IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 17, 2020, accepted September 27, 2020, date of publication September 30, 2020,
date of current version October 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027892

3D Joints Estimation of the Human Body in Single-Frame Point Cloud

TIANXU XU[1], DONG AN[1], ZHONGHAN WANG[1], SICHENG JIANG[1], CHENGNUO MENG[1], YIWEN ZHANG[1], QIANG WANG[2], ZHONGQI PAN[3], (Senior Member, IEEE), AND YANG YUE[1,2], (Member, IEEE)

[1]Institute of Modern Optics, Nankai University, Tianjin 300350, China
[2]Angle AI (Tianjin) Technology Company Ltd., Tianjin 300450, China
[3]Department of Electrical and Computer Engineering, University of Louisiana at Lafayette, Lafayette, LA 70504, USA

Corresponding author: Yang Yue (yueyang@nankai.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0703500, in part by the Key Technologies Research and Development Program of Tianjin under Grant 20YFZCGX00440, and in part by the Fundamental Research Funds for the Central Universities, Nankai University, under Grant 63201178 and Grant 63191511.

ABSTRACT Joint estimation of human body in point cloud is a key step for tracking human movements. In this work, we present a geometric method to achieve detection of the joints from a single-frame point cloud captured using a Time-of-Flight (ToF) camera. Three-dimensional (3D) human silhouette, as global feature of the single-frame point cloud, is extracted based on the pre-processed data, the angle and aspect ratio of the silhouette are subsequently utilized to perform pose recognition, and then 14 joints of human body are derived via geometric features of 3D silhouette. To verify this method, we test on an in-house captured 3D dataset containing 1200-frame depth images, which can be categorized into four different poses (upright, raising hands, parallel arms, and akimbo). Furthermore, we test on a subset of the G3D dataset. By hand-labelling the joints of each human body as the ground truth for validation and benchmarks, the average normalized error of our geometric method is less than 5.8 cm. When the distance threshold from the ground truth is 10 cm, the results demonstrate that our proposed method delivers improved performance with an average accuracy in the range of 90%.

INDEX TERMS Depth camera, human pose detection, joint detection, sensor systems and applications.

I. INTRODUCTION

Human behavior recognition aims to interpret human behavior by a computer, and is one of the most important technologies in computer vision. By analyzing human behavior in image sequences and identifying behavior categories, human behavior recognition is widely used for intelligent monitoring [1], and video analysis [2]. In general, human behavior can be regarded as the continuous evolution of the spatial configuration of rigid segments connected by joints [3]. If the human skeleton can be extracted and tracked reliably, the human behavior then can be classified by action recognition. At present, the detection of human joints has been widely used in the fields of virtual reality [4], automatic driving [5] and elderly care system [6].

The first step of human behavior analysis is to capture the human pose. Before the debut of the depth camera, one typically utilized 2D images or video taken by traditional cameras. However, there are few limitations in this process that need to be addressed. For example, typical factors that are associated with 2D image and may affect joint extraction and recognition, include light conditions, changing background environment, variable clothing, and human body occlusion [7].

Recent advent of low-cost depth sensors provides an alternative solution to extract human joints. Depth sensors can capture three-dimensional (3D) depth data of the scene, and are more robust to lighting change, thus provide more useful information to restore 3D human skeleton. These advantages shift the research focus of computer vision to depth camera.

The 3D information obtained by depth camera can be expressed in different forms, such as depth image, point cloud, voxel mesh, etc. The extraction of 3D features from human body can be divided into global and local features. The former includes the edge features of human body, silhouette, optical flow information, etc. The global feature contains

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

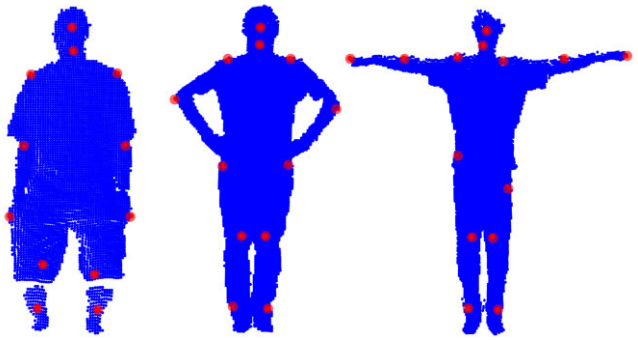

FIGURE 1. Some results of the joint extraction for different poses.

relatively rich information, but it is also sensitive to the environmental noise and the viewpoint change. The latter is to extract the relatively independent unit or the interested image block of human body for analysis. It is insensitive to noise and occlusion, but it is missing the global information and high computational complexity.

The objective of our work is to detect joints of the human body in a single-frame point cloud, and some examples are shown in Fig. 1, the point cloud is acquired by a depth camera. Compared with an ordinary camera, the depth camera is not affected by the illumination and complexity of background. Moreover, the point cloud is relatively simple in terms of expression, and maintains the most primitive geometric information in 3D space without any discretization. Our proposed method uses global geometric information to extract joints of human body, it requires neither a large amount of dataset to train, nor a template model as a prior condition. There is no doubt that it can significantly save computational cost.

In our proposed method, the main technical contributions are listed below: (1) data is pre-processed with filtering, rotation, and silhouette extraction; (2) geometrical figure of merit for common pose classification is defined, which can distinguish the pose by the angle between the head and the farthest points on the positive and negative x-axis, and a further classification is carried out according to the aspect ratio of ranges for the x-axis and the y-axis; (3) the geometric features of the 3D silhouette are utilized to locate 14 joints of the single-frame point cloud in four different poses; (4) the algorithm is performed on the public G3D dataset [8] and the in-house captured dataset with an < 5.8 cm average distance error, demonstrating a high recognition rate and improved accuracy compared with the other methods.

The rest of this article is organized as follows: Section II reviews the related work on human joints extraction. Section III presents the 3D joints extraction of the human body in single-frame point cloud. Section IV analyzes and compares the experimental results. Section V discusses the limitations of the proposed method. Section VI concludes this paper with potential future work.

II. RELATED WORK

Joint estimation is inspired by Johansson's classic mobile light display experiment, which simply replaces the whole human body with a group of points [9]. Existing approaches of joint estimation can be broadly grouped into three main categories: feature points detection, 3D human body model and deep learning-based approaches. Methods that detect feature points usually use the relevant global characteristics of the human body. Body constraints can be applied to first roughly mark each major component in the depth image, then the specific positions of the joints are determined by using Iterative Closest Point (ICP) to search approximate position of the component [10]. Compared with the whole image search method, this method not only improves the search speed, but also avoids falling into the local minimum of the algorithm. Kong *et al.* [11] presented a hybrid framework for automatic joint detection using a depth camera. By constraining the geodesic distance between every two joints, implicit and dominant joints can be detected. In the process of finding joint points, Dijkstra's algorithm is often used to detect the joints. Plagemann *et al.* [12] used it to find the corner points of the human body, whereas leverages local shape description to distinguish these corner points. Their angles are also estimated, so as to obtain the position information of each part of the human body. Baak *et al.* [13] applied a modified Dijkstra's algorithm to find the extreme points in the human body silhouette, and the direction of the extreme point is calculated by tracking the search path in reverse. Wang *et al.* [14] could regress 3D coordinates of mesh vertices at different resolutions from the latent features of point clouds by developing a spatial-temporal mesh attention convolution to predict the locations at the high resolution, based on the features of point clouds at the low resolution.

The method based on human body model can also obtain the joints. Wei *et al.* [15] fitted the 3D point cloud obtained by the depth camera with the human body model, and continuously optimized to obtain the accurate joints positions of the human body. However, their human body model is relatively simple with the trunk and limbs being represented by simple cylinders. Liang *et al.* [16] further expanded this method by using the fusion data of depth point cloud obtained by multiple Kinects. The surface model of human body was obtained and fitted in the way of spherical harmonic function. This further improves the accuracy of joint point position. Kim *et al.* [17] created a behavior template set, which contained weighted human joint data with invariant attributes to scaling and rotation. To further improve the accuracy of the detection results, graphical models that impose kinematic constraints were utilized to improve estimation of full-body pose [18], [19]. Marin *et al.* [20] built a given parametric model of the human for non-rigid registration with robustness to a large variety of nuisances. Xu *et al.* [21] achieved a sparse set of key points annotated in the low-quality point clouds to guide the deformation of a high-fidelity human body model, which was automatically fitted to 2D joints of the human body using CNN based methods.

At present, many learning-based methods also focus on human joint estimation. Shotton *et al.* [22] transformed the difficult pose estimation problem into a pixel classification

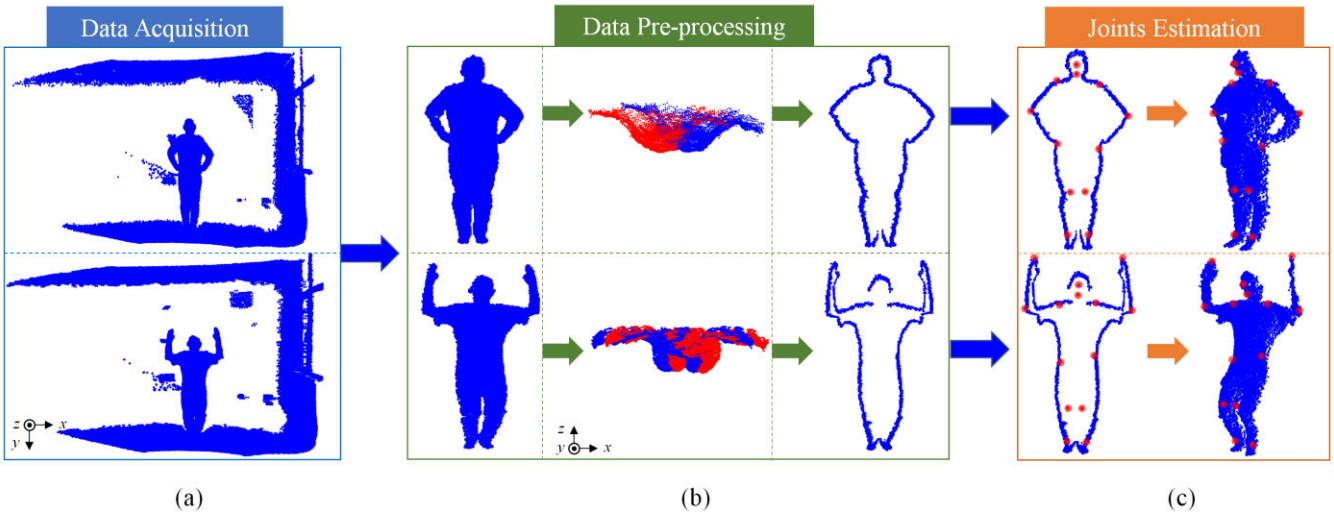

FIGURE 2. Overview of the proposed approach. The proposed approach consists of three stages: data acquisition, data pre-processing and joint estimation. (a) The point clouds are directly obtained from the depth camera. (b) Data pre-proposing mainly involves three parts: firstly, the irrelevant points are filtered, then the orientation of the human point cloud is adjusted, and finally the 3D silhouette is extracted. (c) 14 joints of human body are extracted by using the geometric feature of human silhouette.

model. The method of random forest was used to classify the pixels with different characteristics, and the single depth map was represented by the middle body part. Finally, 3D coordinates of each joint for the different body part were obtained. Zhou *et al*. [23] used 3D body voxel to represent 3D human pose, and then applied CNN to predict the possibility of 3D joints in each voxel. Sun *et al*. [24] used ConvNet to get the 3D heatmap representation of the input data, and then combined the heatmap with the joint regression task to perform the integration operation. In this way, the non-differentiable post-processing and quantization errors caused by the heatmap representation are avoided. Li *et al*. [25] addressed the problem of multiscale graph, using dynamic neural networks to predict 3D skeleton-based human motions. Marín-Jiménez *et al*. [26] proposed a Deep Depth Pose (DDP) model, where 3D human pose in a depth image was designed by linear combination of some predefined pose prototypes, and the human body 3D joints were determined. Zhang *et al*. [27] used PointNet++ network to estimate human joint points, but it required 2D joints as prior information, and could not directly obtain joints on the point cloud. Jiang *et al*. [28] adopted the parametric human body model SMPL, and used the same network to locate joints, however, due to the smoothness of the model, the influence of noise and clothes is not considered. Although the learning-based method can directly obtain the 3D positions of the body joints from the depth image, the accuracy of extraction results is still affected due to the scarcity of training data and the complexity of calculation.

Some recent studies used micro-Doppler radar to determine human behavior with radio frequency signals, however it still cannot provide spatial information of the subject [29], [30]. Li *et al*. [31] proposed a temporal Range-Doppler PointNet-based method to analyze human behavior. Human echoes could be first transformed to 3D point sets, and then sent to the hierarchical PointNet model for classification.

III. METHOD

We propose an architecture for 3D joints estimation of human body. A purely geometric approach is used to obtain the 14 common joints of the human body from a single-frame point cloud in the coordinate system of depth camera. The specific overview of the proposed approach is shown in Fig. 2. This system consists of three main modules: data acquisition, data pre-processing and joint estimation. Unlike laser scanners, human point cloud can be obtained through a simple and low-cost depth camera in the data acquisition module. The data is then exported to a computer for processing if needed. Our ultimate goal is to estimate the positions of 14 joints from each frame, and then represent each joint by its position of (x, y, z) in 3D Cartesian coordinate system that is expressed in meter. The x, y, and z axes are the body axes of the depth sensor. This is a right-handed coordinate system, where the positive z-axis extends to the direction in which the sensor array points. The positive y-axis extends downward, and the positive x-axis extends to the right (with respect to the sensor array). The three coordinates for joint position are presented in Fig. 2. It is worth to note that, we focus on extracting joints using a simple geometrical method from a single-frame point cloud containing four human poses (upright, raising hands, parallel arms, akimbo). Our algorithm aims to extract joints of human body, the pipeline of the algorithm is shown in Fig. 3. However, lots of information in the point cloud is redundant, and the viewpoint direction of point cloud needs to be adjusted. Before extracting joints of the human body, it is necessary to process the point clouds in advance, which can improve the recognition accuracy of algorithm and save the calculation time.

FIGURE 3. The pipeline of the joints extraction. The processing flow of this method includes data acquisition, data pre-processing, pose classification, joints extraction, and error calculation.

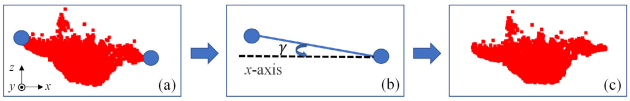

FIGURE 4. An example of adjusting point cloud orientation, (a) Top view of the point cloud (blue dots represent the farthest points on the x-axis), (b) Rotation angle of the straight line composed of two dots to the x-axis, (c) Top view of point cloud with correct direction.

A. DATA PRE-PROCESSING

The depth camera can obtain scene information with a large field of view up to several meters. In order to avoid the complexity of the joint extraction process, we first segment the human body from the whole scene point cloud. Given the depth information, one effective way to remove irrelevant background from the point cloud is to utilize a conditional filter. It is a very simple and common filtering method that can delete points in the point cloud, which do not meet the conditions specified by the user. An example of filtering human point cloud and extracting silhouette is shown in Fig. 2(b). By setting the segmented ranges along the x, y, and z axes, one can highlight foreground human point cloud by performing the filtering operation once.

Due to the huge amount of point cloud acquired by the instrument, not all points are useful for joint extraction. To make the problem tractable, silhouette, as a global feature, is frequently used in image processing. This method can save the computing resources, and retain the core pose information. We adopt the public algorithm to extract silhouette features of a point cloud in the PCL library, and the results are shown in Fig. 2(b). This method is described as following: Firstly, the local neighborhood of the source point is defined by the k-nearest neighbor method. Secondly, the tangent plane of the neighborhood is fitted with the least square method, and the point normal vector is calculated. Then, the neighborhood points which are projected onto the tangent plane form a vector with the source point, each vector is saved into an array. By taking a certain vector as the reference, the angle between the other and the reference one is calculated accordingly, and the angles are listed in a descending order. If the maximum difference between the adjacent angles is greater than a certain threshold, the point is defined as the boundary point. After the 3D silhouette is obtained, all operations in this section are performed on the 3D silhouette.

Given that our method is designed to handle incomplete shapes, the input to our method is simply an oriented point cloud, unlike a complete 3D human body model. When the viewpoint of the camera is not aligned with the direction of the human body being facing, it will cause partial occlusion, which increases the difficulty of recognition. Therefore, it is necessary to rotate the human body so that it is in the same direction as the viewpoint. Fig. 4(a) illustrates that we project the raw point cloud along the positive direction of the y-axis to obtain a top view of the point cloud. The farthest two points on the x-axis are picked out in the top view denoted blue dots in Fig. 4(b). The straight line is determined between the two points of the projected point cloud. Then it's simple to compute the angle γ between the x-axis and the straight line, which is equal to the rotation angle of point cloud. When the angle is acute, rotate the angle clockwise, and rotate counterclockwise otherwise. The final correction result is shown in Fig. 4(c).

B. JOINTS ESTIMATION

The proposed method estimates joints from four types of human poses, upright, raising hands, parallel arms and akimbo. It is sufficient for size measurement of human body and other applications. Without occlusion, when working with the four poses, 14 common joints are defined for representing these poses: head, left / right (L/R) feet, L/R shoulder, L/R hand, L/R elbow, L/R knee, L/R waist, and neck. The extraction process of joints is also slightly different for different poses. In order to solve the problem of pose classification, we pre-compute three points to assist pose classification: head and the farthest points which locate on the positive and negative x-axis of the point cloud. According to the threshold of body proportion, the head is segmented to get the centroid point. Meanwhile, L/R feet can be obtained using the same method for subsequent processing. α, which is the angle between the head and the farthest points on the positive and negative x-axis of the point cloud, is calculated. According to the different degrees of arm opening in the four poses, one determines which pose the current point cloud belongs to. Fig. 5(a) describes algorithm flow chart of human pose classification.

When α is equal or less than τ_a, which represents angle threshold, it is judged as an upright pose as shown in Fig. 5(b). However, when this condition is not met, it is difficult to strictly distinguish the angles of the three poses. Therefore, other methods are used to assist. Firstly, β is the ratio of the farthest distance on the x-axis to the farthest distance on the y-axis. When it is equal or greater than τ_b, it is a pose with parallel arms as depicted in Fig. 5(c). Secondly, by taking advantage of the symmetry of the pose, the farthest point on the x-axis is picked out, and its y-coordinate is denoted as y_2. y_1 is the minimum values of y-axis in raw point cloud. τ_c is used as the criterion for judging whether the pose is the raising hands or the akimbo, are shown in Fig. 5(d) and Fig. 5(e) respectively. If the distance between y_1 and y_2 is equal or greater than τ_c, it is the pose of raising hands. Otherwise, it is the akimbo pose. The thresholds τ_a, τ_b and τ_c are respectively set to 70, 0.7, 0.85 in our experiments if not specified otherwise.

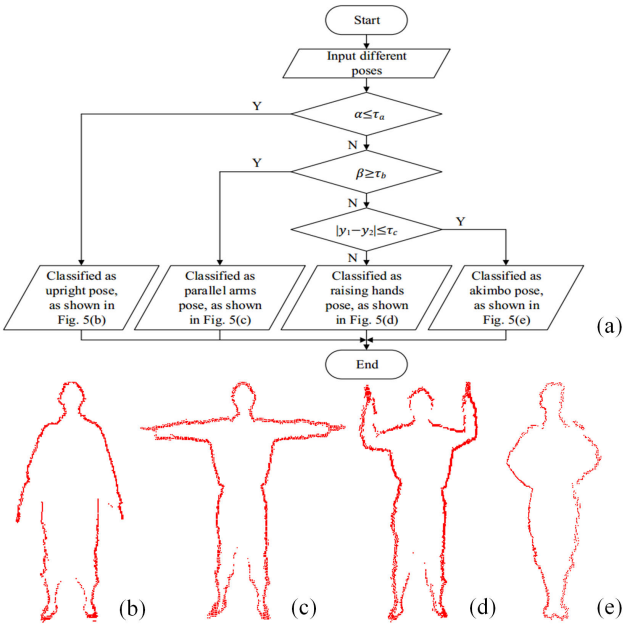

FIGURE 5. Recognition of point cloud poses. (a) Algorithm flow chart of human pose classification, (b)-(e) Schematic diagrams of four poses corresponding to upright, parallel arms, raising hands and akimbo, respectively.

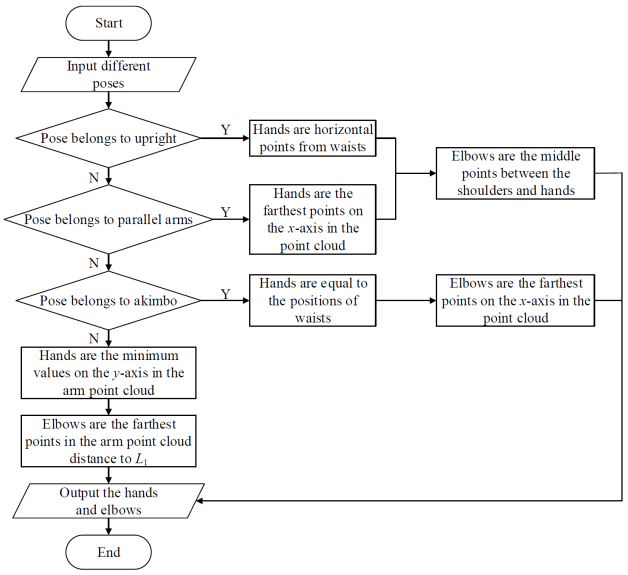

FIGURE 6. Algorithm flow chart of extracting hands and elbows from different poses, L_1 represents the straight line between the shoulder and the head.

1) PARALLEL ARMS POSE

In the current pose, the waist silhouette is segmented in terms of the waist proportion threshold. Let P denotes the 3D point cloud of the waist. In order to make the calculation easier, we first define an xy-plane using a parametric model, and then project the points of waist silhouette onto the plane. Because the pose of human body is generally symmetrical structure, it is easy to calculate the central line of the plane where the middle point of the farthest points on the positive and negative direction of x-axis is located. Using the central line, the projected point cloud can be divided into two parts: the left waist and the right waist. To find the waist joints, we traverse the points in the left and right waist respectively to compute the distances from the central line. Here, L_m denotes the central line, the waist joints are calculated as follows:

$$\{\boldsymbol{p_i}|\boldsymbol{p_i}\in\boldsymbol{P_W}, \boldsymbol{D_{PL}}\,(\boldsymbol{J_W}, \boldsymbol{L_m}) = \arg\min \boldsymbol{D_{PL}}\left(\boldsymbol{p_i}, \boldsymbol{L_m}\right) \quad (1)$$

where P_W denotes the left or right waist silhouette, p_i is a point of P_W, $D_{PL}(\cdot)$ represents the distance from a point to a line, and J_W is the coordinates of waist joint. By calculating the distance between each point of the left and the right waists to L_m, the point with the shortest distance from L_m is then found, and it is re-projected back to the 3D coordinate to get the L/R waist, which serves as part of the references for the detection of the other joints.

Starting with the waists, we can easily find the L/R shoulder by extending the vertical lines from waists along the negative y-axis to get the intersections with the 3D silhouette. After obtaining the 3D coordinates of the shoulders, the location of the middle point, denoted as p_{MSP}, can be calculated between the L shoulder and the R shoulder. This parameter will be favorable for detecting the neck and knees. As shown in Fig. 6, the L/R hand is defined as the farthest points on the x-axis of the 3D silhouette. The neck is the midpoint between the head and p_{MSP}. The L/R elbow is located in the same way as the midpoint of the hand and shoulder. The knee recognition depends on whether the knee is upright or bent. We first segment the 3D silhouette below the waist, and further divide it into two parts, the left leg and the right leg. When the human knee is upright, the L/R knee is located on the straight line formed by p_{MSP} and L/R feet. The specific location satisfies the condition that the distance from the knee to the p_{MSP} are three times the distance to the L/R feet. And when the knee is bent, the knee is regarded as the farthest point on the negative z-axis of the L/R leg. To solve the crucial issue how to recognize the state of the human knees, we first calculate the z-coordinate range of the L/R leg, which is denoted by Δz_k. Similarly, the y-coordinate range can be obtained, which is represented by Δy_k, and the above process is shown in Fig. 7. When the ratio of Δz_k to Δy_k is less than the threshold value, the knees are in the upright state. Otherwise, they are in the bent state.

2) AKIMBO POSE

The method provides a consistent means for finding the locations of some joints above. The difference is that the farthest points on the x-axis of the 3D silhouette are not L/R hand, but the L/R elbow. As the subject maintains the akimbo pose, the position of the waist overlaps the position of the hand. It is reasonable to assume that the positions of the two joints are equal in this pose. Fig. 6 displays the estimation process of the hand and elbow in the akimbo pose.

3) RAISING HANDS POSE

Due to the change of arm pose, the initially proposed method is not suitable for finding the hand and elbow in this pose.

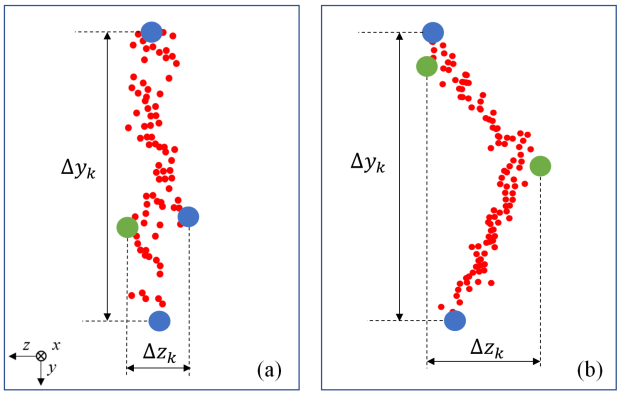

FIGURE 7. Extraction of the knee. (a) Left view of the knee silhouette in the state of upright knees (blue and green dots respectively represent the farthest points on the y-axis and z-axis, Δz_k, and Δy_k are respectively the z-coordinate and y-coordinate range of the knee silhouette), (b) Left view of the knee silhouette in the state of bent knees.

We can take the waist as the base point, and portion the points whose x-coordinate is smaller or larger than the waists, then cluster them into left or right arm silhouette. As a result, the point with minimum y-coordinate in the left or right arm silhouette are L/R hand. The L/R elbow is considered as the farthest points of left or right arm silhouette distance to the straight line named L_1 between L/R shoulder and head. Fig. 6 illustrates the extraction process of hand and elbow in the raising hands pose.

4) UPRIGHT POSE

Due to the occlusion of hands in this pose, the waists are invisible implicit point. We can find the concave points within the range of the waist silhouette to get waist position, and it does not affect the other joints cached on this basis. The intersection of the waist and the silhouette along the x-axis, which is collinear with the waist on the x-axis, is denoted as the L/R hand, the extraction process is shown in Fig. 6.

IV. EXPERIMENTS AND RESULTS

The proposed geometric method was implemented in C++ using PCL. We test on the part of public G3D dataset and the dataset captured by ourselves. In our dataset, we capture four main poses (upright, raising hands, parallel arms and akimbo) of 10 people (5 men and 5 women), which contains 1200 frames depth images with a Time-of-Flight (ToF) camera (Camboard Pico Monstar). Since the method uses no temporal information, we are interested only in static pose. In the acquired images, the resolution of each frame is 352 × 287 pixels. While maintaining a pose, the knees have two forms: upright and bent respectively. We record 5 times for every pose, each time subject remains in the same position. When a group of poses is finished, each subject puts on different clothes and repeats the previous pose. Each person changes three types of clothes. To evaluate our method, we hand-labelled with 14 joint positions. Fig. 8 gives some depth images of four poses in the dataset.

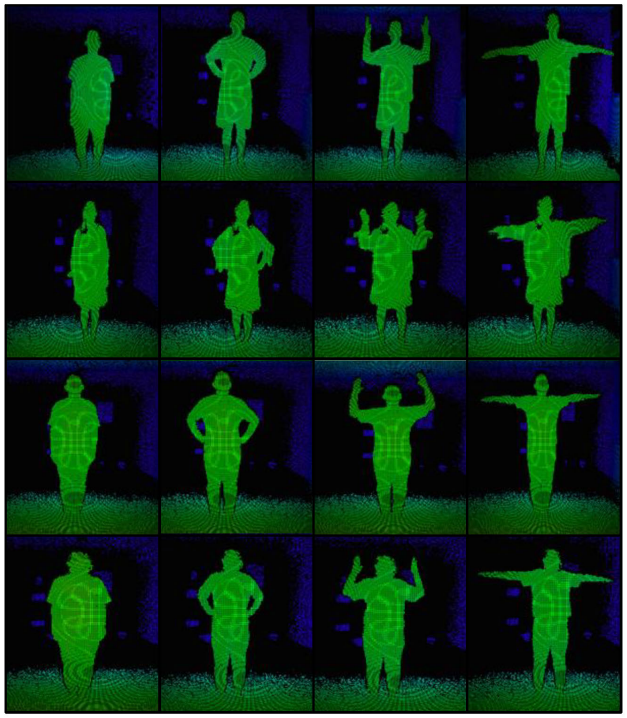

FIGURE 8. Some depth images of human body in different poses collected by a ToF camera. For each subject, four main poses (upright, raising hands, parallel arms and akimbo) are captured, and the knees have two forms: upright and bent respectively in each pose.

The results obtained from our dataset via the proposed algorithm are shown in Fig. 9. The first column is the point cloud of different poses. The silhouette obtained in the pre-processing is shown as green points in the raw point cloud. The second column is the joints calculated by the proposed algorithm displayed on the 3D human silhouette. Although the joints we obtain can be displayed in the silhouette, there are z offsets between some joints and the raw point cloud surface. To locate the positions of the joints on the surface of the body, we find the intersection of the joint and the point cloud along the z-axis, which is collinear with the joint on the z-axis, to determine a final position of joint, the results are shown in the third column. Additionally, 14 joints are hand-labelled on the point cloud after pre-processing to provide the ground truth joints. Due to the limitation of recognizable poses, we select 22 data in the G3D dataset. The extracted joints include 10 people who perform 20 types of actions. The same analysis processing is performed, and the result is shown in Fig. 10.

To further prove the feasibility of the proposed algorithm, the overall accuracy of joints is calculated as listed in Table 1. Compared with Kong *et al.* [11] and Shotton *et al.* [22], we set 6 cm and 10 cm respectively as a reasonable threshold to judge whether the joint is detected. When a joint within 6 cm is chosen as the threshold in Table 1, the overall accuracy is decreased. In the proposed method, because the difference in the clothing of human has a greater impact on the silhouette, which introduces a large detection error of the waists. Besides, the shoulders use the waists as the base point,

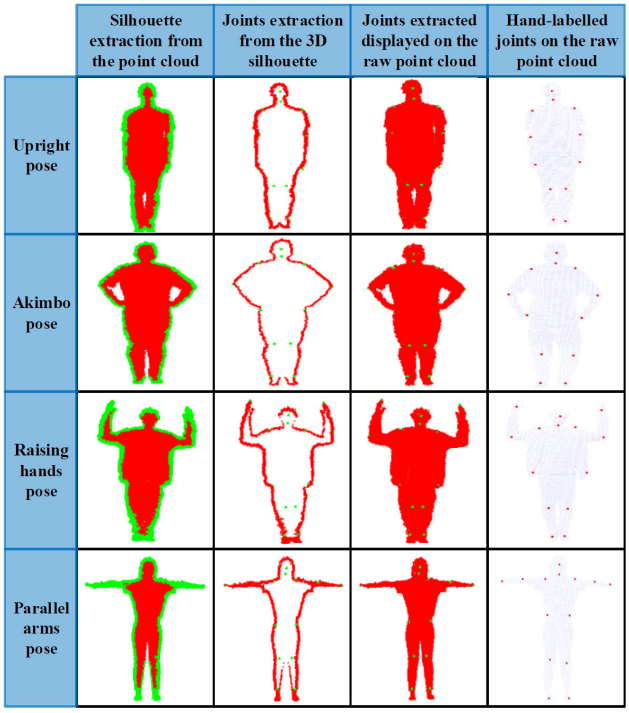

FIGURE 9. The results of extracting 3D joints from our dataset. The first column is extracted silhouette from the raw point cloud, the second column is the joints map extracted by the algorithm from the silhouette, the third column is the extracted joints displayed on the raw point cloud, and the fourth column is the hand-labelled joints map.

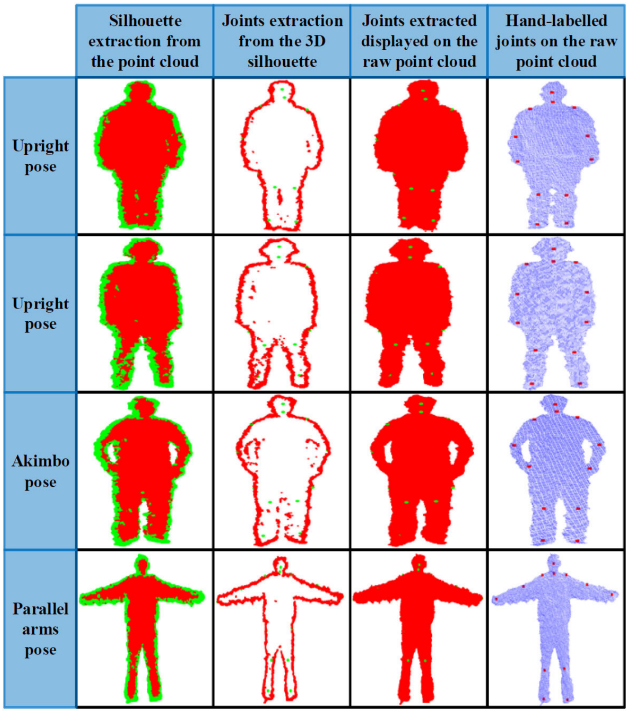

FIGURE 10. The results of extracting 3D joints from G3D dataset. The first column is extracted silhouette from the raw point cloud, the second column is the joints map extracted by the algorithm from the silhouette, the third column is the extracted joints displayed on the raw point cloud, and the fourth column is the hand-labelled joints map.

when the knees are upright, the knees are determined by the shoulders and the foot. Therefore, the waists indirectly

TABLE 1. Overall accuracy of joints (%).

Joints	Method			
	ε=10 cm		ε=6 cm	
	I	Ours	II	Ours
Head	90.00	98.74	-	97.90
L Shoulder	73.00	99.15	88.30	94.89
R Shoulder	74.00	98.73	88.00	94.51
L Elbow	75.00	100	87.20	92.92
R Elbow	76.00	100	86.30	92.37
L Hand	66.00	94.12	-	70.59
R Hand	66.00	94.54	-	80.25
L Knee	58.00	90.83	84.00	53.75
R Knee	57.00	90.83	86.00	38.75
L Feet	74.00	99.58	-	87.39
R Feet	73.00	99.58	-	83.61
Neck	87.00	97.48	81.3	70.16
L Waist	75.00	62.18	86.7	39.50
R Waist	74.50	52.94	86.7	26.89
Average	72.75	91.34	86.05	73.11

L and R are abbreviations of left and right, respectively. "-" denotes the value is not given. I and II indicate (Shotton *et al.*[22]) and (Kong *et al.*[11]), respectively.

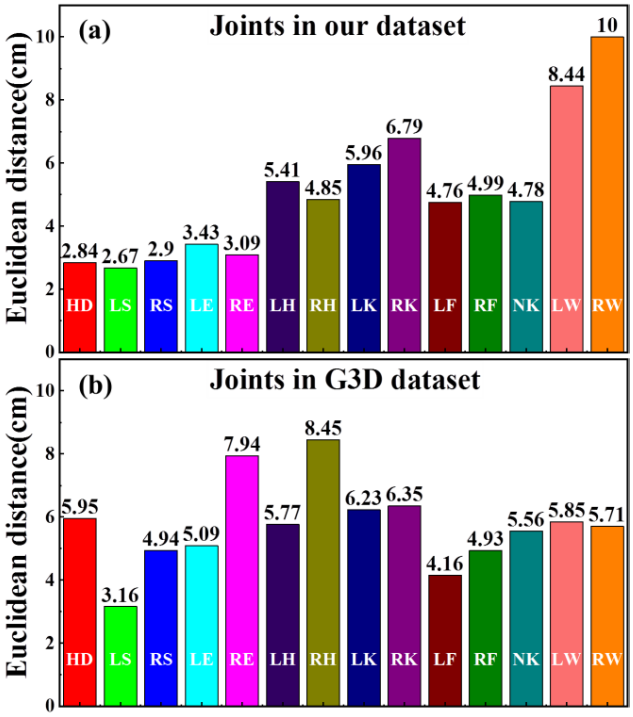

FIGURE 11. Average error of the 14 joints obtained on human examples in our dataset and G3D dataset, the abbreviation of joint name is consistent with each joint.

affect the detection of the knees, and we need to optimize the detection method of the waists in the subsequent work. When a joint within 10 cm is chose as the threshold in Table 1, the accuracy of waists decreases a little, but it is clear that the method improves the overall accuracy by 18.59 %. We believe that the success rate of identifying the correct joint will be tremendously improved, in cases when the detection of the waists is optimized.

In addition, we quantify average normalized 3D Euclidean distance error as evaluation criteria per joint, which is shown

in Fig. 11. In our evaluation, the error distance is measured with respect to the ground truth hand-labelled on the acquired data. In Fig. 11(a), because of loose-fitting clothes, the extraction of the waists on the silhouette has large deviation, when in the upright pose, the waists indirectly affect the detection of the knees and hands, and the process of automatic correction can be further included in the future work. Fig. 11(b) shows the average error of the joints in G3D dataset, the appearance of the clothes is more close-fitting, and the waist error is small. On the other hand, the error of the elbows and the hands is significantly enhanced. The reason behind is that the hand-labelled is based on the point cloud, and our method is to extract the joints on the silhouette. In addition, due to the lack of the denoising procedure and available samples, so the error of the other joints will increase. Finally, the average error of all parts results on our dataset and G3D dataset, achieving 5.07 cm and 5.72 cm respectively. The above results illustrate that the algorithm is feasible to extract joints from four positions without occlusion.

V. DISCUSSION

Besides the experimental results shown above, it is necessary to discuss some limitations in each step of our approach.

Initialization of the parameters. We assume that regardless of gender, each part of the point cloud is segmented according to the proportion of normal human criteria in this paper, and some thresholds (Section III), such as τ_a, τ_b and τ_c, are equal to the maximum values of the frequency statistics by calculating the corresponding distribution of thresholds in the dataset. Compared with the learning-based method, the accuracy and robustness of mega data need to be further improved.

Diversification of the poses. In order to achieve subsequent measurement of human body, we apply four poses to locate the joints of the human body. In future work, the spatial position and length information between the joints can be used as the prior condition so that improve the detection accuracy of joints and expand the diversity of the pose.

Computational time. Since we use geometric algorithm to extract feature points, the joints extraction process takes approximately 103 ms and the step of classifying the poses takes 2.66 ms. In comparison, the data pre-processing part is more time-consuming, with an average of 9497.07 ms. As the raw point cloud contains more points, it takes a lot of time for the calculation of silhouette. This part can also be processed offline. The system runs on a regular desktop computer (Intel Core i3-6100K CPU running at 3.7 GHz, and 8 GB of RAM). If we use GPU-enhanced computer, the computing time will be greatly reduced.

VI. CONCLUSION

In this paper, we have proposed a geometric method to extract 3D joints from a single-frame point cloud of human body collected by a depth camera. The method, neither does it require many training sets and predefined templates, nor does it need to limit the appearance or wearing of the subject under test. In our proposed method, we first classify the pose by the angle between the head and the farthest points on the x-axis, and a further classification is carried out according to aspect ratio of x-axis and y-axis range. By making use of advantageous geometric features, 14 joints can then be extracted in four different poses from the point cloud. To further verify the feasibility of our proposed algorithm, we test our algorithm using both the dataset captured by ourselves and the public G3D dataset, the experimental results demonstrated that our proposed method achieves improved accuracy in comparison to the others, the average normalized error is less than 5.8 cm. For future work, we plan to further study the extraction of joints in different poses, especially in the case of occlusion. The calculation time of silhouette can be optimized. Furthermore, we can combine the measurement technology to achieve multi-size measurement of human body based on a single-frame point cloud.

REFERENCES

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[2] W. Lin, X. He, W. Dai, J. See, T. Shinde, H. Xiong, and L. Duan, "Key-point sequence lossless compression for intelligent video analysis," *IEEE MultimediaMag.*, vol. 27, no. 3, pp. 12–22, Jul. 2020, doi: 10.1109/MMUL.2020.2990863.

[3] K. M. Knutzen, "Kinematics of human motion," *Amer. J. Hum. Biol.*, vol. 10, no. 6, pp. 808–809, 1998.

[4] S. Lu, L. Cai, X. Ding, and F. Gao, "A combined strategy of hand tracking for desktop VR," in *Proc. PCM*, Hefei, China, 2018, pp. 256–269.

[5] U. E. Manawadu, M. Kamezaki, M. Ishikawa, T. Kawano, and S. Sugano, "A hand gesture based driver-vehicle interface to control lateral and longitudinal motions of an autonomous vehicle," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 1785–1790.

[6] M. Liang and Y. Hu, "Application of human body posture recognition technology in robot platform for nursing empty-nesters," in *Proc. 6th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2020, pp. 91–95.

[7] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. IJCAI*, Beijing, China, Jun. 2013, pp. 1–7.

[8] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 7–12.

[9] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, Jun. 1973.

[10] Y. Zhu, B. Dariush, and K. Fujimura, "Controlled human pose estimation from depth image streams," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.

[11] L. Kong, X. Yuan, and A. M. Maharjan, "A hybrid framework for automatic joint detection of human poses in depth frames," *Pattern Recognit.*, vol. 77, pp. 216–225, May 2018.

[12] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 3108–3113.

[13] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1092–1099.

[14] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang, "Sequential 3D human pose and shape estimation from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7275–7284.

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

[16] L. Shuai, C. Li, X. Guo, B. Prabhakaran, and J. Chai, "Motion capture with ellipsoidal skeleton using multiple depth cameras," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 2, pp. 1085–1098, Feb. 2017.

[17] H. Kim, S. Lee, Y. Kim, S. Lee, D. Lee, J. Ju, and H. Myung, "Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system," *Expert Syst. Appl.*, vol. 45, pp. 131–141, Mar. 2016.
[18] L. He, G. Wang, Q. Liao, and J.-H. Xue, "Depth-images-based pose estimation using regression forests and graphical models," *Neurocomputing*, vol. 164, pp. 210–219, Sep. 2015.
[19] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 755–762.
[20] R. Marin, S. Melzi, E. Rodolà, and U. Castellani, "FARM: Functional automatic registration method for 3D human bodies," *Comput. Graph. Forum*, vol. 39, no. 1, pp. 160–173, Feb. 2020.
[21] Z. Xu, W. Chang, Y. Zhu, D. Le, H. Zhou, and Q. Zhang, "Building high-fidelity human body models from user-generated data," *IEEE Trans. Multimedia*, early access, Jun. 10, 2020, doi: 10.1109/TMM.2020.3001540.
[22] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Jun. 2011, pp. 1297–1304.
[23] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 901–914, Apr. 2019.
[24] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 529–545.
[25] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 214–223.
[26] M. J. Marín-Jiménez, F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "3D human pose estimation from depth maps using a deep combination of poses," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 627–639, Aug. 2018.
[27] Z. Zhang, L. Hu, X. Deng, and S. Xia, "Weakly supervised adversarial learning for 3D human pose estimation from point clouds," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 5, pp. 1851–1859, May 2020.
[28] H. Jiang, J. Cai, and J. Zheng, "Skeleton-aware 3D human shape reconstruction from point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5431–5441.
[29] F. Jin, R. Zhang, A. Sengupta, S. Cao, S. Hariri, N. K. Agarwal, and S. K. Agarwal, "Multiple patients behavior detection in real-time using mmWave radar and deep CNNs," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2019, pp. 1–6.
[30] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "Mm-pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10032–10044, Sep. 2020.
[31] M. Li, T. Chen, and H. Du, "Human behavior recognition using range-velocity-time points," *IEEE Access*, vol. 8, pp. 37914–37925, 2020.

TIANXU XU received the B.S. and M.S. degrees from Zhengzhou University, Zhengzhou, Henan, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree in optical engineering with the Institute of Modern Optics, Nankai University, Tianjin, China.

DONG AN received the bachelor's degree in optical information science and technology from Anhui University, Hefei, Anhui, China, in 2019. He is currently pursuing the Ph.D. degree in optical engineering with the Institute of Modern Optics, Nankai University, Tianjin, China.

ZHONGHAN WANG is currently pursuing the bachelor's degree in electronic science and technology with Nankai University, Tianjin, China.

SICHENG JIANG is currently pursuing the bachelor's degree in optoelectronic information science and engineering with Nankai University, Tianjin, China. His current research interest includes computer vision.

CHENGNUO MENG is currently pursuing the bachelor's degree in optoelectronic information science and engineering with Nankai University, Tianjin, China. Her current research interest includes computer vision.

YIWEN ZHANG received the B.S. degree in optical information science and engineering from the Dalian University of Technology, Dalian, Liaoning, China, in 2018. She is currently pursuing the M.S. degree in optical engineering with the Institute of Modern Optics, Nankai University, Tianjin, China. Her research interests include machine learning and fiber network systems.

QIANG WANG, photograph and biography not available at the time of publication.

ZHONGQI PAN (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, China, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA.

He is currently a Professor with the Department of Electrical and Computer Engineering. He also holds BORSF Endowed Professorship in electrical engineering II, and BellSouth/BORSF Endowed Professorship in telecommunications. He has authored/coauthored 160 publications, including five book chapters and 18 invited presentations/papers. He also has five U.S. patents and one China patent. His research interests include photonics, including photonic devices, fiber communications, wavelength-division-multiplexing (WDM) technologies, optical performance monitoring, coherent optical communications, space-division-multiplexing (SDM) technologies, and fiber sensor technologies. He is an OSA Senior Member.

YANG YUE (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and optics from Nankai University, Tianjin, China, in 2004 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2012.

He is currently a Professor with the Institute of Modern Optics, Nankai University. He has published over 170 peer-reviewed journal articles and conference proceedings, three edited books, one book chapter, more than ten invited papers, more than 30 issued or pending patents, and more than 80 invited presentations. His current research interests include intelligent photonics, optical communications and networking, optical interconnect, detection, imaging and display technology, integrated photonics, free-space, and fiber optics.

Dr. Yue is a member of the IEEE Communications Society (ComSoc), the IEEE Photonics Society (IPS), the International Society for Optical Engineering (SPIE), the Optical Society of America (OSA), and the Photonics Society of Chinese-American (PSC). He also served as a Committee Member. He also served as a Session Chair for ∼30 international conferences. He is an Associate Editor of IEEE Access. He is also an Editor Board Member of three other scientific journals. He also served as a Guest Editor for seven journal special issues. He also served as a Reviewer for more than 50 prestigious journals and OSA Centennial Special Events Grant 2016.