Pattern Recognition 77 (2018) 216–225

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A hybrid framework for automatic joint detection of human poses in depth frames

Longbo Kong[b], Xiaohui Yuan[a,b,*], Amar Man Maharjan[b]

[a] College of Information Engineering, China University of Geosciences, Wuhan, 430074, China
[b] Department of Computer Science and Engineering, University of North Texas, Denton, TX 76210, USA

ARTICLE INFO

Article history:
Received 16 December 2016
Revised 13 December 2017
Accepted 30 December 2017
Available online 3 January 2018

Keywords:
Human pose detection
Joint detection
Human body model
Geodesic features

ABSTRACT

Human pose detection has been an active research topic, and many studies have been done to address different problems in the topic. However, very few methods are proposed to detect joints in the human body. In this paper, we proposed a novel hybrid framework to detect joints automatically by using depth camera. In the proposed method, joints are categorized into two classes: implicit joints and dominant joints. Implicit joints are the joints on the torso, such as neck and shoulders. Dominant joints include elbows and knees. In the hybrid framework we proposed, a loose skeleton model is used to locate implicit joints, and data-driven method is applied to detect dominant joints. The highlight of the proposed work is that geodesic features of the human body are used to build the skeleton model and detect joints. To evaluate our work, experiments are conducted on the dataset recorded by a Microsoft Kinect and compared with state-of-art methods. The results demonstrate that the proposed work can deliver stable and accurate detection results of joints.

1. Introduction

Human pose detection and tracking have been widely used in many applications, such as human-computer interaction, robot control, 3D animation creation, and home entertainment. In such video-based detection and tracking system, the feature point of different body parts and joints such as head, elbows, and hands, provides the information of poses and activities of people. Many methods [1–4] have been proposed to tracking feature points from RGB videos that can be used for human pose detection and tracking. When detecting human motions, joints provide accurate information about the poses. Joint detection provides vital information for characterizing human pose and serves as a foundation for a wide range of computer vision applications such as physical training and pose analysis, healthcare, entertainment, etc [5–9]. For instance, knowing the precise location of human joints enables estimation of poses and movements, which facilitates personalized training for applications in rehabilitation and combat tactics instruction. Despite recent developments in markerless human tracking, few methods, to our best knowledge, have been proposed for accurately detecting human joints. Models are usually used for identifying joints in point cloud by matching a point cloud to a model and inheriting the pre-marked joints. However, the accuracy of the detected joints is suboptimal due to the misalignment, which affects the precision of tracking human movements. Alternatively, methods that classify point cloud into body parts have been proposed, which leverage a database of key poses and are robust to the deformation from clothes. Yet, the variety of key poses greatly affects the performance. To overcome the aforementioned problems, we proposed a method that integrates data-driven and model-based strategies to improve the accuracy of joint detection. Accurate detection of joints is critical for motion and activity analysis of a human and vital for delivering accurate results of detecting and tracking human motions.

There are various methods proposed to detect feature points (end points) from a different type of videos, e.g. infrared video and depth video (time-of-flight video). Comparing with visual (RGB) and infrared videos, depth video shows a great advantage, which provides distance information that is important to overcome the confusion of body parts and occlusions. Depth video is the projection of the point cloud captured by depth camera (such as Microsoft Kinect). Each point carries the distance between the target and the camera. With the intrinsic parameters of the depth camera, position in 3D space of each point in point cloud is acquired. There are two major strategies for detecting human poses using depth videos, model-based and data-based strategies. The idea of model-based strategy is to fit a predefined 3D human model to the point cloud acquired by the depth camera. Data based strategy

* Corresponding author at: Department of Computer Science and Engineering, University of North Texas, Denton, TX 76210, USA.
E-mail address: xiaohui.yuan@unt.edu (X. Yuan).

https://doi.org/10.1016/j.patcog.2017.12.020
0031-3203/

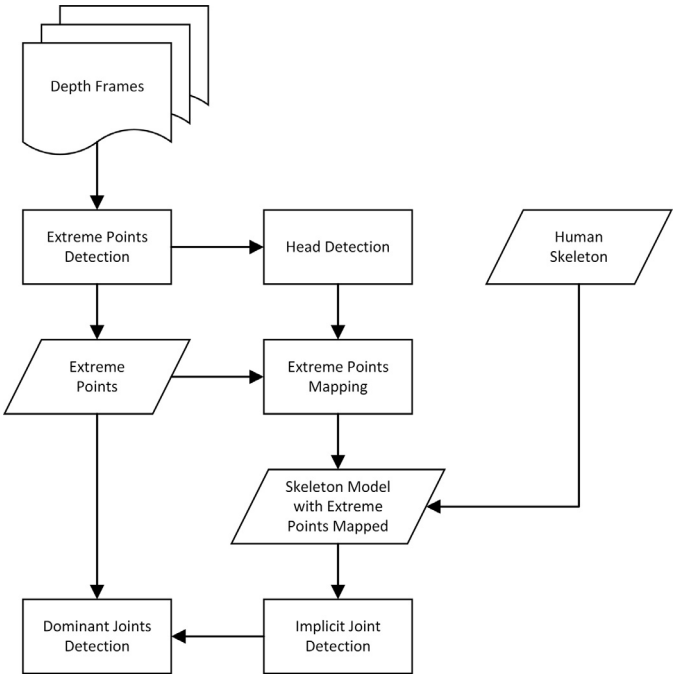


Fig. 1. Overview of the workflow of the proposed method.

uses machine learning algorithm to classify points to body parts or detect feature points using based on different features. However, both strategies face challenges in accurate detection of the joints of human body.

The objective of our work is to accurately detect joints on the human body by combining body model and automatic feature point detection. Fig. 1 represents the workflow of the proposed framework. The human body model maps the detected extreme points to the corresponding body parts on the model. The skeleton model detects the position of implicit joints. The dominant joints are detected after implicit joints and extreme points are located by a shortest path based methods. The main contribution of this work is a hybrid framework to detect joints on the human body to achieve robustness to different body shapes or proportions, pose variations and occlusions. Another contribution of this work is the idea of using geodesic features of the human body to build a model to guide the detection of human pose estimation.

The rest of this article is organized as follows: Section 2 reviews the related work on human pose detection. Section 3 presents the hybrid framework for detecting joints on the human body. Section 4 discusses and compares the experimental results. Section 5 concludes this paper with future work.

2. Related work

Markerless human body pose detection is an active research area in computer vision [10] and many studies have been conducted that employ optical and depth videos (see [11] and references therein). Feature points detection, 3D human body model fitting or alignment, and pixel classification are the most widely used strategies. Methods that detect feature points usually use the silhouette of the human body as a graph and apply shortest distance methods such as Dijkstra's algorithm to locate feature points. The detected feature points are used to represent poses of the human body. Plagemann et al. [12] used Dijkstra's algorithm to extract feature points from the human body and calculated the orientation of each feature point by using backtracking method. Then, the author applied a body part identification method to classify detected feature points to different body parts. It extracted the shape of the area around each detected feature point and compared with shape features of different body parts from the trained data set. However, joints remained undetected in [12]. Baak et al. [13] applied a modified Dijkstra's algorithm to detect feature points. In [13], the detected feature points were used as query input to search possible poses from the database. However, the feature points to this type of method include no joints. Because Dijkstra' algorithm is used to locate the extreme points (end points of body parts).

Many methods [5,6,14–19] using human body model have been developed to detect human poses. Zhang et al. [18] proposed an approach based on the data-driven Markov Chain Monte Carlo (DDMCMC) framework for human pose estimation. A three-level tree structure was used as a human body model and the pose estimation is formulated as a Bayesian inference problem. The tree-structure state space was parsed into a lexicographic order and searched by the DDMCMC technique. Cui et al. [19] integrated low- and high-dimensional tracking approaches into a framework using a probabilistic fusion formulation. The low-dimensional approach uses motion models, whereas the high-dimensional approach tracks movements by sampling the pose space. Zuffi et al. [7] proposed a method that used realistic and part-based 3D human models for human pose detection. The part-based 3D human model allows each body part in the model rotating and transforming independently to fit the data. A cost function is defined to calculate how smoothly two adjacent parts can be connected. Ye et al. [8] used a set of pre-captured motion examples for initial pose estimation and the acquired data is matched to these examples for initial pose estimation and semantic labeling. Most of the model-based methods define body parts only and use the intersections of body parts as joints. Recently, Sigalas et al. [6] proposed the top view re-projection method to align body model to the point cloud of the human body. In [6], the author re-projected the points inside the cylinder body part model to the top surface of the cylinder. The ratio of the number of re-projected points to the total number of points inside the cylinder model is computed as a re-projection ratio. The best hypothesis position of a body model is determined by selecting the minimum top view score (TVR) which includes the re-projection ratio as the key factor, along with other adjusting factors. The joints were defined on fixed positions of each cylinder body part model. Handrich et al. [15] proposed a hybrid method to detect human pose. In their work, geodesic distances were used to detect feature points of hands and elbows, then individual models were applied to detect feature points of head and shoulders. A skeleton model was applied to reject incorrect detection results. Schwarz et al. [16] applied Dijkstra's algorithm to detect end points of limbs as primary feature points and fit a skeleton model to the primary feature points. Joints were defined by the skeleton model, and the positions of joints were estimated by the model fitting procedure. Shen et al. [20] adopted a hierarchical method that identifies head and torso by template fitting followed by detecting shoulders and hips. A multi-cue fusion method is developed to fit a skeleton model, which extracts 2D cues from synchronized multiview images and integrated them into a 3D representation. A volumetric reconstruction method was implemented to merge relevant information into a coherent structure. Zhu et al. [5] proposed a template-based method to detect body parts, where joints were defined as the intersections of two connected body part templates. Ganapathi et al. [21] applied a model based local optimizing method to match the human body model to the point cloud, and a data-driven method to detect body parts which were used to initialize the local optimizing method. It is clear that model-based or model involved methods usually define their human models by defining the body parts and degrees of freedom of corresponding parts, or kinematic relationship between connected body parts. However, these methods suffer from

different body proportions, cumulative deviation of model fitting and local maxima during the model fitting.

Similar to the model-based methods, most learning based methods also focus on large body parts detection. Because comparing to joints, body parts have a larger area of the human body. The Larger area of the human body means more points on the point cloud to learn and classify, therefore, results that are more accurate can be generated. In learning based methods, joints are also usually defined as the intersections of connected body parts. Wei et al. [22] applied classification method in their system to detect the initial pose and registered the human skeleton model to the depth frame. After the skeleton model was initialized, a tracking method was invoked to track 3D poses via a Maximum A Posteriori (MAP) framework. The joints were defined by the skeleton model and initialized by the classification method at the beginning. Shotton et al. [9] proposed two methods to estimate human poses, body part classification (BPC) and offset joint regression (OJR). Both methods estimated the positions of joints, and cast votes for the position of joints. Buys et al. [23] used an underlying kinematic model and a pixel classification method to extract the skeleton information from the depth frames. The kinematic model was used to constrain the extracted skeleton components and generate the final skeleton. In learning based methods, body parts can be classified due to training procedure on large scale of the dataset. However, the noisy or incomplete dataset could cause incorrect classification results. The boundary of different body parts sometimes can be ambiguous which can cause unstable or incorrect skeleton extraction results. Most recently, Nishi and Miura [24] proposed a method to generate datasets of human depth images with body part labels to enhance learning-based pose recognition, which complements the existing datasets by providing many unusual poses such as lying and crouching. Twelve-class labels were proposed, which include eleven body parts and the background.

By reviewing the all three different strategies, methods that lack accurate joint detection could suffer ambiguous, unstable or even incorrect detection results. Accurate detection of joints remains an open question. To overcome this issue, we proposed a hybrid framework that combines human body model and geodesic features of human body together to detect and estimate the position of joints.

3. Method

3.1. Overview

The proposed method categorizes joints into two types, implicit joints, and dominant joints. Implicit joints include the joints that are close to the torso. In the proposed method, neck, left and right shoulders, left and right hips and waist are defined as implicit joints. Dominant joints include the joints on the limbs of a human body. Left and right elbows and knees are categorized as dominant joints. In practice, the dominant joints are easier to detect than the implicit joins due to the rigidity of the bone segments of the limbs of the human body. Implicit joints are difficult to detect. Because implicit joints are part of the torso, and the deformation of these joints are less significant than that of the dominant joints. On the other hand, dominant joints carry more information about human motion than the implicit joints. As the connections between torso and limbs of the human body, it is still necessary to detect the position of the implicit joints.

The proposed method employs a skeleton model of human body. Skeleton model defines the geodesic features of implicit joints. Extreme points (feature points on the tip of body parts) are detected and then mapped to the corresponding parts of the skeleton model. Implicit joints are found based on the skeleton model. The global shortest paths from the head to other extreme points are used to provide candidates for joints. Finally, a data-driven method is applied to each limb, to detect possible dominant joints. The dominant joints are determined by the voting the results of possible joints from each limb and joint candidates on the global shortest paths.

3.2. Extreme point detection and mapping

3.2.1. Extreme point detection

Extreme points on a human body include the head, hands, and feet. As part of the feature points, the spatial distribution of extreme points provides general information of human pose. Let P denote the 3D point cloud of a human body. The 3D point cloud is calculated from the captured depth frame. Starting from a randomly selected point in the 3D point cloud, denoted as p_0, the geodesic distance between any other point to p_0 is defined as the shortest geodesic distance to p_0. The geodesic distance between a given point to p_0 is calculated as follow:

$$D_g(p_0, P(x,y)) = \sum D_g(P(x_p, y_p), P(x_q, y_q)). \quad (1)$$

In the above equation, $P(x_p, y_p)$ and $P(x_q, y_q)$ are neighboring points on the shortest path between $P(x, y)$ and p_0. $D_g(\cdot)$ represents the geodesic distance between two points on the point cloud P. To calculate the distances from each point on the point cloud to p_0, we adopted an iterative method to go through the whole point cloud. Start from p_0, each point on the point cloud calculate the distance between itself to its eight nearest neighbors. The shortest distance from each of the eight neighbors to the p_0 will be updated. For a new point, its distance value is calculated; for a point that bears a distance, the distance is updated when a shorter distance is found. A distance map is generated by computing the distances to all points in the point cloud, and a point with the longest distance (i.e., an extreme point), denoted with E_1, is represented as follows:

$$E_1 = arg\max D_g(E_0, P(x,y)). \quad (2)$$

To avoid the same extreme points being repeatedly found, when an extreme point is identified, its geodesic distance to any existing extreme point is set to zero. Therefore, when a new extreme point is found, it must have the longest geodesic distance to all the existing points. Thus, five distance maps are usually required. In the proposed method, the extreme points include head, hands, and feet. Let M^i denotes the distance map, where i is the distance map index. The final updated distance map is as follows:

$$M(x,y) = min(M^1(x,y), M^2(x,y) \ldots M^n(x,y)). \quad (3)$$

Furthermore, Eq. (2) is rewritten in a more general form:

$$E_i = arg\max D_g(E_{i-1}, P(x,y)), i > 0. \quad (4)$$

To handle the self-occlusion, we compute the difference of the depth value between adjacent points when a distance map is updated. If the difference is less than a threshold δ, the two points are considered as lying on the same surface of the human body; otherwise, they are considered as belonging to different body parts. If the two points belong to different body parts, their geodesic distance remains unchanged. Therefore, for a point in P, its geodesic distance is only updated according to its neighboring points on the same body part. Fig. 2 shows two examples of the results of extreme point detection.

3.2.2. Extreme point mapping

When extreme points are detected, there is no correspondence between extreme points and body parts on the skeleton model. Without knowing the correspondence between extreme points and the skeleton model, it is difficult to detect the positions of joints.

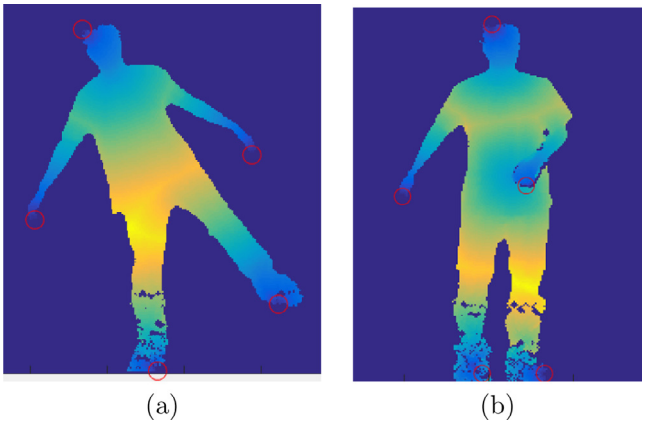


Fig. 2. Examples of extreme points. The red circles marked the position of the extreme points. (a) shows the result without self-occlusion, (b) shows the result with self-occlusion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Thus, mapping the extreme points to the human body model ensures that the data-driven method works with the human body model. The mapping method starts from mapping an extreme point to the head. The shape feature of the area around each extreme point is used to compare with an ellipse model of the head. The area with the highest likelihood is used as a head. To map the other extreme points to the human body model, the geodesic relationship between hands and feet is used. In the skeleton model, the geodesic distance between the head and the hands are shorter than the geodesic distance between the head and the feet, that is

$$D_g(p_{head}, p_{hand}) < D_g(p_{head}, p_{foot}). \tag{5}$$

With the above constraints, the extreme points of feet and hands are separated. To determine if an extreme point of hand corresponds to the left or right hand, we assume that the geodesic distance between the left hand and the left shoulder is shorter than the geodesic distance between the left hand and the right shoulder, and the same logic is applied to the right hand. The relationship between the left and the right hands can be described as follows:

$$\begin{aligned} &D_g(p_{Lh}, j_{Ls}) < D_g(p_{Lh}, j_{Rs}), \\ &D_g(p_{Rh}, j_{Rs}) < D_g(p_{Rh}, j_{Ls}), \end{aligned} \tag{6}$$

where p_{Lh} and p_{Rh} represent the extreme points of left and right hands, respectively, p_{Ls} and p_{Rs} represent the estimated joints of left and right shoulders, respectively. Estimating the position of shoulders is presented in Section 3.3.2. The relationship between the left and the right hands is also suitable for the left and the right feet.

3.3. Joint detection

3.3.1. Skeleton model

As part of our hybrid framework, the skeleton model estimates the positions of the implicit joints and provide constraints for data-driven joint detection algorithm. The traditional model-based human pose detection methods [5,25–27] define the human body model with a collection of body parts and DOFs (degrees of freedom) or joints with articulated structure and DOFs of joints. Our method defines the human body model only by defining the overall structure and general geodesic features of the human body model. For implicit joints, relative position and size are defined. On the other hand, the only relative position is defined for each dominant joint. Fig. 3 shows the skeleton model used in our method.

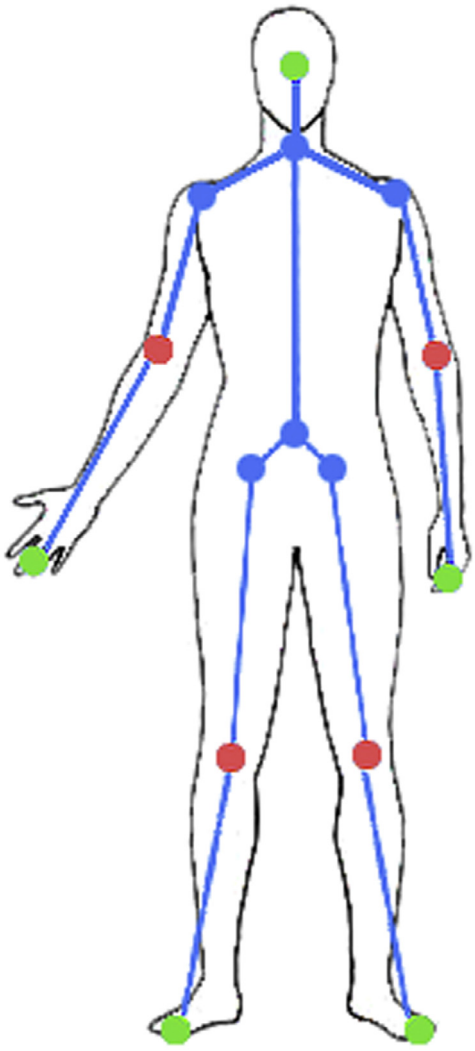


Fig. 3. The skeleton model used in our method. The green dots represent the extreme points. Blue dots represent implicit joints (neck, waist, shoulders and hips). Red dots represent dominant joints (elbows and knees). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In this figure, there are three types of point, the green points represent the extreme points, the blue points represent the implicit joint, and the red ones represent the dominant joints.

3.3.2. Estimation of the implicit joints

The implicit joints as part of the torso are more difficult to detect than the dominant joints such as elbows. However, because implicit joints locate within the torso, they have much-limited DOFs than the dominant joints. As a result, the model-based estimation methods provide reliable results for estimating the position of implicit joints. When estimating the position of implicit joints, we take full advantage of the geodesic features of the human skeleton to focus on the possible positions of joints. The estimation procedure follows a top-to-bottom order. The position of the neck is estimated first based on the position of the head. Given the length between neck and shoulder, denoted as l_{ns}, the left and right shoulders are defined as follows:

$$\{p_i \mid p_i \in P; D_g(p_{neck}, p_i) = l_{ns}; \tag{7}$$

$$D_g(p_{head}, p_i) > D_g(p_{head}, p_{neck})\}$$
$$\{p_j \mid p_j \in P; D_g(p_{neck}, p_j) = l_{ns};$$
$$D_g(p_{head}, p_j) > D_g(p_{head}, p_{neck})\}, \quad (8)$$

where $i \neq j$ and $p_i, p_j = arg\max_{i,j}(A(p_i, p_j))$.

In the above definition, p_i and p_j are two points in P, $A(\cdot)$ is the function to calculate the Euler angle between p_i and p_j. Eq. (9) ensures left and right shoulder are separated as much as possible. The hips are defined in a similar way to the shoulders because the structure of neck-shoulders and waist-hips are both triangle structure based on the skeleton structure of the human body. Thus, given the distance between the waist and the hips l_{wh}, the hips are defined as follows:

$$\{p_m \mid p_m \in P; D_g(p_{waist}, p_m) = l_{wh}; \quad (9)$$

$$D_g(p_{head}, p_m) > D_g(p_{head}, p_{waist})\}$$
$$\{p_n \mid p_n \in P; D_g(p_{waist}, p_n) = l_{wh};$$
$$D_g(p_{head}, p_n) > D_g(p_{head}, p_{waist})\}, \quad (10)$$

where $m \neq n$ and $p_m, p_n = arg\max_{m,n}(A(p_m, p_n))$.

Here, we assume that the geodesic distance from head to any shoulder is greater than that of the head to the neck, and the geodesic distance from head to any hip is greater than that of the head to the waist. The waist is defined as:

$$p_{waist} \in \{p_k \mid D_g(p_{head}, p_k) = l_w; \quad (11)$$
$$|D_g(p_{Ls}, p_k) - D_g(p_{Rs}, p_k)| < \mu\},$$

where l_w is the given distance from the head to the waist, μ denotes the threshold of the difference between the geodesic distance from the left and the right shoulder to the waist. The skeleton model requires the waist to have a close distance to the left and right shoulders. This ensures the scope of the waist is within the torso instead of arms. Fig. 4 illustrates the process and constraints for estimating the positions of implicit joints.

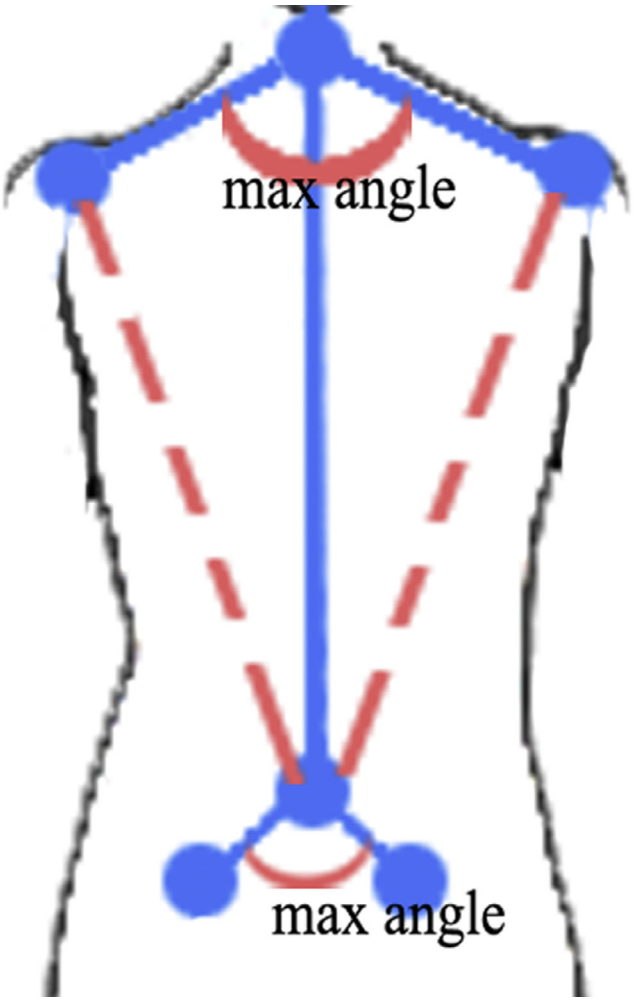


Fig. 4. Illustration of the constraints for estimating shoulders, waist, and hips.

3.3.3. Detection of the dominant joints

The dominant joints are elbows and knees. The data-driven method is used to detect these joints because dominant joints usually cause a greater magnitude of deformation of the limbs in contrast to the implicit joints. In our method, a method that integrates two detection strategies is developed to ensure accurate and stable detection results. A global shortest path based strategy is employed to detect candidates for the dominant joints, and a specific detection for each elbow and knee is employed. The detection results of elbows and knees are averaged results from both the shortest path based method and specific detection method.

The global shortest path based method uses the distance map similar to the distance maps used in Section 3.2.1. The distance map starts from the centroid point of the head, denoted as p'_{head}, and calculate the geodesic distance to all the other points in the point cloud. The shortest paths from p'_{head} to all extreme points can be generated during the updating procedure of the distance map. For each shortest path, given the start and end points of a path, denoted with p_i and p_n, respectively, the joint candidates on it should satisfy the following condition:

$$p_i, p_j \ldots p_n = \arg\min_{i,j\ldots n}(D_g - D_e), \quad (12)$$

and

$$D_g = D_g(p_i, p_j) + \ldots + D_g(p_{n-1}, p_n),$$

$$D_e = D_e(p_i, p_j) + \ldots + D_e(p_{n-1}, p_n).$$

The objective is to minimize the difference between the cumulative Euclidean distance and the geodesic distance of the path. To limit the number of joint candidates, the following restrictions are enforced:

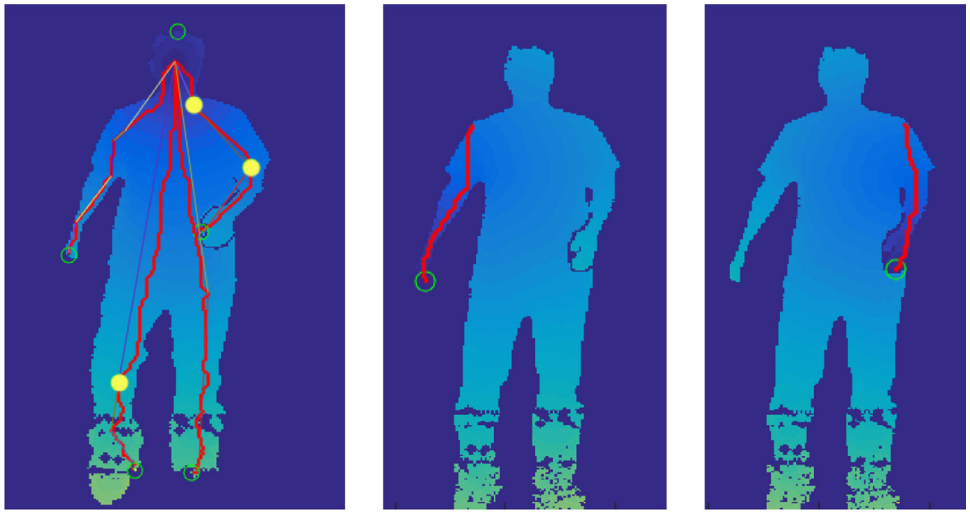
$$\forall p_i, A(p_i) < \beta \quad \text{and} \quad R_g(p_i) > \alpha, \quad (13)$$

where $A(p_i)$ represents the Euler angle formed by p_i and its two adjacent points p_{i-1} and p_{i+1}, and β and α are defined as threshold variables. The $R_g(p_i)$ is the geodesic distance ratio on p_i, defined as:

$$R_g(p_i) = \frac{\min(D_g(p_{i-1}, p_i), D_g(p_i, p_{i+1}))}{D_g(p_{i-1}, p_i) + D_g(p_i, p_{i+1})}. \quad (14)$$

The restrictions ensure that the joint candidates show how curvy the path is, and the points that close to the end points of the path are not found as candidates. Because the sharper the angle is and the greater the geodesic distance ratio is, the more contribution of the corresponding joint candidate makes to bend the limb. An example of the shortest path from the head to the other extreme points is shown in Fig. 5(a).

The objective of the specific joint detection is to detect the possible joint positions for each limb. A local shortest path from the corresponding extreme point to its closest implicit joint (e.g. shoulder or hip) is created. For example, the shortest path from the left hand to the left shoulder is created for detecting the position of the left elbow. Given the start and end points p'_{start} and p'_{end} of the shortest path on each limb, the detected joint must satisfy the following condition:



(a) (b) (c)

Fig. 5. Examples of the shortest paths. (a) the five shortest paths from head to all extreme points; (b) the shortest path from left shoulder to left hand; (c) the shortest path from right shoulder to right hand. Yellow dots in (a) are detected joints candidates. (b) and (c) are the shortest paths for specific detection.

Table 1
Detection Rate of Implicit Joints (%).

Feature Points	Neck	Shoulder		Hip	
		left	right	left	right
Detection Rate	88.3	88.0	88.5	83.6	83.1

$$p_k = \arg\min_k (D_e(p'_{start}, p_k) + D_e(p_k, p'_{end})).$$
$$A(p_k) < \beta \quad \text{and} \quad A(p_k) = \angle p_{start} p_k p_{end}. \tag{15}$$

This is to prevent random detection when the limb stretches straight. The position of each dominant joint on limbs is the average position of the joint candidates on the corresponding limb from Eq. (12) and the detected joint from Eq. (15). Furthermore, when a dominant joint cannot be detected, a geodesic middle point on the shortest path of the corresponding limb is used instead. Examples of the shortest path for the specific detection are shown in Fig. 5.

4. Experiments and evaluation

To evaluate our method, we record 10 videos with a Microsoft Kinect camera. In the acquired videos, the resolution of each frame is 512 × 424 pixels. The acquired videos contain various human poses such as walking, kicking, turning the upper body, and jumping. The reference points for joints are manually annotated in the 3D point cloud. Examples of the detection are depicted in Fig. 6, and three different views are shown for each result to give a 3D view of the joints.

4.1. Detection rate

Because the proposed method focuses on joint detection, it is necessary to evaluate the detection rate of joints. Table 1 lists the overall detection rate of the implicit joints.

Since the implicit joints are mostly estimated by the human skeleton model, the failure cases are mostly caused by the inaccurate head detection. The detection rate of hips is slightly lower than that of the shoulders because hands and other body parts occluded the hips in some of the frames in the data set. Geodesic features are used when estimating the implicit joints by the skeleton model, the areas of the hip with the corresponding geodesic distance value is not detectable when the areas are occluded by other body parts.

Table 2
Detection Rate of Dominant Joints (%).

Feature	Waist	Elbow		Knee	
Points		left	right	left	right
Detection Rate	86.7	90.1	89.3	89	90.2

Table 3
Overall Accuracy of Joints in terms of detection rate (%).

Neck	Waist	Shoulder		Elbow		Hip		Knee	
		left	right	left	right	left	right	left	right
81.3	86.7	88.3	88	87.2	86.3	83.6	84.1	84	86

The detection rate of the dominant joints is greater than that of the implicit joints partly due to the shortest path based and specific detections. In Table 2, we discuss the situations of the significant deformation occurring in the joints area. In practice, when the Euler angle of a bent limb is greater than 145°, it is considered as significant deformation, which can be detected by the proposed method. It is assumed that when a limb is fully stretched straight, the Euler angle on the corresponding dominant joint is 180°.

4.2. Accuracy of joint detection

In our evaluation, if a joint is within 6 cm of the selected reference point, then the detection is considered correct. The overall accuracy of all joints are listed in Table 3.

In Table 3, the accuracy of implicit joints (neck, waist, shoulders, and hips) are close to their detection rate. Because in the proposed method, the skeleton model finds the most suitable points for shoulders and hips after the geodesic constraints are calculated. Comparing to fixed structure human body model, our model can reduce the error distance for shoulders and hips. On the other hand, the overall accuracy of dominant joints is lower than their detection rate. Because when an elbow or knee is not detectable, a geodesic middle point is placed, and the middle points have bigger

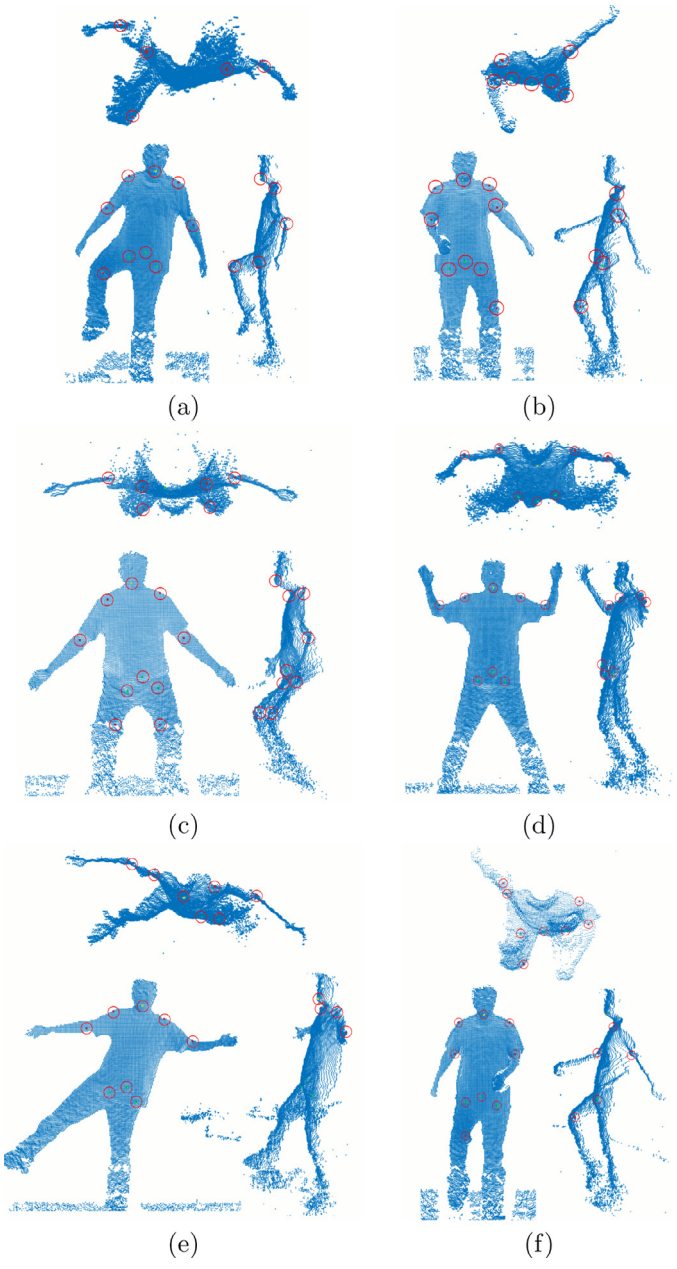


Fig. 6. Examples of detection results. The detected joints are marked with red circles. Each result is displayed in three views: top view at the top, front view at the bottom left and side view at the bottom right. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

error distances. A phenomenon that we realized from the experiments is that the deformation of the cloth on the testing object could affect the detection of shortest paths. Therefore, the deformation of cloth could affect the accuracy of dominant joint detection. As a result, only major joints are detected and minor joints such as ankles and wrists are left behind in the proposed method to ensure the accuracy.

We compare the proposed method against with our previous work [28], and the accuracy of elbows in the proposed method is 17.8% higher than our previous work. However, we also realize that the accuracy of shoulders is 5.4% lower. The Fig. 7 shows the comparison of the accuracy of elbows and shoulders between [28] and the proposed method. Because the method in [28] only detects shoulders and elbows as the result of joint detection, only

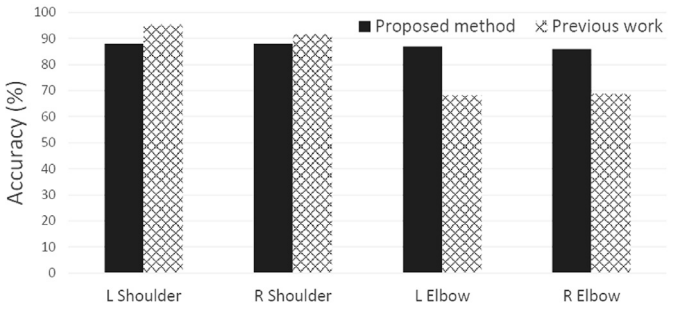


Fig. 7. Comparison between our previous work [28] (red shadow bars) and the proposed method (solid blue bars). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

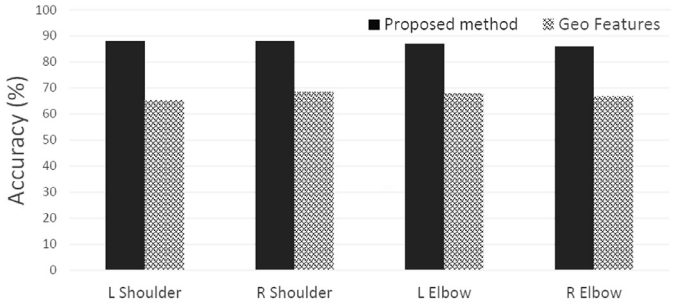


Fig. 8. Comparison between the method in [15] (green shadow bars) and the proposed method (solid blue bars). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the comparison data of elbow and shoulders are listed in Fig. 7. The major factor that causes the drop of accuracy on shoulders is that a general skeleton model is used in the proposed method. The accuracy of estimation of shoulders is affected by the detection of the head. In [28] a specific head-shoulder template is used to detect the positions of head and shoulders. Comparing the two different type of models, head-shoulder template can detect head more accurately than the ellipse head model, but it also produces large error distance in some cases, especially when the testing object give complex poses.

We also run the method of [15], which combines model-based estimation and data-driven detection to extract human poses to compare with our method. Because in [15] only shoulders and elbows are detected, only the accuracy of shoulders and elbows are listed in the Fig. 8. The average accuracy of elbows and shoulders of our method is 21.79% higher than the accuracy of [15]. In [15], the positions of shoulders are estimated by calculating the average position of selected points with a certain distance to the head and centroid of the torso, fixed searching range is defined for selecting points. In our method, an adaptive skeleton is applied, which improves the accuracy of shoulder detection. When detecting elbows, the shortest paths provide a better set of joint candidates, and the shortest paths have fewer chances to be affected by the edges of the clothes on human bodies. Comparing to [15], the average accuracy of elbows in our method is 19.25% higher.

4.2.1. Error distance

Error distance is calculated as the Euclidean distance between detected points and reference points. On a small area on the surface of the human body, it is close to the geodesic distance.

The average and the max error distance and listed in Table 4. The average error distance of waist and hips are higher than the neck and shoulders, due to the cumulative error caused by the model. Furthermore, hips have no clear boundary on the human

Table 4
Error distance (in cm).

	Neck	Sho.	Elbows	Waist	Hips	Knees
Avg Err	4.2	4.1	3.3	5.2	5.5	4.2
Max Err	6	6.1	6.8	7.4	8.8	6.7

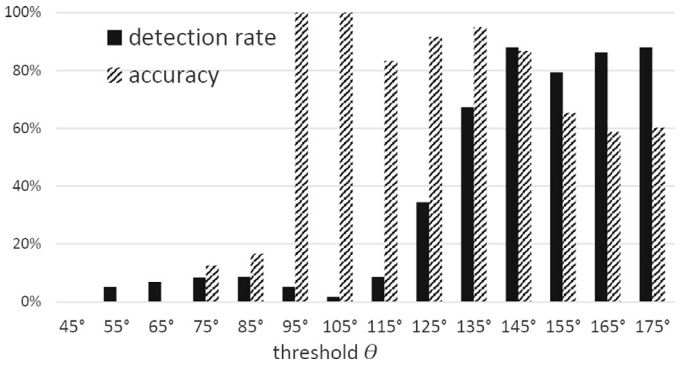


Fig. 9. Detection rate and accuracy of dominant joints using different threshold θ.

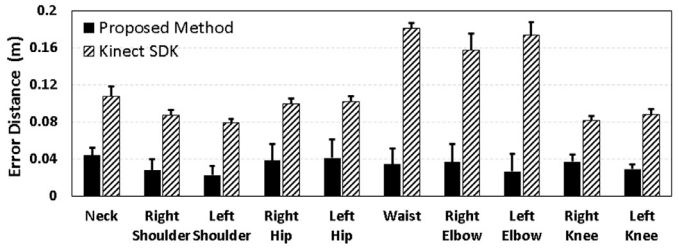


Fig. 10. The average error distance of the detected joints using our proposed method (solid bars) and the Microsoft Kinect SDK (textured bars).

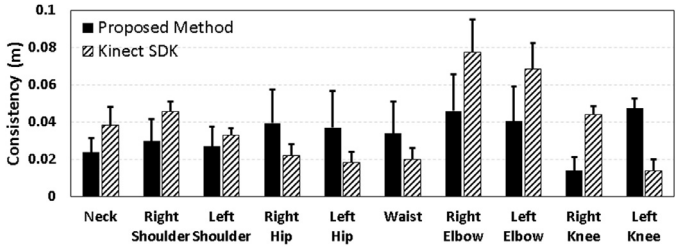


Fig. 11. Average consistency of the detected joints using our proposed method (solid bars) and the Microsoft Kinect SDK (textured bars).

body, but shoulders have the clear boundary, which makes them easier to find. Elbows and knees have smaller average error distance than the dominant joints, due to the mixture of two detection methods. The average and max error distances of the shoulders of [15] are 5.7cm and 10.2cm, respectively, and for elbows, the average and max error distances are 4.8cm and 10.1cm, respectively. The max error distance in [15] is mainly caused by the deformation of the edges of clothes.

4.2.2. Analysis of parameters

In our analysis of parameters, 200 frames that contain 15 different poses were used. In our proposed method, a threshold δ is used to verify if two adjacent points belong to the same surface. Table 5 lists the average accuracy of joint detection with different δ values. It is clear that the system achieves the highest accuracy (87%) among all possible thresholds when δ is at 45 mm. When a lower threshold is used, more points on the same body surface are mistaken as points on the different body surface. On the other hand, as this threshold is increased, points on a different surface

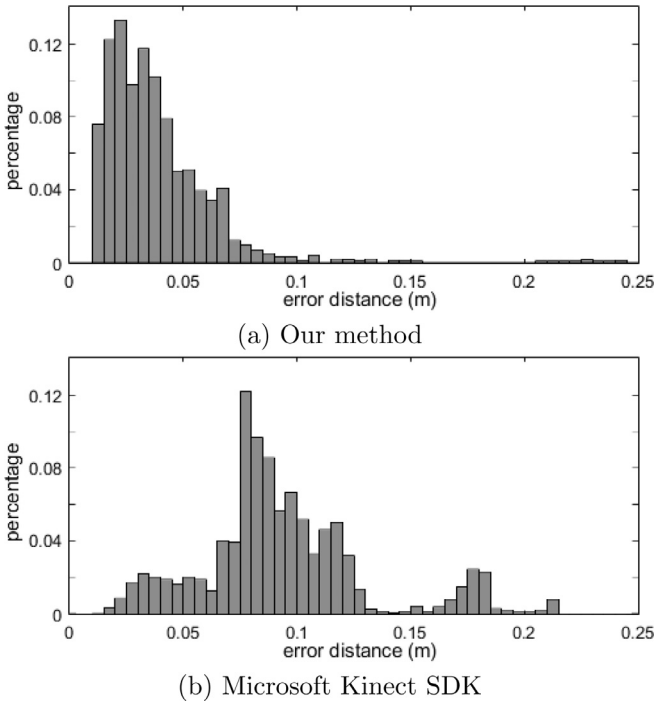


Fig. 12. Histograms of error distance.

Table 5
Accuracy (%) of detecting joints with different δ (mm).

δ	15	25	35	45
Acc.	15.21	77.17	65.21	87.67
δ	55	65	75	85
Acc.	67.93	65.21	58.69	59.29

are considered to be on the same surface, which, consequently, degrade the accuracy. The choice of threshold δ affects the procedure of updating the distance map and, hence, it influences the accuracy of detecting both implicit and dominant joints. In the rest of our experiments, the threshold δ is 45 mm.

Another threshold used in our method is θ for selecting candidates for dominant joints, which is the angle of the two vectors formed by three adjacent points. The three adjacent points are selected by the geodesic distance ratio. In general, a small angle allows a fewer number of candidates to be selected. We conducted experiments with different θ and evaluated the average accuracy and detection rate as shown in Fig. 9. As θ increases, the detection rate increases, and best detection rate was achieved with θ at 145° and 175°. The accuracy, however, varies fluctuated with the increment of θ. When θ was at 95° and 105°, the accuracy reached nearly 100%. This is due to the low detection rate. Within a few successful detections, the joints were accurate. By considering both detection rate and accuracy, we set θ to 145° in the rest of our experiments.

4.2.3. A comparison study with Microsoft Kinect SDK

We conducted a comparison study with Microsoft Kinect SDK following the study in [29] and evaluated the accuracy and consistency of our proposed method. Fig. 10 illustrates the average error distance of the detected joints. The error distance is measured with respect to the ground truth marked manually on the acquired data. It is shown that the average error distance of the detection of joints using Microsoft Kinect SDK is 11.56cm; whereas that of our proposed method is 3.36cm. The largest errors in the results of SDK are related to waist and elbows, which are in the

range of 16cm and above. Our proposed method demonstrated much-reduced error distance. The error bars in Fig. 10 depict the standard deviation (STD) and the average STDs for our proposed method and the SDK are 1.36cm and 0.8cm, respectively. It is evident that the proposed method exhibited much-improved accuracy in comparison to Microsoft Kinect SDK.

We also evaluated the consistency of joint detection. The consistency is gauged by the distance to the initial detection of each joint. That is, the joint detection of a consistent method deviates slightly, if any, regardless of the poses. Fig. 11 illustrates the bar plot of consistency with respect to the ten joints. Our method exhibited greater consistency for six joints and SDK achieved better consistency for hips, waist, and left knee. The overall average consistencies for our method and the SDK are 3.38cm and 3.8cm, respectively. The error bars in Fig. 11 show the standard deviations. The consistencies of the two methods are comparative with a slight advantage to our method.

Fig. 12 illustrates the histograms of error distance. The distribution of our method is condensed to the lower end and the distribution of the SDK is scattered across the entire scale. The skewness of our method is 1.575 and the skewness of the SDK is 1.091, which indicates that the error distance distribution of our method is statistically better than that of the SDK.

5. Conclusions

In this paper, we proposed a hybrid framework for accurate joint detection for human pose estimation. In our proposed method, joints are categorized into two classes including implicit joints and dominant joints. Model-based and data-driven strategies are used to estimate and detect the position of joints in the human body. Both strategies take advantage of the geodesic features of the human body to locate the joints accurately. Our experimental results demonstrated that an integrated method provides more stable and accurate results. Furthermore, the data-driven method that uses global shortest path and local shortest path can be widely used in different types of methods for human pose detection. The geodesic distances between the extreme points and the joints can be used for tracking and estimating the position of joints when the joints are occluded. Complex and multi-layer self-occlusions could cause failure of detection in our method. Our method failed to detect the joints when the body parts and limbs are occluded. In our future work, we plan to employ temporal information to improve the detection accuracy and robustness. The geodesic distance between a joint and an extreme point is useful for tracking the joints.

References

[1] R. Pinho, J. Tavares, Tracking features in image sequences with Kalman filtering, global optimization, Mahalanobis distance and a management model, Comput. Model. Eng. Sci. 46 (1) (2009) 51–75.

[2] R. Pinho, J. Tavares, M. Correia, An improved management model for tracking missing features in computer vision long image sequences, WSEAS Trans. Inf. Sci. Appl. 1 (4) (2007) 196–203.

[3] R. Pinho, J. Tavares, M. Correia, A movement tracking management model with Kalman filtering, global optimization techniques and Mahalanobis distance, Lect. Ser. Comput. Comput. Sci. 4A (2005) 463–466.

[4] J. Tavares, A. Padilha, Matching lines in image sequences using geometric constraints, in: the 7th Portuguese Conference on Pattern Recognition, Portugal, 1995.

[5] Y. Zhu, B. Dariush, K. Fujimura, Controlled human pose estimation from depth image streams, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.

[6] M. Sigalas, M. Pateraki, P. Trahanias, Full-body pose tracking-the top view reprojection approach, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1569–1582.

[7] S. Zuffi, M. Black, The stitched puppet: a graphical model of 3D human shape and pose, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3537–3546.

[8] M. Ye, X. Wang, R. Yang, L. Ren, M. Pollefeys, Accurate 3D pose estimation from a single depth image, in: 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 731–738.

[9] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A.A.B.A. Kipman, Efficient human pose estimation from single depth images, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2821–2840.

[10] Y. Zhou, J. Han, X. Yuan, Z. Wei, R. Hong, Inverse sparse group lasso model for robust object tracking, IEEE Trans. Multimedia 19 (8) (2017) 1798–1810.

[11] L.L. Presti, M.L. Cascia, 3D skeleton-based human action classification: a survey, Pattern Recognit. 53 (2016) 130–147.

[12] C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Real-time identification and localization of body parts from depth images, IEEE Int. Conf. Rob. Autom. (2010) 3108–3113.

[13] A. Baak, M. Muller, G. Bharaj, H. Seidel, C. Theobalt, A data-driven approach for real-time full body pose reconstruction from a depth camera, in: IEEE International Conference on Computer Vision, Barcelona, 2011, pp. 1092–1099.

[14] E. Weng, L. Fu, On-line human action recognition by combining joint tracking and key pose recognition, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4112–4117.

[15] S. Handrich, A. Al-Hamadi, A robust method for human pose estimation based on geodesic distance features, in: IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 906–911.

[16] L. Schwarz, A. Mkhitaryan, D. Mateus, N. Navab, Estimating human 3D pose from time-of-flight images based on geodesic distances and optical flow, in: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, 2013, pp. 700–706.

[17] M. Vasconcelos, J. Tavares, Human motion segmentation using active shape models, in: International Conference on Computational and Experimental Biomedical Sciences, 2015, pp. 237–246.

[18] X. Zhang, C. Li, W. Hu, X. Tong, S. Maybank, Y. Zhang, Human pose estimation and tracking via parsing a tree structure based human model, IEEE Trans. Syst. Man Cybern.: Syst. 44 (5) (2014) 580–592.

[19] J. Cui, Y. Liu, Y. Xu, H. Zhao, H. Zha, Tracking generic human motion via fusion of low- and high-dimensional approaches, IEEE Trans. Syst. Man Cybern.: Syst. 43 (4) (2013) 996–1002.

[20] J. Shen, W. Yang, Q. Liao, Part template: 3D representation for multiview human pose estimation, Pattern Recognit. 46 (7) (2013) 1920–1932.

[21] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 755–762.

[22] X. Wei, P. Zhang, J. Chai, Accurate realtime full-body motion capture using a single depth camera, ACM Trans. Graph. - Proc. ACM SIGGRAPH Asia 31 (6) (2012). Article No. 188

[23] K. Buys, C. Cagniart, A. Baksheev, T.D. Laet, J.D. Schutter, C. Pantofaru, An adaptable system for RGB-D based human body detection and pose estimation, J. Vis. Commun. Image Represent. 25 (1) (2014) 39–52.

[24] K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognit. 71 (2017) 402–413.

[25] Y. Zhu, B. Dariush, K. Fujimura, Kinematic self retargeting: a framework for human pose estimation, Comput. Vision Image Understanding 114 (12) (2010) 1362–1375.

[26] H.-D. Yang, S.-W. Lee, Reconstructing 3D human body pose from stereo image sequences using hierarchical human body model learning, in: 18th International Conference on Pattern Recognition (ICPR'06), volume 3, 2006, pp. 1004–1007.

[27] M.W. Lee, I. Cohen, A model-based approach for estimating human 3d poses in static images, IEEE Trans. Pattern Anal. Mach. Intell. 28 (6) (2006) 905–916.

[28] X. Yuan, L. Kong, D. Feng, Z. Wei, Automatic feature point detection and tracking of human actions in time-of-flight videos, IEEE/CAA J. Autom. Sin. 4 (4) (2017) 677–685.

[29] Q. Wang, G. Kurillo, F. Ofli, R. Bajcsy, Evaluation of pose tracking accuracy in the first and second generations of Microsoft Kinect, in: IEEE International Conference on Healthcare Informatics, Dallas, TX, 2015.

Longbo Kong received a B.S. degree in computer science from Dalian University of Technology, China in 2010, a M.S. degree in computer science from the University of North Texas, USA. in 2013, and a Ph.D. degree in computer science and engineering from the University of North Texas, USA. in 2017. He is a member of the Computer Vision and Intelligent Systems Laboratory. His research interests include computer vision, data mining, and artificial intelligence.

Xiaohui Yuan received a B.S. degree in electrical engineering from the Hefei University of Technology, Hefei, China in 1996 and a Ph.D. degree in computer science from the Tulane University, USA in 2004. He is currently an Associate Professor at the Department of Computer Science and Engineering in the University of North Texas. His research interests include computer vision, data mining, machine learning, and artificial intelligence. His research findings have been reported in over one hundred peer-reviewed papers. Dr. Yuan is a recipient of Ralph E. Powe Junior Faculty Enhancement award in 2008 and the Air Force Summer Faculty Fellowship in 2011, 2012, and 2013.

Amar Man Maharjan received a B.S. degree in computer science and a M.S. degree in computer science and information technology from the Tribhuvan University, Nepal, in 2005 and 2009, respectively. He is currently a Ph.D. candidate and a member of the Computer Vision and Intelligent Systems Laboratory at the University of North Texas (UNT). He is a recepient of the UNTs Multicultural Scholastic Award 2017. His research interests include computer vision, data mining, and artificial intelligence.