missing data Cao et al. developed a contraction operation for generalized discrete geometry data by local Delaunay triangulation and topology refinement, which handles missing data without explicit surface reconstruction [7]. Tagliasacchi et al. proposed an algorithm to extract skeleton curves from point clouds with a large amount of missing data based on generalized rotational symmetry axis of an oriented point set [33]. Huang et al. developed a L1-media skeleton construction algorithm, which can be directly applied to unoriented raw point scan with significant noise, outliers and large area of missing data, without strong requirements for the quality of input point clouds [13]. Zhang et al. proposed ℓ_0 -regularization-based skeleton optimization method from continuous point set of dynamic human body [35]. By integrating spatiotemporal constraints, the method recovers missing points in the skeletons, corrects the outliers in the skeleton, and maintains the motion characteristics.

Other skeleton extraction methods Liang et al. proposed a framework for skeletonization of point clouds by using a discretization scheme of differential operators and applied to geometric understanding of point clouds [21]. The framework defined a discretized Laplace–Beltrami operator on point clouds, which effectively combines local information with global information. Zhang et al. proposed to reconstruct a skeletonization of trees using an enhanced PyrLK optical flow method [36]. The method circumvents the issue of manual interaction and inaccuracy and facilitates automatic tree modeling by reconstructing a 3D skeletonization model of trees with realistic sense. Mei et al. proposed an incomplete point cloud skeletonization of trees from laser scanning data, by using a hybrid model consisting of an L1 intermediate skeleton algorithm and minimum spanning tree algorithm [22].

In summary, the existing skeleton extraction methods from static surfaces suffer from three shortcomings:

- Imperfect point clouds with missing data cannot be well treated [7,21,33].
- Closed loops and glitches exist in extracted skeletons [6,7,15].
- Incomplete skeletons or skeletons with error branches may be produced [35].

2.2 Skeleton extraction from consecutive surfaces

Topology matching methods Topology matching is particularly useful for interactive retrieval of 3D objects. Hilaga et al. proposed a topology matching technology by comparing multi-resolution Reeb map, calculating the similarity between polyhedral models, establishing the corresponding relationship between various parts of the object, and achieving accurate search of 3D shape data set [12]. Dey et al. defined a medial geodesic function and derived an approximate algorithm of curve skeletons; the generated curve skeletonization together with proper attributes can model different types of real-world objects [9]. Chen et al. proposed a 3D object shape description based on generalized cylinders, which can better approximate the entity [4]. Zimovnov et al. proposed an effective algorithm for calculating the three-dimensional distance transformation of voxels inside visual shells to form the first approximation of curve skeletons [37]. Zheng et al. regarded the curve skeletons of shapes as a global description feature and assumed that the skeleton structure of the captured shape be consistent for a period of time [38]. Other scholars extract skeletons by determining motion postures based on the observation of continuous frames of body movements [5,35].

Animation driven methods Wang et al. proposed a shape correspondence method based on base point driving, which can extract joint object skeleton from 3D mesh shape, and can be applied to skeleton driven animation [34]. Aguiar et al. proposed a robust framework, which automatically extracts motion skeleton and surface skin weight from any mesh animation [8]. Using this framework, deformation mesh sequences can be automatically switched to fully assembled virtual objects completely. Pantuwong et al. proposed an algorithm for automatically generating inverse kinematic skeleton of characters and did not require the input 3D character model which have a certain attitude or direction [25]. Le et al. introduced an example-based assembly method for automatically generating a linear hybrid skin model with skeleton structure [19]. James et al. extended the method of skinned characters to a general setting of skinned deformable mesh animation and provided an algorithm of automatic and progressive skin approximation, which is particularly effective for pseudo-joint motion [14]. Baran et al. proposed an automatic character animation method based on static character meshes and general skeletons, by attaching skele-

tons to the surface of characters for realistic animation [3]. Pang et al. proposed to extract skeletons from animation surfaces by using a global skeleton alignment method, which can spread the key skeletons to the initial skeletons [26].

In summary, the existing skeleton extraction methods from consecutive surfaces suffer from two shortcomings.

- Inter-frame consistency skeletons are difficult to produce.
- Most of methods are unsuitable for the data of point cloud with outliers, noises, and much missing points.

3 Details of STC

We introduce details of STC in this section. Figure 1 shows a flowchart of STC, and Algorithm 1 shows an algorithm of STC. STC mainly consists of four stages, each of which is detailed in the following subsections.

Denote $\|\mathbf{A}\|_F$ to be the Frobenius norm of a matrix \mathbf{A} , and denote $\|\mathbf{v}\|_2$, $\|\mathbf{v}\|_0$ to be the ℓ_2 norm, the ℓ_0 pseudo-norm of a vector \mathbf{v} , respectively. Denote $[\mathbf{A}]_{i,j}$ to be the element of the i th row, j th column of \mathbf{A} , and denote $[\mathbf{A}]_j$ to be the j th column of \mathbf{A} .

3.1 Data preprocessing

The first stage of STC is data preprocessing, which consists of three steps (see Lines 1–3 of Algorithm 1).

Multiview image collection We set different frame rates according to different actions (see Table 1) and collect multiview images of a moving human body of each action using the light field acquisition device (see Fig. 2), which contains 50 industrial cameras with a given frame rate (see Fig. 1a).

Point cloud generation, normalization, and alignment We generate a 3D dense point cloud of human body using patch-based multiview stereopsis (PMVS [11]) and then perform a normalization and alignment scheme to scale to a unit box and move to the origin on the point cloud so that all point clouds of human body of an action sequence share similar sizes, geometric centers and orientations (see Fig. 1b–e).

Triangular mesh reconstruction In order 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, whose color intensity (normal, resp.) is determined by averaging the color intensity (normal, resp.) of multiple points in a box whose size is set as $0.03 \times 0.03 \times 0.03$, and perform Poisson surface reconstruction [17] to obtain a triangular mesh of human body (see Fig. 1f).

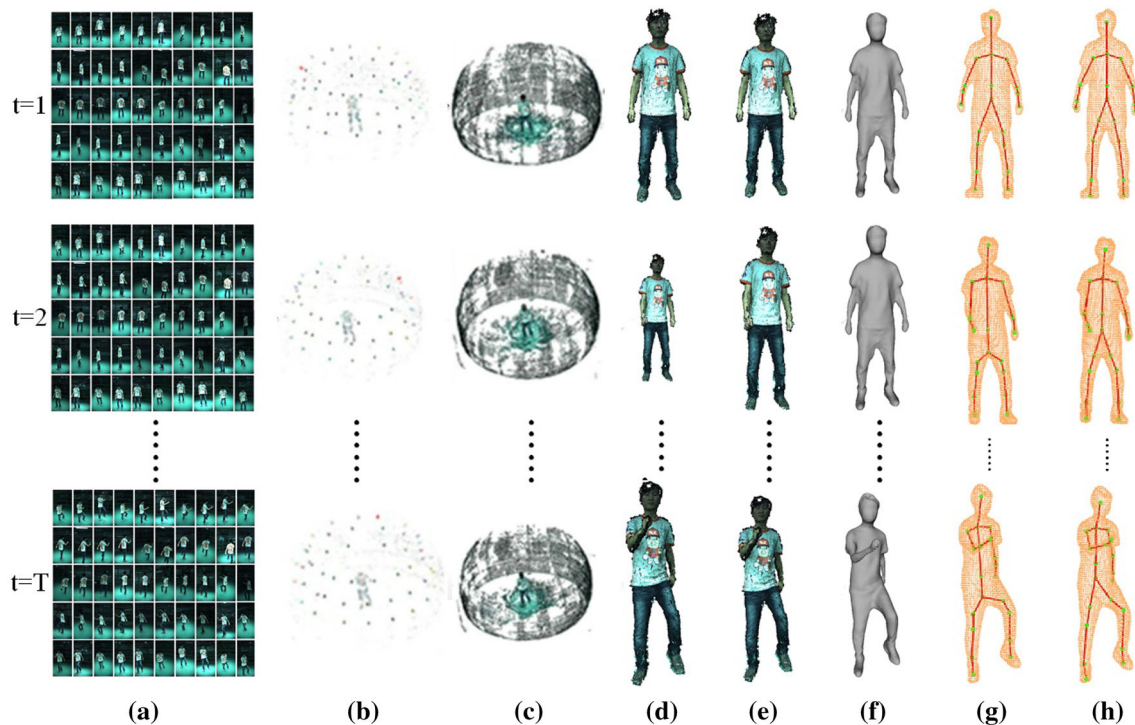


Fig. 1 A flowchart of spatial–temporal consistency model. **a** Multiview image collection; **b** sparse point cloud; **c** dense point cloud of whole scene; **d** dense point cloud of human body; **e** normalization and alignment; **f** Poisson surface reconstruction; **g** initial skeleton extraction; and **h** skeleton adjustment

Table 1 Different frame rates for different actions

| Actions | Frame rates (fps) |
|------------------|-------------------|
| Arm Stretching | 5 |
| Walking | 30 |
| Arms&Legs Moving | 30 |

3.2 Initial skeleton extraction

The second stage of STC extracts initial skeletons from point clouds of each frame individually based on a semantic segmentation of triangular meshes of human body (see Fig. 3). Specially, this stage consists of four steps (see Lines 4–7 of Algorithm 1) which are introduced as follows.

Pseudo-skeleton generation (Fig. 3b, c): We segment the mesh into several semantic patches using [18] (Fig. 3b) and generate “pseudo-skeletons” using the centroid of each patch (Fig. 3c). Those pseudo-skeletons differ from standard human body skeletons in two aspects: pseudo-skeletons always have different numbers with standard skeletons within each body component, and may have incorrect locations compared with standard skeletons. We solve the first issue using the following two steps and leave the second issue until Sect. 3.4.

CShoulder, Waist, and head determination (Fig. 3d): 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.¹ Similarly, Waist is recognized as the unique pseudo-skeleton which achieves degree three.² Then, we determine head as the only point which both achieves degree one and is adjacent to CShoulder.

LShoulder and RShoulder determination (Fig. 3e): After determining CShoulder and Waist skeletons, we observe that the patch corresponding to CShoulder includes LShoulder and RShoulder additionally. To determine their locations, we first select the leftmost adjacent patch (i.e., left upper arm) and rightmost adjacent patch (i.e., right upper arm) of current patch by human body topology connection obtained when model segmentation. Then, we divide points of current patch into three subpatches with equal cardinality according to a distance rule: the first (second, resp.) subpatch is

¹ When more than two pseudo-skeletons achieve the maximum degree simultaneously, we select the pseudo-skeleton with the greatest z coordinate. Our experiments show that the semantic segmentation method of [18] always produces semantic patches with exactly a pseudo-skeleton whose degree achieves the maximum value (four).

² When more than two pseudo-skeletons achieve degree three simultaneously, we select the pseudo-skeleton with the smallest z coordinate.



Fig. 2 A light field acquisition system

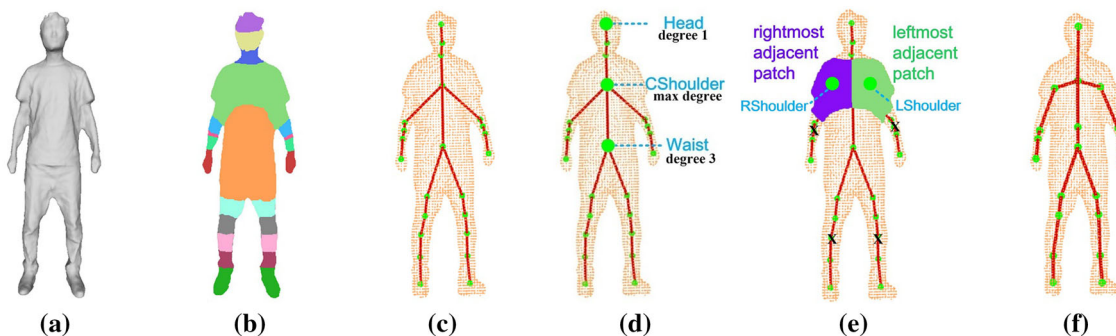


Fig. 3 A flowchart of initial skeleton extraction. **a** Input point cloud; **b** semantic segmentation; **c** pseudo-skeleton generation; **d** Determining CShoulder, Waist, head; **e** determining LShoulder, RShoulder; and **f** standard skeleton completion

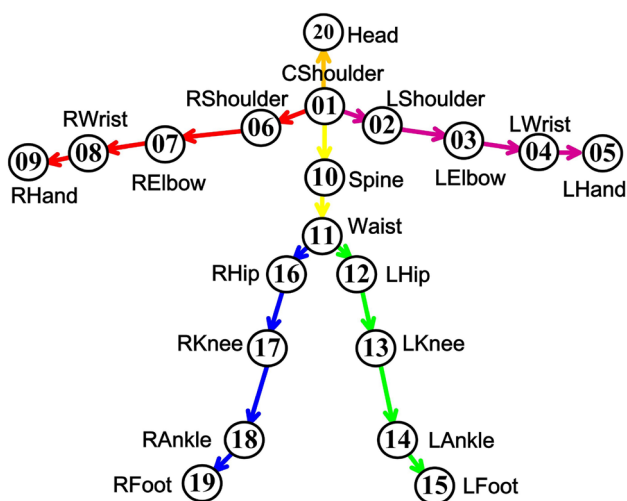


Fig. 4 Standard human model: a directed rooted tree of 20 body skeletons (black spheres) and six body components **Torso**, **Head**, **LArm**, **RArm**, **LLeg**, **RLeg** (colored line segments)

the point set which achieves closest distances to the leftmost (rightmost, resp.) adjacent patch. Finally, we determine LShoulder and RShoulder using the centroid of the first and second subpatches, respectively.

Standard skeleton completion (Fig. 3f): So far we obtain several pre-defined pseudo-skeletons and four skeletons (CShoulder, Waist, LShoulder, RShoulder), whose number may differ from standard skeletons. 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). Note each component corresponds to a number of standard skeletons. If the number of pseudo-skeletons of a component exceeds the number of corresponding standard skeletons, we select the shortest edge among all edges of the component and replace both of two endvertices of that edge with their center; if the number of pseudo-skeletons of a component is less than the number of corresponding standard skeletons, we select the longest edge among all edges of current component and add its center as a new pseudo-skeleton. Either of two tricks is repeated until the number of pseudo-skeletons equals the standard number of current component.

3.3 Skeleton matching

The third stage of STC is to match skeleton points between consecutive frames, i.e., to establish the correspondence between skeletons of different frames so that each skeleton of different frames is correctly matched. We first align CShoulder of all frames so that all CShoulder skeletons share the same coordinate. Next, we find a correspondence between two arms (and two legs) of pairwise adjacent frames, by comparing the total distance from LArm and RArm of the next frame with respect to LArm of the previous frame, i.e., we denote $\mathbf{x}_{t,i} \in \mathbb{R}^3$ to be the 3D coordinates of the i th

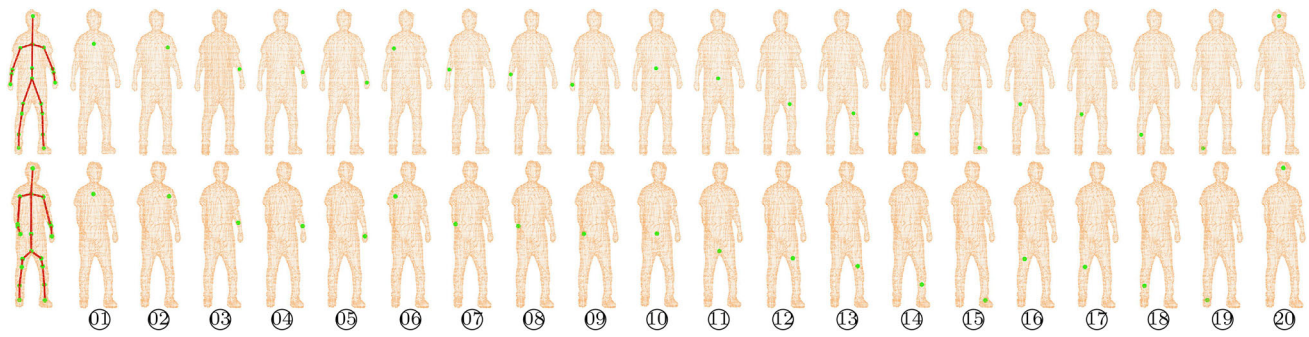


Fig. 5 Effect of inter-frame skeleton matching. Column 1: the 16th frame (top) and 21st frame (bottom) of Walking. Columns 2–21: 01–20 skeletons of the 16th frame (top) and 21st frame (bottom) of Walking (marked by green). All the skeletons of the 21st frame are correctly matched

skeleton of the t th frame; if

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

holds, then the skeletons of the arm of the $(t+1)$ th frame are correctly matched; otherwise we switch the skeletons of two arms of the $(t+1)$ th frame from LArm to RArm. The correspondence between two legs is computed in a similar fashion.

Algorithm 1: Algorithm for STC

input : Multiview image sequence, $\alpha = 0.5$, $\epsilon = 0.02$,
 $\lambda_{\max} = 6000$, $\mu = 2$, $l_{\max} = 300$
output : optimized skeletons $\mathbf{X}^{(k)}$, $k = 1, 2, 3$
initialize: $\lambda = 0.001$
// Data preprocessing
 1 Point cloud generation using PMVS;
 2 Point cloud normalization and alignment;
 3 Triangular mesh reconstruction;
// Initial skeleton extraction
 4 Pseudo-skeleton generation;
 5 Determine CShoulder and Waist;
 6 Determine LShoulder and RShoulder;
 7 Standard skeleton completion;
// Skeleton matching
 8 Align CShoulder of all frames;
 9 Find the correspondence between two arms (and two legs) of pairwise adjacent frames;
// Spatial-temporal consistency adjustment model
 10 **for** $l = 1, \dots, l_{\max}$ **do**
 11 Solve the $\mathbf{X}^{(k)}$ -subproblem of (3) by

$$[\mathbf{X}^{(k)}]_j = \begin{cases} (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T [\mathbf{Y}^{(k)}]_1 & \text{if } j = 1 \\ (\alpha \mathbf{I} + \lambda \mathbf{D}^T \mathbf{D})^{-1} (\alpha \mathbf{z}_{jk} + \lambda \mathbf{D}^T [\mathbf{Y}^{(k)}]_j) & \text{if } j \geq 2 \end{cases}$$

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

 12 Solve the $\mathbf{Y}^{(k)}$ -subproblem of (3) by

$$[\mathbf{Y}^{(k)}]_{ij} = \begin{cases} [\mathbf{D}\mathbf{X}^{(k)}]_{ij} & \text{if } [\mathbf{D}\mathbf{X}^{(k)}]_{ij} \geq \frac{1}{\sqrt{\lambda}} \\ 0 & \text{otherwise} \end{cases}, \quad k = 1, 2, 3;$$

 13 Update penalty parameter: $\lambda \leftarrow \min(\mu\lambda, \lambda_{\max})$;
 14 **end**

Figure 5 shows the effect of the skeleton matching of the 16th frame and the 21st frame of Walking. The left-most subfigure of Fig. 5 denotes the 16th frame (top) and 21st frame (bottom) before matching. The top row represents all skeleton points of the 16th frame, while the bottom row represents all skeleton points of the 21st frame after matching.

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

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

3.4 Skeleton adjustment

The fourth stage of STC adjusts skeletons by using a spatial–temporal consistency adjustment model (see Fig. 1h). Our motivation arises from two observations. For one thing, positions 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 former and latter frames (see Fig. 6 which demonstrates the smooth change of the positions of a skeleton); for another, for each frame, semantic segmentation produced by [18] is imprecise: most non-root skeletons locate far from the corresponding “parent skeletons” defined by Fig. 4 except four ending skeletons (LHand, RHand, LFoot, RFoot) which locate close from 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 [18] cannot distinguish that skeleton from its parent skeleton. Based on the argument above, we propose the spatial–temporal consistency adjustment model (2), which consists of an inter-frame consistency term and an intra-frame consistency term ($\alpha > 0$ is a parameter for balancing two terms). The first term enforces the medium representation of skeletons of almost all frames, with a *median representation matrix* $\mathbf{D} \in \mathbb{R}^{T \times T}$; the second term enforces a framewise fine-tuning over all non-root skeletons for approaching or keeping away from the corresponding parent skeletons, with pre-given parameter $\beta_j, j = 2, \dots, 20$. That parameter determines whether each skeleton approaches or keeps away from its parent skeleton. Specially, for most of skeletons, we set $\beta_j = 1$ to guarantee that each of them approach its parent skeleton; for skeletons ④, ⑤, ⑧, ⑨, ⑭, ⑮, ⑲ which belong to four limbs LArm, RArm, LLeg, RLeg, respectively, as two skeletons locating at the end of each limb are closer to the parent node, we set $\beta_j = -1$ to keep them away from each corresponding parent skeleton.

To solve model (2), we introduce auxiliary matrices $\mathbf{Y}^{(k)} \in \mathbb{R}^{T \times 20}$ for replacing $\mathbf{DX}^{(k)}, k = 1, 2, 3$, and transform (2) into (3)

$$\min_{\{\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}\}_{k=1}^3} \sum_{k=1}^3 \left(\|\mathbf{Y}^{(k)}\|_0 + \alpha \sum_{j=2}^{20} \left\| [\mathbf{X}^{(k)}]_j - (1 - \epsilon) \cdot [\mathbf{X}_{\text{init}}^{(k)}]_j - \beta_j \epsilon \cdot [\mathbf{X}_{\text{init}}^{(k)}]_{\text{parent}(j)} \right\|_2^2 + \lambda \|\mathbf{Y}^{(k)} - \mathbf{DX}^{(k)}\|_F^2 \right). \tag{3}$$

by using naive Lagrange multiplier method, where $\lambda > 0$ is the penalty parameter. We then solve (3) by alternating solving two subproblems of $\mathbf{X}^{(k)}$ and $\mathbf{Y}^{(k)}$ (see Lines 10–13 of Algorithm 1).

4 Experimental results and analysis

In this section, we evaluate the effectiveness of STC by comparing it with state-of-the-art methods. The experiments are conducted on an Intel(R) Core(TM) i5-8250U 1.8 GHz CPU with 8 GB RAM using MATLAB R2016a. We collect multiview color images of three actions by using 50 industrial cameras with 2.2 million pixels through the light field acquisition system (Fig. 2), and the captured images are of

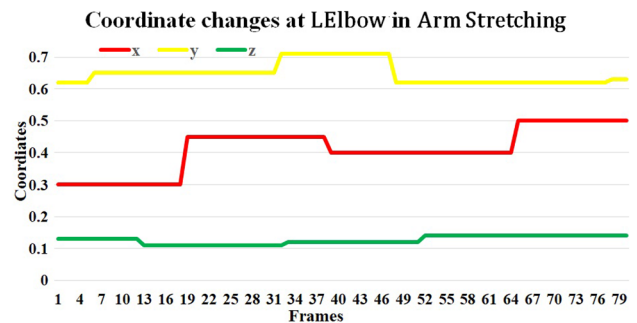


Fig. 6 Coordinate changes at LElbow on the 20th, 33rd, 49th, and 66th frames of Arm Stretching

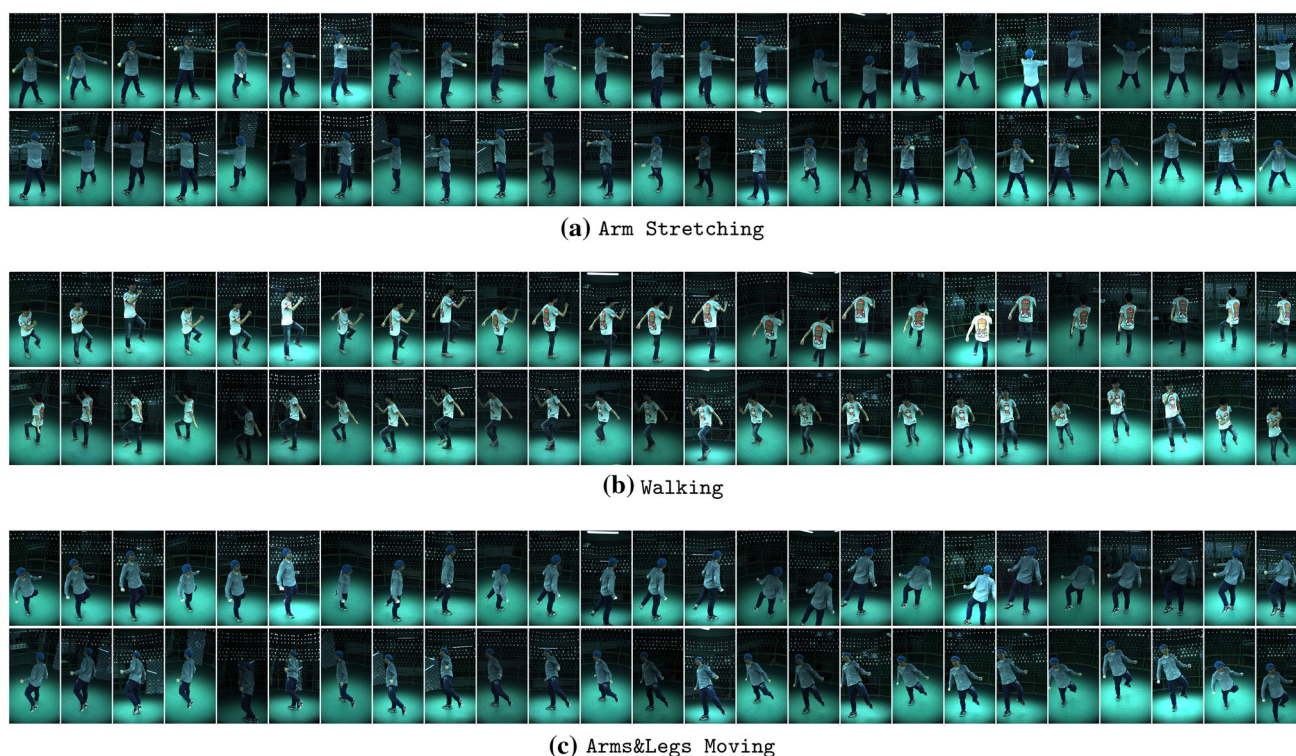


Fig. 7 Multiview images of three actions captured by the light field acquisition system

2048 × 1088 resolution. The three actions are detailed as follows:

- **Arm Stretching:** which consists of 40,000 images collected from 50 different perspectives in 160 s, with frame rate 5 fps (see Fig. 7a).
- **Walking:** which consists of 90,000 images collected from 50 different perspectives in 60 s, with frame rate 30 fps (see Fig. 7b).
- **Arms&Legs Moving:** which consists of 120,000 images collected from 50 different perspectives in 80 s, with frame rate 30 fps (see Fig. 7c).

4.1 Ablation study on skeleton adjustment

While STC consists of three stages, the skeleton adjustment stage plays a key role for the whole STC framework. Thus, we first show comparative results before skeleton adjustment and after skeleton adjustment in Figs. 8, 9, 10, 11, and 12.

According to Fig. 8, the results before skeleton adjustment tend to produce Waist of lower height (see the red boxes), and produce LWrist (RWrist, resp.) which is closer to LEIbrow (REIbrow, resp.) (see the green boxes); moreover, both LLeg and RLeg before skeleton adjustment exhibit abnormal lengths (see the blue boxes). In contrast, results after skeleton adjustment reflect a promising location of body skeletons.

According to Figs. 9 and 10, skeleton adjustment produces a more accurate prediction of LWrist, RWrist, Waist, LAnkle, and RAnkle (the red boxes). Moreover, the length of both LLeg and RLeg is more consistent over those frames after using skeleton adjustment (the blue boxes).

According to Figs. 11 and 12, the results before skeleton adjustment produce the skeletons that are not compatible with the actual skeleton of the human body, such as LHand, RHand, LWrist, RWrist, LFoot, and RFoot (the red boxes). In contrast, the skeletons after adjustment are more accurate (the green boxes).

Overall, by using the skeleton adjustment, STC produces skeletons which are tidier, smoother, and are closer to real body skeletons, and hence achieves good results from different action sequences.

4.2 Comparative results with state-of-the-art methods

We select four state-of-the-art methods for comparative experiment and introduce them as follows:

- Tagliasacchi et al. [32] propose a average curvature skeleton extraction method. The authors formulate the skeletonization problem via MCF. While the classical application of MCF is surface fairing, Tagliasacchi et al. take advantage of its area-minimizing characteristic to

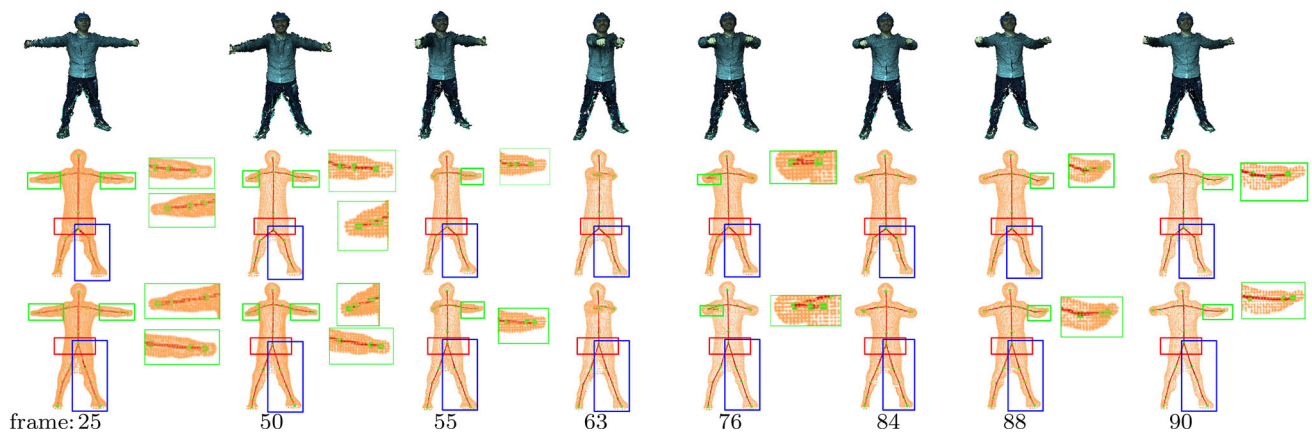
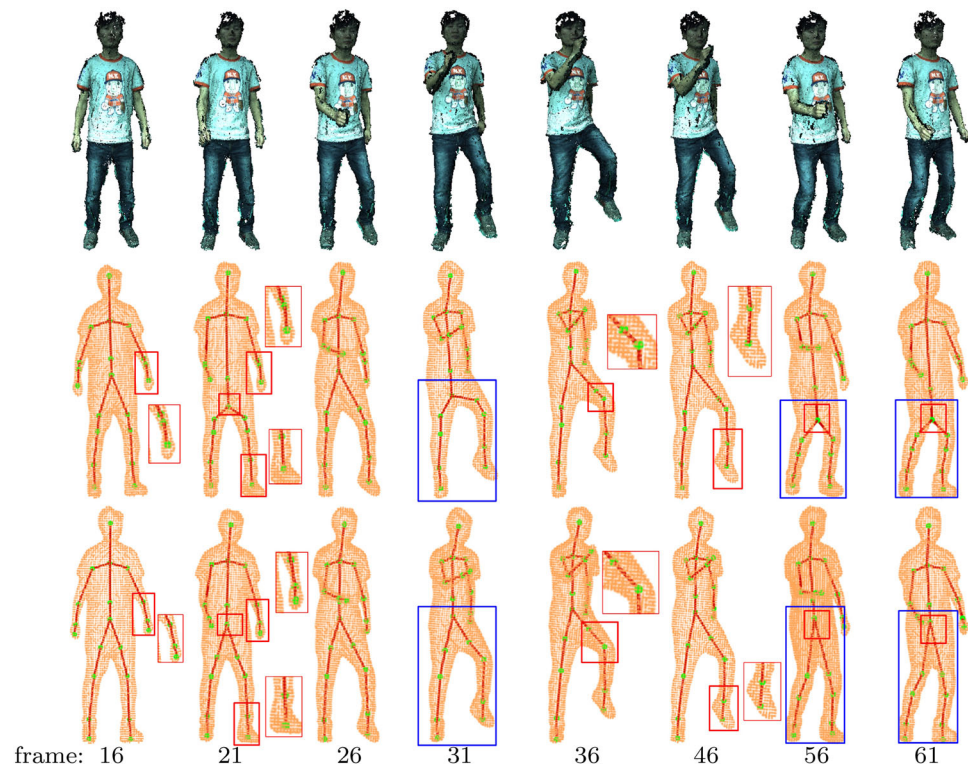


Fig. 8 Top: input data of the 25th, 50th, 55th, 63rd, 76th, 84th, 88th, and 90th frames of Arm Stretching. Middle: results before skeleton adjustment. Bottom: results after skeleton adjustment

Fig. 9 Top: input data of the 16th, 21st, 26th, 31st, 36th, 46th, 56th, and 61st frames of Walking. Middle: results before skeleton adjustment. Bottom: results after skeleton adjustment



drive the curvature flow toward the extreme so as to collapse the input mesh geometry and obtain a skeletal structure.

- Cao et al. [7] propose a Laplacian contraction method. The authors develop a contraction operation that is designed to work on generalized discrete geometry data, particularly point clouds, via local Delaunay triangulation and topological thinning. The method is robust to noise and can handle moderate amounts of missing data, allowing skeleton-based manipulation of point clouds without explicit surface reconstruction.

- Huang et al. [13] propose an ℓ_1 median skeleton extraction method by introducing ℓ_1 -medial skeleton as a curve skeleton representation for 3D point cloud data. They adapted ℓ_1 -medians locally to a point set representing a 3D shape that gives rise to a one-dimensional structure, which can be seen as a localized center of the shape.
- Zhang et al. [35] propose an ℓ_0 -regularization-based skeleton optimization method from consecutive point sets of kinetic human body to extract consecutive skeletons by using the temporal constraint and a spatial constraint.

Fig. 10 Top: input data of the 71st, 86th, 91st, 101st, and 111th frames of Walking. Middle: results before skeleton adjustment. Bottom: results after skeleton adjustment

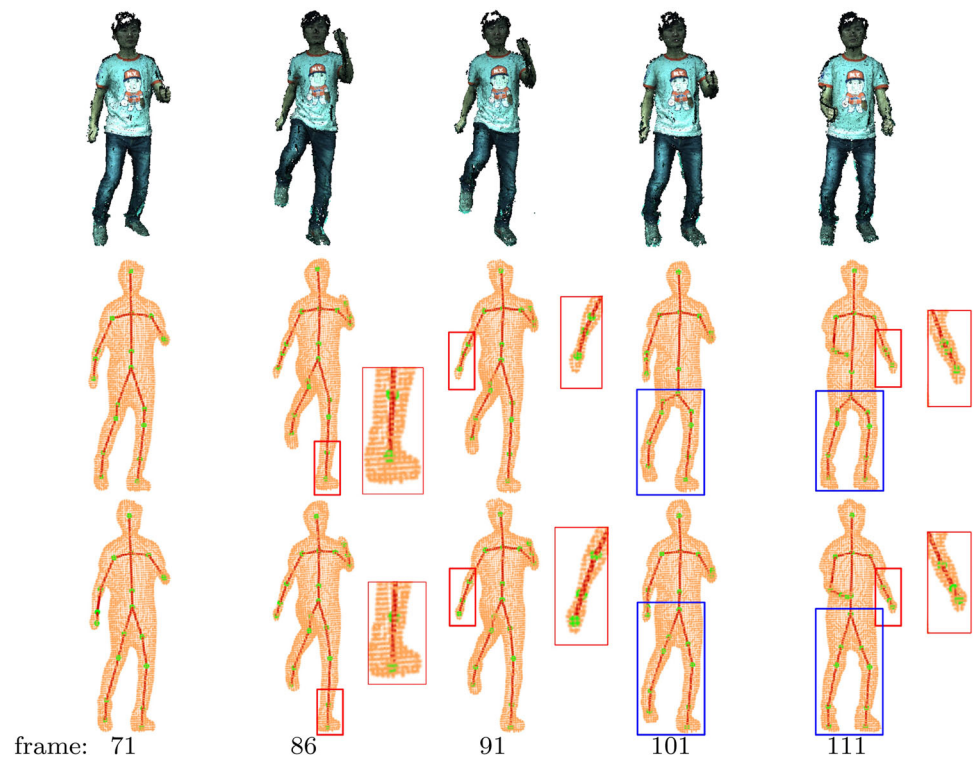
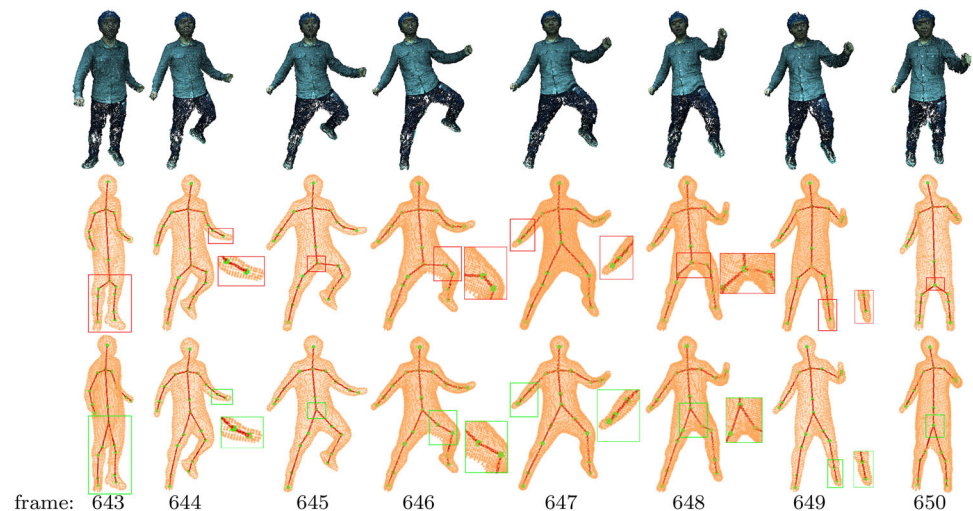


Fig. 11 Top: input data of the 643rd–650th frames of Arms&Legs Moving. Middle: results before skeleton adjustment. Bottom: results after skeleton adjustment



Figures 13, 14, and 15 show qualitative results of STC and four state-of-the-art methods for Arm Stretching, Walking, and Arms&Legs Moving, respectively. We summarize the main shortcomings of comparative methods as follows.

Tagliasacchi et al. [32] always produce incomplete skeletons on arms obviously (see the 61st, 96th, frames of Fig. 14, the 643rd, 644th skeletons of Fig. 15, marked by red boxes), as well as inconsistent connection points of four limbs and Torso (see the 50th, 55th, 63rd frames of Fig. 13, the 36th, 46th, 61st, 91st, 86th frames of Fig. 14, and the 645th, 647th,

648th, 649th, 650th, and 651st skeletons of Fig. 15, marked by blue boxes).

Cao et al. [7] suffer from missing of skeletons, especially on LArm-Torso junction, RArm-Torso junction, LLeg-Torso junction, RLeg-Torso junction (see the 25th, 50th, 55th, 63rd, 76th, 84th, 88th, and 90th frames of Fig. 13, the 36th, 46th, 56th, 61st, 91st, and 86th frames of Fig. 14 and the 643rd–649th frames of Fig. 15, marked by blue boxes) and great prediction errors on LArm-Torso junction, RArm-Torso junction (see the 61st frames of Fig. 14), LKnee-Torso junction (see the 63rd frame of Fig. 13), the RArm skeleton (see the

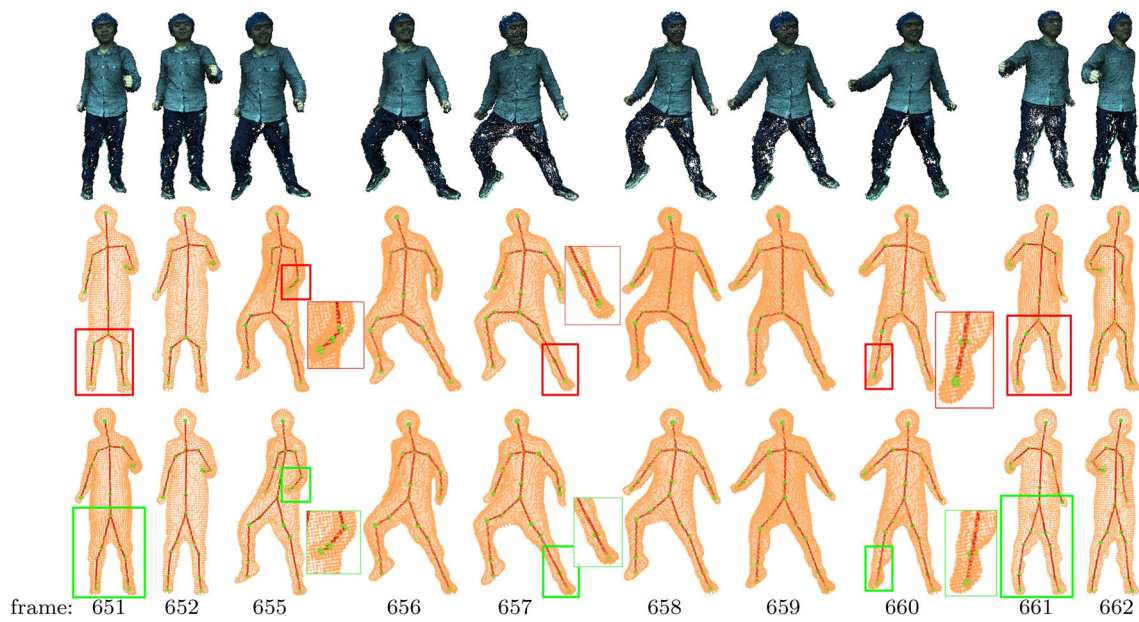


Fig. 12 Top: input data of the 651st, 652nd, 655th, 656th, 657th, 658th, 659th, 660th, 661st, and 662nd frames of Arms&Legs Moving. Middle: results before skeleton adjustment. Bottom: results after skeleton adjustment

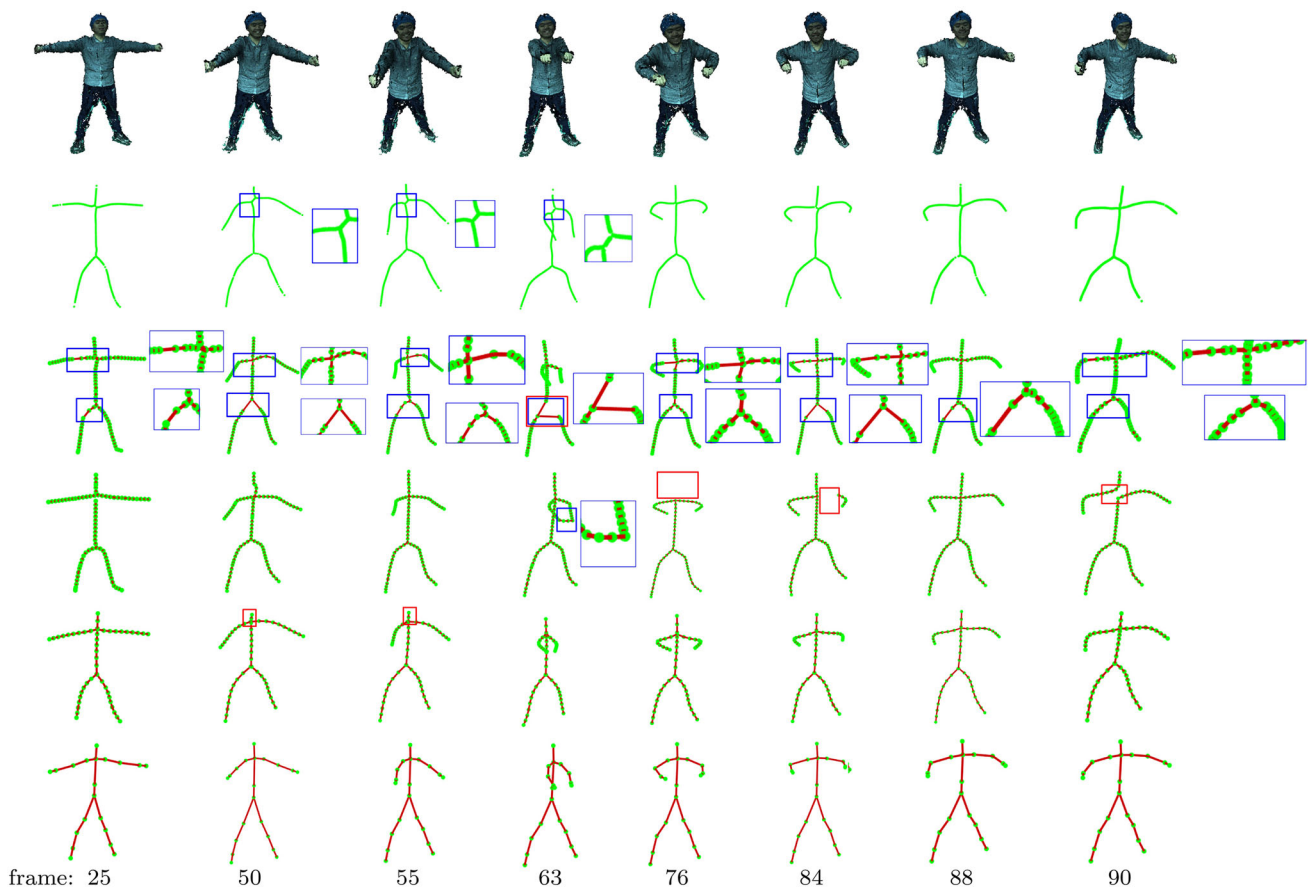
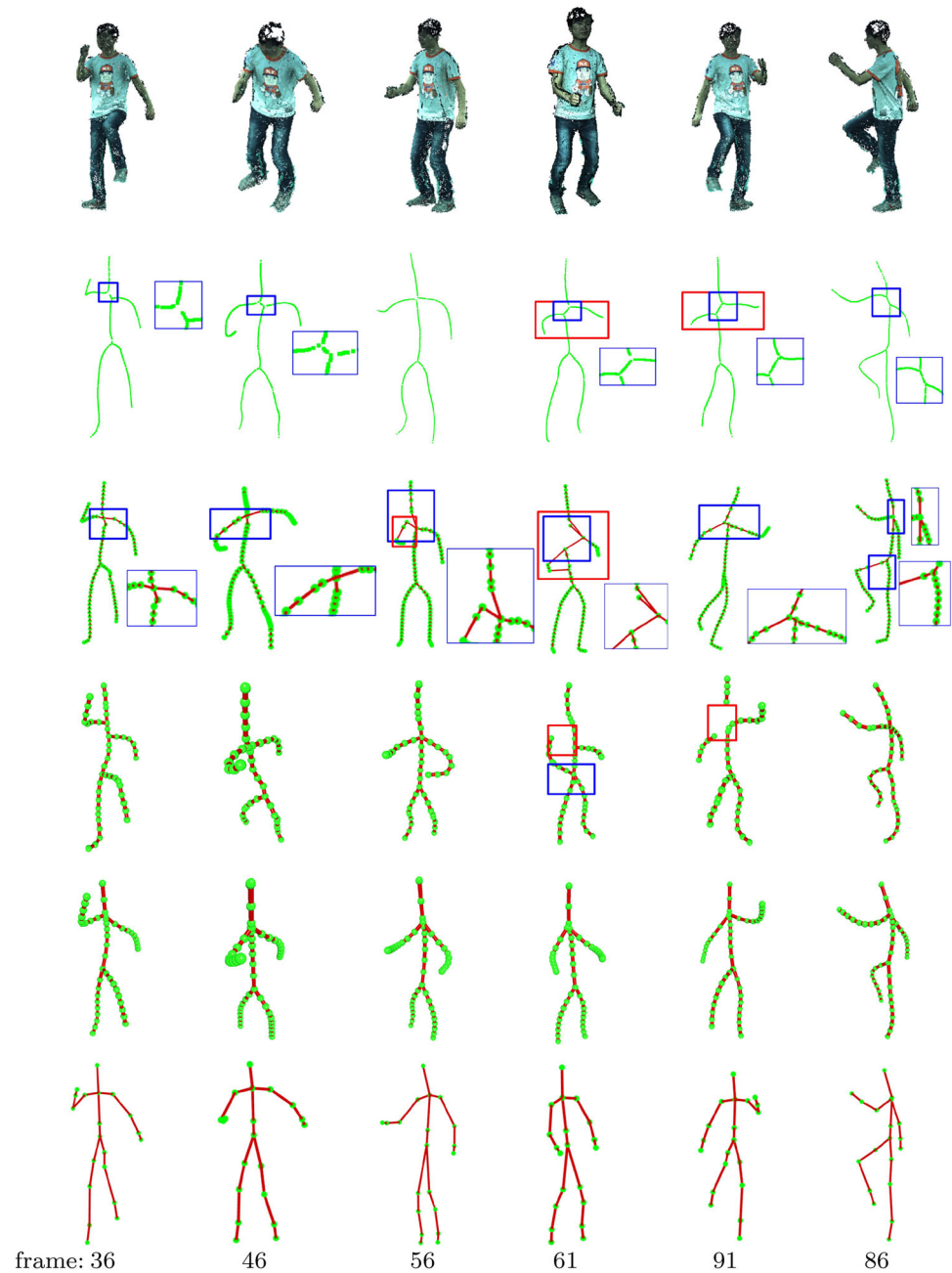


Fig. 13 Qualitative results of Tagliasacchi et al. [32] (row 2), Cao et al. [7] (row 3), Huang et al. [13] (row 4), Zhang et al. [35] (row 5), and STC (row 6) of 25th, 50th, 55th, 63rd, 76th, 84th, 88th, and 90th frames of Arm Stretching

Fig. 14 Qualitative results of Tagliasacchi et al. [32] (row 2), Cao et al. [7] (row 3), Huang et al. [13] (row 4), Zhang et al. [35] (row 5), and STC (row 6) of 36th, 46th, 56th, 61st, 91st and 86th frames of Walking



56th frame of Fig. 14), and RLeg-Torso junction (see the 63rd frame of Fig. 13 marked by red boxes). The skeletons are incomplete on LArm of the 643rd frame of Fig. 15 (marked by red boxes).

Huang et al. [13] suffer from obvious problems such as missing of skeleton points (see the 76th, 84th, and 90th frames of Fig. 13, the 61st, 91st frames of Fig. 14, the 643rd, 645–648th, 650th, and 651st frames of Fig. 15 marked by red boxes), missing of branches (the 76th of Fig. 13, the 645–648th frames of Fig. 15), incorrectness of connection between branches (the 63rd of Fig. 13, the 61st frame of Fig. 14 marked by blue boxes, and the 643rd, 645th frames

of Fig. 15 marked by blue boxes), as well as incorrectness of branches (see the 651st frames of Fig. 15 marked by orange boxes).

Zhang et al. [35] occasionally produce incomplete skeletons on the head (the 50th, 55th frames of Fig. 13 marked by red boxes), incomplete skeletons on two arms (see the 644th, 649th, 650th, and 651st frames of Fig. 15), incorrect branches (see the legs of the 646th, 647th frames of Fig. 15 marked by blue boxes), and the skeletons are not an accurate representation of the human body. In particular, for the 46th, 56th, 61st frames of Fig. 14, Zhang et al. [35] produce bent left arms (which should be straight) and incorrect propor-



Fig. 15 Qualitative results of Tagliasacchi et al. [32] (row 2), Cao et al. [7] (row 3), Huang et al. [13] (row 4), Zhang et al. [35] (row 5), and STC (row 6) of 643rd–651st frames of Arms&Legs Moving

tion of human body (i.e., shorter legs); for the 643rd, 644th, 645th frames skeletons of Fig. 15, Zhang et al. [35] produce very small difference between extracted skeletons exist in the results of Zhang et al. [35], while those frames of the human pose produce great different movement.

In contrast, STC produces more accurate skeletons generally, without the appearance of incorrect 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 makes the final standard skeletons more accurate, more tidy, and more conformable to the original input surfaces, more in line with the actual human body skeleton points distribution. Therefore, STC is better than many traditional skeleton extraction methods and is more convenient for subsequent posture estimation, human body modeling and operation.

4.3 Analysis and discussion

It should be noted that STC has many shortcomings and requires improvement. First, the effect of skeleton extraction of STC heavily depends on the effect of mesh segmentation. Failure of model segmentation may occur when point cloud is seriously missing. Segmentation errors occur when body parts are in contact. As a result, the extracted skeleton is inaccurate. Secondly, compared with deep learning-based methods [23,27], STC cannot treat singular poses or poses with sudden changes due to the lack of training set. This issue may be circumvented by exploiting motion principles or detecting anchor landmarks (e.g., head) of special actions.

5 Conclusion

We propose a 3D human body standard skeleton extraction method from consecutive surfaces, by using a spatiotemporal consistency model. Our model can be applied to 3D human body standard skeletons extraction from meshes which are

reconstructed from either multiview images of moving body or 3D scanned human motion surfaces, without requiring manual intervention. The model produces more complete, tidier, more accurate 3D human body standard skeletons, and facilitates subsequent posture estimation, human modeling and operation.

In the future work, we shall consider generalizing our method to (semi-)supervised fashion, so that singular poses can be inferred. Moreover, we shall consider action specific periodicity estimation for improving skeleton extraction accuracy.

Compliance with ethical standards

Conflict of interest The authors declare no conflicts of interest regarding the publication of this manuscript.

References

- Au, K.C., Tai, C.L., Chu, H.K., Cohen-Or, D., Lee, T.Y.: Skeleton extraction by mesh contraction. *ACM Trans. Graph.* **27**(3), 1–10 (2008)
- Bærentzen, J.A., Abdrashitov, R., Singh, K.: Interactive shape modeling using a skeleton-mesh co-representation. *ACM Trans. Graph.* **33**(4), 1–10 (2014)
- Baran, I., Popović, J.: Automatic rigging and animation of 3D characters. *ACM Trans. Graph.* **26**, (3), article no. 72 (2007)
- Chuang, J.H., Ahuja, N., Lin, C.C., Tsai, C.H., Chen, C.H.: A potential-based generalized cylinder representation. *Comput. Graph.* **28**(6), 907–918 (2004)
- Chun, C., Jenkins, O.C., Mataric, M.J.: Markerless kinematic model and motion capture from volume sequences. *Comput. Vis. Pattern Recognit.* **2**, 475–482 (2003)
- Chuang, M., Kazhdan, M.M.: Fast mean curvature flow via finite elements tracking. *Comput. Graph. Forum* **30**(6), 1750–1760 (2011)
- Cao, J., Tagliasacchi, A., Olson, M., Zhang, H., Su Z.: Point cloud skeletons via Laplacian based contraction. In: 2010 Shape Modeling International Conference. IEEE, pp. 187–197 (2010)
- De Aguiar, E., Theobalt, C., Thrun, S., Seidel, H.P.: Automatic conversion of mesh animations into skeleton-based animations. *Comput. Graph. Forum* **27**(2), 389–397 (2008)
- Dey, T.K., Sun, J.: Defining and computing curve-skeletons with medial geodesic function. *Symp. Geom. Process.* **6**, 143–152 (2006)
- Fêdor, M.: Application of inverse kinematics for skeleton manipulation in real-time. In: Proceedings of the 19th Spring Conference on Computer Graphics. ACM, pp. 203–212 (2003)
- Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(8), 1362–1376 (2010)
- Hilaga, M., Shinagawa, Y., Komura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3D shapes. *ACM SIGGRAPH*, pp. 203–212 (2001)
- Huang, H., Wu, S., Cohen-Or, D., Gong, M., Zhang, H., Li, G., Chen, B.: L1-medial skeleton of point cloud. *ACM Trans. Graph.* **32**(4), 1–8 (2013)
- James, D.L., Twigg, C.D.: Skinning mesh animations. *ACM Trans. Graph.* **24**, 399–407 (2005)
- Jiang, W., Xu, K., Cheng, Z., Martin, R.R., Dang, G.: Curve skeleton extraction by coupled graph contraction and surface clustering. *Graph. Models* **75**(3), 137–148 (2013)
- Katz, S., Tal, A.: Hierarchical mesh decomposition using fuzzy clustering and cuts. *ACM Trans. Graph.* **22**(3), 954–961 (2003)
- Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Symposium on Geometry Processing, pp. 61–70 (2006)
- Kleiman, Y., Ovsjanikov, M.: Robust structure-based shape correspondence. *Comput. Graph. Forum* **38**, 7–20 (2019)
- Le, B.H., Deng, Z.: Robust and accurate skeletal rigging from mesh sequences. *ACM Trans. Graph.* **33**(4), 1–10 (2014)
- Li, X., Woon, T.W., Tan, T.S., Huang, Z.: Decomposing polygon meshes for interactive applications. *Interactive 3D Graphics and Games*, pp. 35–42 (2001)
- Liang, J., Lai, R., Wong, T.W., Zhao, H.: Geometric understanding of point clouds using Laplace–Beltrami operator. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 214–221 (2012)
- Mei, J., Zhang, L., Wu, S., Zhen, W., Liang, Z.: 3D tree modeling from incomplete point clouds via optimization and L_1 -MST. *Int. J. Geogr. Inf. Sci.* **31**(5), 999–1021 (2017)
- Mehrizi, R., Peng, X., Tang, Z., Xu, X., Metaxas, D., Li, K.: Toward marker-free 3D pose estimation in lifting: a deep multi-view solution. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 485–491 (2018)
- 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, pp. 243–248 (2011)
- Pantuwong, N., Sugimoto, M.: A novel template-based automatic rigging algorithm for articulated-character animation. *Comput. Anim. Virtual Worlds* **23**(2), 125–141 (2012)
- Pang, Z., Yong, Z., Xiao, C.: Effective skeletons extraction for animated surfaces based on geometry propagation. *Comput. Anim. Virtual Worlds* **26**(3–4), 301–309 (2015)
- Schwarz, S., Pollard, T.: 3D human pose estimation from deep multi-view 2D pose. In: 24th International Conference on Pattern Recognition, pp. 2326–2331 (2018)
- Straka, M., Hauswiesner, S., Rüther, M., Bischof, H.: Skeletal graph based human pose estimation in real-time. In: Proceedings of the British Machine Vision Conference, pp. 1–12 (2011)
- Sharf, A., Lewiner, T., Shamir, A., Kobbelt, L.: On-the-fly curve-skeleton computation for 3D shapes. *Comput. Graph. Forum* **26**(3), 323–328 (2010)
- Singh, V., Silver, D., Cornea, N.: Real-time volume manipulation. In: Proceedings of the 2003 Eurographics/IEEE TVCG Workshop on Volume Graphics. ACM, pp. 45–51 (2003)
- 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*. ACM, pp. 141–154 (1997)
- Tagliasacchi, A., Alhashim, I., Olson, M., Hao, Z.: Mean curvature skeletons. *Comput. Graph. Forum* **31**(5), 1735–1744 (2012)
- Tagliasacchi, A., Zhang, H., Cohen-Or, D.: Curve skeleton extraction from incomplete point cloud. *ACM Trans. Graph.* **28**(3), 1–9 (2009)
- Wang, K., Razzaq, A., Wu, Z., Feng, T., Ali, S., Jia, T., Wang, X., Zhou, M.: Novel correspondence-based approach for consistent human skeleton extraction. *Multimed. Tools Appl.* **75**(19), 1–22 (2015)
- Zhang, Y., Shen, B., Wang, S., Kong, D., Yin, B.: L0-regularization-based skeleton optimization from consecutive point sets of kinetic human body. *ISPRS J. Photogramm. Remote Sens.* **143**, 124–133 (2018)
- Zhang, D., Liang, S., Zhang, C., Jia, J.: 3D tree skeleton reconstruction based on enhanced PyrLK optical flow algorithm. *J. Comput. Aid. Des. Comput. Graph.* **27**(7), 1247–1254 (2015)

37. Zimovnov, A., Mestetskiy, L.: Curve-skeleton extraction from visual hull. In: International Conference on Computer Vision Theory and Applications, pp. 666–671 (2015)
38. Zheng, Q., Sharf, A., Tagliasacchi, A., Chen, B., Zhang, H., Sheffer, A., Cohen-Or, D.: Consensus skeleton for non-rigid space-time registration. *Comput. Graph. Forum* **29**(2), 635–644 (2010)

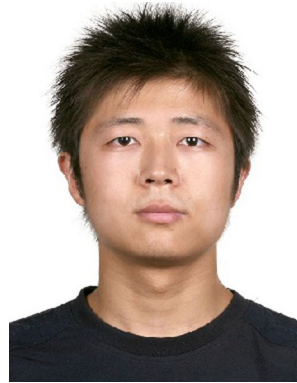
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Yong Zhang received the Ph.D. degree in computer science and technology from Beijing University of Technology, China, in 2010. He is currently an Associate Professor with Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing, China. His research interests include intelligent transportation system, big data analysis and visualization, computer graphics.



Fei Tan received the B.S. degree in computer science and technology from Qinghai University, China, in 2016 and received the M.S. degree in computer science and technology from Beijing University of Technology, China, in 2019. She is currently working in Jiuquan Satellite Launch Center, China.



Shaofan Wang received the B.S. and Ph.D. degrees in computational mathematics from Dalian University of Technology, Dalian, China, in 2003 and 2010, respectively. He is an Associate Professor from Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing, China. His research interest includes pattern recognition and machine learning.



Baocai Yin received the M.S. and Ph.D. degrees in computational mathematics from Dalian University of Technology, Dalian, China, in 1988 and 1993, respectively. He is currently a Professor of the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. His research interests include multimedia, image processing, computer vision, and pattern recognition.