Efficient Body Motion Quantification and Similarity Evaluation Using 3-D Joints Skeleton Coordinates

Aouaidjia Kamel, Bin Sheng, Ping Li, Jinman Kim, and David Dagan Feng, *Fellow, IEEE*

***Abstract*—Evaluating whole-body motion is challenging because of the articulated nature of the skeleton structure. Each joint moves in an unpredictable way with uncountable possibilities of movements direction under the influence of one or many of its parent joints. This paper presents a method for human motion quantification via three-dimensional (3-D) body joints coordinates. We calculate a set of metrics that influence the joints movement considering the motion of its parent joints without requiring prior knowledge of the motion parameters. Only the raw joints coordinates data of a motion sequence are needed to automatically estimate the transformation matrix of the joints between frames. We also consider the angles between limbs as a fundamental factor to follow the joints directions. We classify the joints motion as global motion and local motion. The global motion represents the joint movement according to a fixed joint, and the local motion represents the joint movement according to its first parent joint. In order to evaluate the performance of the proposed method, we also propose a comparison algorithm between two skeletons motions based on the quantified metrics. We measured the comparative similarity between the 3-D joints coordinates on Microsoft Kinect V2 and UTD-MHAD dataset. User studies were conducted to evaluate the performance under different factors. Various results and comparisons have shown that our method effectively quantifies and evaluates the motion similarity.**

***Index Terms*—Human–computer interaction, motion quantification, similarity evaluation, three-dimensional (3-D) human motion representation.**

Manuscript received December 4, 2018; revised February 21, 2019; accepted May 4, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61872241 and Grant 61572316, in part by the National Key Research and Development Program of China under Grant 2017YFE0104000 and Grant 2016YFC1300302, in part by the Macau Science and Technology Development Fund under Grant 0027/2018/A1, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 18410750700, Grant 17411952600, and Grant 16DZ0501100. This paper was recommended by Associate Editor R. Roberts. *(Corresponding author: Bin Sheng.)*

A. Kamel is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

B. Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the MoE Key Laboratory of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

P. Li is with the Department of Computing, Hong Kong Polytechnic University, Hong Kong (e-mail: lipingfire@ieee.org).

J. Kim and D. D. Feng are with the Biomedical and Multimedia Information Technology Research Group, School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia (e-mail: dagan.feng@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2019.2916896

I. Introduction

UNDERSTANDING human behavior is indispensable for automating many tasks. Human body motion analysis is a branch of understanding human behavior, and it is fundamental for various applications, such as surveillance [1]–[3], human–computer interaction [4]–[6], health care of the elderly [7], [8], human gait analysis [9], [10], and robotics [11]–[16]. A variety of motion capture systems have been built to capture the human motion using different techniques, including wearable devices and sensors, multiple cameras, or even a single camera. Some of the systems combine many devices together to improve the capturing quality and generate accurate positions of the human body joints. Recently, the fast development in machine learning algorithms [17], [18] greatly improved the quality of motion data. Microsoft Kinect [19] can accurately estimate three-dimensional (3-D) human body joints to form a skeleton model from a single depth image using the method in [20]. Real-time motion capture from just a single RGB image is proposed in [21] via a deep learning convolutional neural network (CNN) model to estimate accurate 3-D human body poses.

The accuracy of the body key joints generated from the recently developed systems motivated researchers to work on understanding and analyzing the human motion to extract informative data that can be useful for exploitation in several applications. A lot of attempts have been made to recognize the human actions from a 3-D skeleton model through processing the joints coordinates [22]–[24]. However, all those proposed methods can only recognize a limited number of actions from training samples of a dataset because there are unlimited possibilities of the body movements that make it extremely difficult to recognize all possible actions. One big challenge facing the human motion analysis in general is the sensitivity of the joints motion in 3-D space. The articulated nature of the human body complicates the representation and the quantification of the body motion because the movement of any joint can be due to itself or due to the movement of one or many of its parent joints. This restriction makes it ambiguous to distinguish the motions just by looking at the joint movement. For example, the movement of the wrist joint can be due to itself, due to the elbow, due to the shoulder, or due to the movement of the whole body. Moreover, without a clue about the joints direction, the movement of a joint from right to left or left to right will be considered the same if they have the same trajectory shape. Furthermore, it is difficult to know exactly whether the shape of the joint movement trajectory

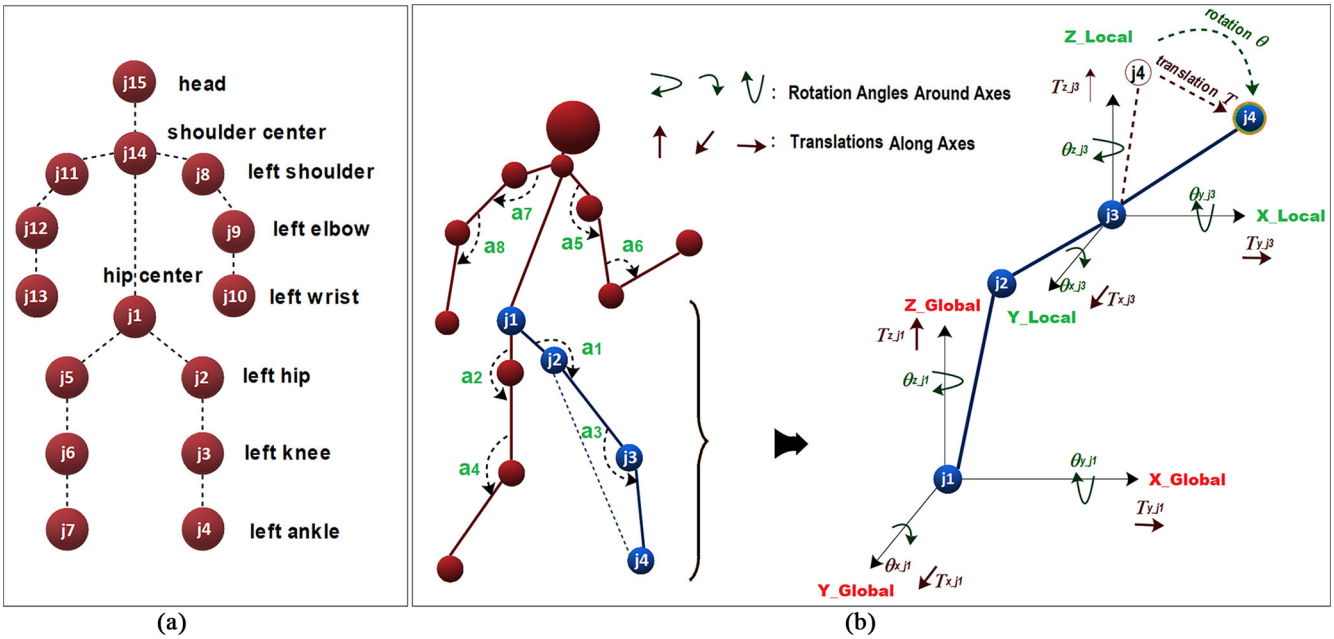


Fig. 1. (a) Skeleton model used for motion quantification. (b) Metrics used in the proposed quantification algorithm: a detailed illustration of the left leg coordinates. The rotation and the translation of the joint $j4$ according to its parent joint ($j3$) are considered as local motion, whereas its rotation and translation according to the hip center ($j1$) are considered as global motion. $a1$, $a2$, ..., $a8$ are the angles between limbs.

is circular or straight in order to decide if the rotation or the translation is more convenient to represent the movement. The human motion analysis requires a detailed and effective representation that can quantify the joints movement direction, angles, etc., so that each movement will have a unique representation. The previous analysis motivated us to think that the best motion evaluation method must cover all the aspects that influence the joints movement and, hence, the whole-body motion.

In this paper, we present a method for human motion quantification. By motion quantification, we refer to providing a numeric representation for the body motion based on a set of a calculated metrics. Accurate quantification is indispensable to distinguish between different human movements even when they are performed in a slightly different way. Our objective is to propose a robust representation for the body motion using 3-D joints coordinates generated from any motion capture system. The proposed algorithm takes a sequence of input skeletons in a form of 3-D joints coordinates then generates a set of metrics for each joint. We consider two types of metrics: 1) motion metrics are calculated between each pair of consecutive frames and 2) angles metrics are calculated within each frame. Since the joints displacement from one frame to another can be represented by a rotation or a translation depending on its trajectory, we consider both measurements to estimate the transformation matrix between each pair of frames from the joints coordinates directly. We also categorize the motion metrics into two categories: 1) the local motion to represent each joint displacement according to its first parent joint in the skeleton hierarchy and 2) the global motion to represent the motion of each joint according to a fixed joint in the skeleton hierarchy which we consider as the origin joint because its motion is not influenced by any of the other joints [hip center in Fig. 1(a)]. The motivation behind analyzing the joint motion locally is that the local motion provides information about the influence of the parent joints to overcome the problem of identifying which joints in the skeleton hierarchy are the cause of the movement. We also consider the angles between skeleton limbs to check how far the joints are moving from each other during the motion.

In order to test the effectiveness of the motion quantification algorithm, we propose a comparison algorithm to evaluate the similarity between two movements of the human body using our quantification. The algorithm takes the outputs metrics from the quantification phase of the two movements then generates a percentage that indicates the level of similarity based on a calculated distance between the metrics of the two skeletons. Using the motion quantification and comparison algorithms together can be of a great benefit for many motion evaluation applications in practice. Fig. 2 gives a general idea about our proposed method. The main contributions of this paper can be summarized as follows.

1) *A motion quantification algorithm* that estimates the movement of the body from only a raw 3-D joints coordinates without the need of prior knowledge about the body joints motion parameters that are usually offered by motion capture systems, which makes it appropriate to be used with any system that generates 3-D body joints.
2) *A motion comparison algorithm* evaluates the similarity between two motions based on the quantification of the joints movements. It assigns a percentage of similarity to each joint in each frame according to a calculated distance. The comparison algorithm can be applied to, e.g., applications that demand movement imitation.
3) *The evaluation results* comprehensively demonstrate the effectiveness of our motion quantification and, hence,

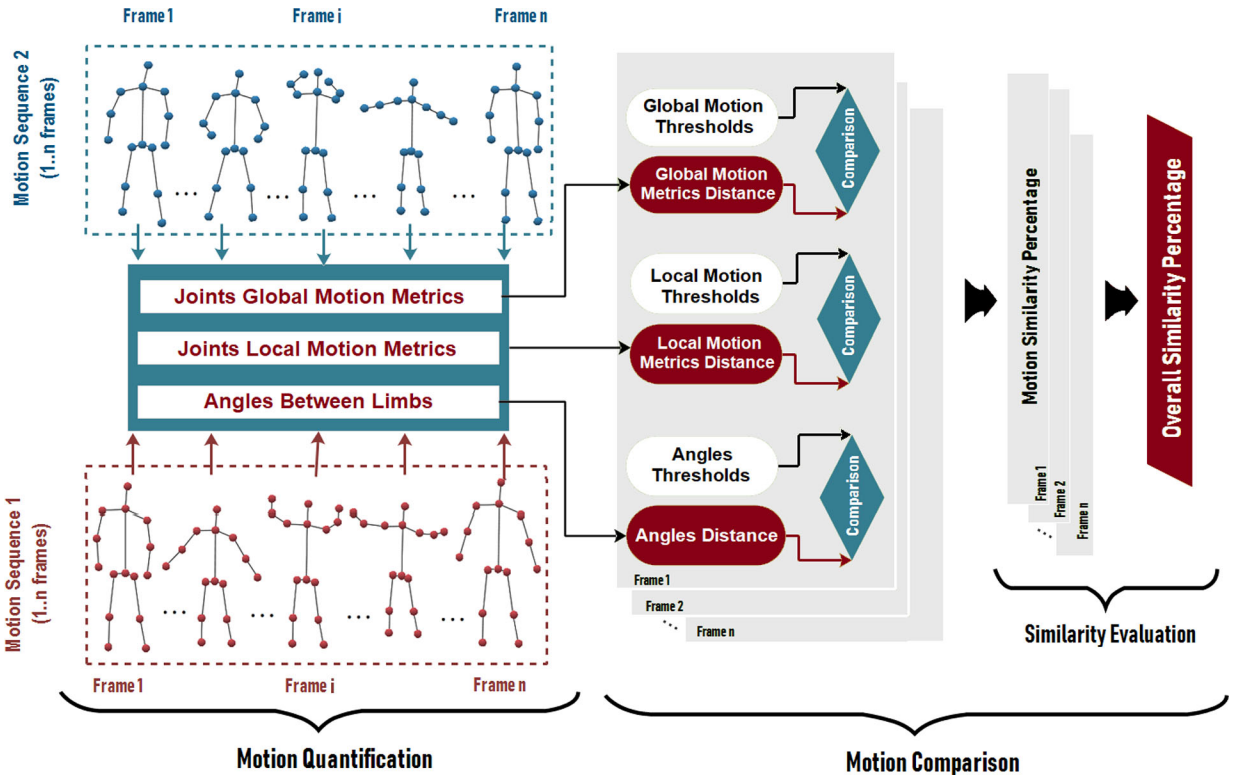


Fig. 2. Framework of the proposed motion quantification and comparison method. The motion metrics of the skeleton sequences are calculated (quantification), followed by a comparison between the two metrics according to given thresholds to evaluate the similarity between the two bodies movements.

comparison algorithm. Besides the evaluation on a benchmark dataset, a user study is carried out through recording sports movements and comparing the users' performance to test the effectiveness of the proposed method in real situations.

The rest of this paper is organized as follows. A review of the related work is presented in Section II. Technical details of our approach are given in Section III followed by the experimental results and user study in Section IV. The proposed work is concluded in Section V.

II. Related Work

In this section, we discuss some of the motion capture systems and methods used for generating 3-D joints coordinates, where our method can be applied. After that we review some of the recent human motion analysis methods followed by a discussion about motion comparison.

A. Motion Capture Systems

Motion capture systems vary by the type of devices used to capture the motion of the human body, including wearable devices and sensors, multiple cameras, a single camera, or involving multiple types of equipment together. Arsenault and Whitehead [25] presented a system composed of ten wearable inertial sensors connected in the form of a network to determine the change in the body movement. Zhang and Zhang [26] proposed an inexpensive human motion capture system prototype using a flexible architecture and distributed computing technology. Tao *et al.* [27] proposed a wearable motion capture microsensor system to capture the motion in real time, then a 3-D model of the human body is employed to reconstruct the motion. The introduction of time-of-flight cameras provided depth cues of the human body, which makes using wearable devices less required for capturing the human motion. Later on, depth cameras became the basis of several proposed motion capture systems. Ganapathi *et al.* [28] proposed an algorithm involves a generative model and a discriminative model to filter a stream of monocular depth images in order to capture the human motion at the frame rate from a single depth camera. Microsoft Kinect [19] is a popular depth sensor that can capture the human motion by generating an accurate body joints from a single depth image using the algorithm proposed in [20]. A matching algorithm between the depth map and 3-D full-body point cloud is proposed in [29] for accurate pose estimation from a single depth image. Multiple depth cameras were exploited in [30] to improve the human body pose recognition using different views. Their method uses multiple depth maps as inputs of a classifier to identify body part region through segmentation, then all the views are merged in a single 3-D point clouds to estimate the body pose. Shuai *et al.* [31] involved multiple Kinect depth cameras to capture the point cloud of the human body from different views with the help of a designed ellipsoid-based skeleton to capture the geometry details of the tracked body. The fusion of multisensors for capturing human motion is presented in [32]. They combined video data with a small number of inertial sensors to overcome the weaknesses of using one type of data.

The recent development in machine learning field motivated researchers to work on capturing human motion from

a single RGB image via estimating the body parts position using trained models with large datasets. Even though pose estimation from a single RGB image is a challenging task due to the lack of depth cues, the complexity of the background, clothing color, and the uncountable possible position of body parts, many recent successful approaches can estimate 3-D body poses from a single RGB image, which makes capturing the human motion easier and cheaper than previous approaches that involve depth cameras and wearable sensors. Moreno-Noguer [33] proposed a human pose estimation method using a single RGB image. The method infers 3-D poses from two-dimensional (2-D) poses, then improves pose estimation precision by representing the problem as a learning 2-D to 3-D human poses using a distance matrix for regression. Predicting 3-D human poses from multiview RGB images is proposed in [34] to avoid using large dataset for training. VNect [21] is a recent successful method that captures 3-D human poses in real time from a single RGB image. The method jointly predicts 2-D and 3-D human body poses using CNNs, then a skeleton kinematic model is used for fitting 2-D to 3-D poses to generate stable angles between joints.

B. Motion Analysis

The development in motion capture systems opened the way to exploit the data for motion analysis and understanding human behavior. A variety of algorithms based on machine learning have been proposed to recognize the human actions from a 3-D skeleton model. Gaglio *et al.* [22] exploited the joints detected by Microsoft Kinect for activity recognition using three machine learning algorithms: 1) K-means clustering to find the joints involved in the activity; 2) support vector machines (SVMs) for postures refinement; and 3) hidden Markov models (HMMs) to model human activities. Du *et al.* [35] transformed the skeleton joints coordinates to an RGB image where each of the three channels of the image represents the motion features along one of the Cartesian coordinates axes. In [23], the skeleton of the human body is divided into five parts, each part is used as an input to five recurrent neural networks which are fused together to classify human actions. Wang *et al.* [24], [36] transformed the skeleton joints trajectories shapes from 3-D space into three images to represent the front view, the top view, and the side view of the joints trajectories shapes. The three views are given to a CNN model for action classification.

To the best of our knowledge, there are very limited proposed approaches for motion quantification or comparison. Chen and Koskela [37] proposed an alignment method between two motion sequences using dynamic time warping, one sequence was recorded by the Kinect and the other sequence was recorded by another motion capture system. Their method calculates a similarity matrix between the two sequences to find the minimum distance for alignment. In [38], a comparison between human and robot motions based on the skeleton data is proposed. The movement of both human and robot are modeled by an HMM. However, their method works only on limited types of movements learned by the model.

III. Method

A. Overview

The challenge facing human motion analysis and evaluation is the lack of a strong representation that can provide a distinctive description of each movement even though when the difference in motion is small. In spite of the fact that the previous approaches mentioned in Section II tried to recognize a limited number of human actions by classification based on predefined movements, the human motion needs a detailed representation to model any type of movement, which is the key factors for many human–machine interaction applications. From the literature of the state-of-the-art motion capture systems and motion analysis methods, we found that the best data type to represent the motion of the human body is a skeleton model of 3-D joints since it provides precise information about the body movement. In this paper, we use the skeleton model presented in Fig. 1(a) for motion analysis and experiments. We chose 15 essential joints which are enough to cover informative details of the human motion. First, we propose an algorithm to quantify the motion based on metrics calculated from the 3-D joints. Second, in order to test the performance of the quantification algorithm, we propose a comparison algorithm between two motions based on the quantification metrics to evaluate the motion similarity between two existing movements.

B. Motion Quantification

There are unlimited possibilities of the joints movement direction because of the articulated nature of the human body, which makes taking the changing in 3-D Cartesian coordinates of the joints as a factor for motion quantification is not enough. As a matter of fact, two similar motions may have different Cartesian coordinates due to the large numerical space of the coordinates values. Even though when two motions have the same joints coordinates, we do not know whether a joint is moving by itself or under the influence of other joints. Furthermore, the movement direction is unknown. We propose a set of metrics that analyze the skeleton joints motion from many aspects for the sake of providing a distinctive representation of each motion, including the rotation and the translation of the joints according to the whole body and according to its parent joints. We also consider the angles between limbs as an essential metric to gather information about the joints direction. The only input data required to apply the quantification algorithm are the raw Cartesian coordinates of the 15 joints [Fig. 1(a)], then our proposed algorithm will estimate the motion metrics automatically.

The mathematic geometric representation of any point displacement in 3-D space can be defined by the rotation and the translation in the form of a transformation matrix. We adopt this mathematic transformation to represent the movement of the joints from frame to another during the motion. The rotation is defined as the changing in the angles around the three axes of the Cartesian coordinates system, and the translation is defined as the changing in the distance along the three axes. We use both metrics to represent the joints motion in 3-D coordinates system. In order to make the hip center the origin

Algorithm 1 Motion Quantification

function QUANTIFY (*Sk*, *N_frms*)
Sk: *Skeleton joints*
Gr, Gt: *Global Rotation and Global Translation*
GM, LM: *Global and Local transformation Matrices*
  for $i = 1$ to *N_frms* **do**
    for $j = 1$ to 15 **do**
      $[Gr(i,j), Gt(i,j)] = \textbf{Kabsch}([Sk_{(i-1,j)}, Sk_{(i,j)}, Sk_{(i+1,j)}])$
      $GM(i,j) = [Gr(i,j) \quad Gt(i,j) \quad [0\ 0\ 0\ 1]^t]$
      $[G\_\theta_x(i,j), G\_\theta_y(i,j), G\_\theta_z(i,j)] = \textbf{Eurler}(Gr(i,j))$
      $LM(i,j) = GM(i,j) \times [LM(i,p_1) \times \ldots \times LM(i,1)]^{-1}$
      $[Lr(i,j), Lt(i,j)] = LM(i,j)$
      $[L\_\theta_x(i,j), L\_\theta_y(i,j), L\_\theta_z(i,j)] = \textbf{Eurler}(Lr(i,j))$
      $G\_Motion(i,j) = [G\_\theta_x(i,j), G\_\theta_y(i,j), G\_\theta_z(i,j), Gt(i,j)]$
      $L\_Motion(i,j) = [L\_\theta_x(i,j), L\_\theta_y(i,j), L\_\theta_z(i,j), Lt(i,j)]$
    end for
    for $k = 1$ to 8 **do**
      $Angles(k) = \textbf{CalcAngle}(Sk(i,k), Sk(i,pr), Sk(i,ch))$
    end for
  end for
  return *G_Motion*, *L_Motion*, *Angles*
end function

of the coordinate system, we normalize the joints coordinates by subtracting the coordinates of the hip center $j1$ (Fig. 1) from their coordinates values. As we previously mentioned, we categorize the motion into global and local. The global motion is evaluated according to the hip center joint, and the local motion is evaluated by taking the first parent joint as the origin of the system. Besides the local and the global motion metrics, we also consider the angles between skeleton limbs as an essential factor for analysis. The angles between limbs provide critical information about the joints direction by knowing how the joints are moving from each other during the motion. Algorithm 1 (motion quantification) presents a detailed illustration of the motion quantification. In the following sections, we illustrate in details the quantification method.

1) Global Motion (Global Transformation Matrix): We define the global motion as the joints movement by considering the hip center joint as the origin of the coordinates system. For the global motion, we only care about how the joints move around the origin without taking into consideration the movement of its parent joints. The rotation and the translation of a joint can be deduced from the transformation matrix. However, the only motion data that we have are the joints coordinates in each frame. We propose an approach to estimate the rotation and the translation of the joints from only the joints coordinates by using the *Kabsch* [39] *algorithm*, which was proposed to estimate the rotation and the translation between two groups of 3-D points given the start and the end coordinates of their displacement, where all the points are assumed to be moved with the same rotation and translation. Each group must contain at least two points. However, during the movement of the human body, the joints do not move in the same way due to the articulated skeleton structure, which

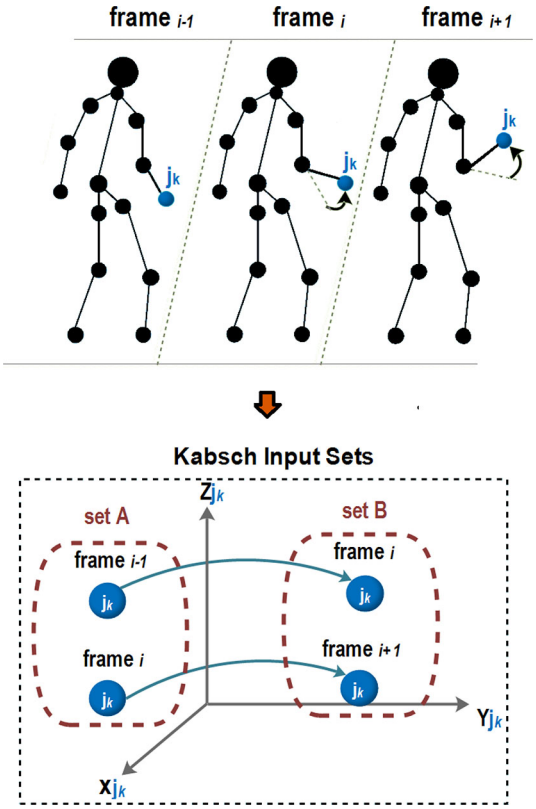


Fig. 3. Two sets (A and B) formed from the same joint to determine its transformation using the Kabsch algorithm. The two sets represent the movement of the same joint from the frame $i-1$ to the frame i and from the frame i to the frame $i+1$.

makes choosing a group of joints that move from a frame i to a frame $i+1$ is impossible.

Since the change in the joints coordinates from a frame to another during the motion is unnoticeable with a high frame rate, we suppose that any joint displacement from the frame $i-1$ to the frame i and from the frame i to the frame $i+1$ has the same transformation matrix. We use the coordinates of each joint to create two sets of its displacement from one frame to another to determine its transformation matrix using the Kabsch algorithm. Each set contains two points. The first set A contains the coordinates of the joint j in the frames $i-1$ and i. The second set B contains the coordinates of the joint j in the frames i and $i+1$. Fig. 3 illustrates the formation of the two sets used to apply the Kabsch algorithm, and (1) defines the problem of determining the rotation and the translation of a joint (point) X of a set A to its corresponding coordinates Y of a set B. Equations (2)–(6) illustrate the steps of the Kabsch algorithm to calculate the rotation matrix R and the translation vector T, and hence, the transformation matrix. The goal of this calculation is to get the global rotation angles $(\theta_x, \theta_y, \theta_z)$ around the three axes, and the global translation (t_x, t_y, t_z) along the three axes (last column of the transformation matrix). In spite of the fact that we involved the frame $i-1$ in the calculation to apply the Kabsch algorithm, our objective is to determine the joint movement from the frame i to the frame $i+1$

$$X = \mathbf{R} \times Y + \mathbf{T} \tag{1}$$

where R is the rotation matrix and T is the translation vector. X is the starting point of the movement and Y is the ending point of the movement

$$\begin{aligned} \text{Set}_A &= \left\{P_A^1, P_A^2, \ldots, P_A^N\right\} \\ \text{Set}_B &= \left\{P_B^1, P_B^2, \ldots, P_B^N\right\} \end{aligned} \tag{2}$$

$$\text{Cent}_A = \frac{1}{N}\sum_{i=1}^{N} P_A^i \quad \text{Cent}_B = \frac{1}{N}\sum_{i=1}^{N} P_B^i \tag{3}$$

where Set_A and Set_B are the starting and the ending sets of 3-D joints. P_A^j is the coordinates of the joint j in the starting frame and P_B^j represents its corresponding joint coordinates in the ending frame. In our case, we use only sets of two joints coordinates (Fig. 3), where in our case $N = 2$ (two joints). Cent_A and Cent_B are the centroid of Set_A and Set_B, respectively

$$H = \sum_{i=1}^{N} \left(P_A^i - \text{Cent}_A\right)\left(P_B^i - \text{Cent}_B\right)^t \tag{4}$$

$$[U, S, V] = \text{SVD}(H) \qquad \boldsymbol{R} = VU^t \tag{5}$$

$$\boldsymbol{T} = -R \times \text{Cent}_A + \text{Cent}_B \tag{6}$$

where H is the covariance matrix. SVD is the *SingularValueDecomposition* function that factorizes the matrix H into three matrices ($U, S,$ and V). The rotation matrix R is calculated from the matrices U and V. T is the translation vector where $T = [t_x \quad t_y \quad t_z]^t$. It represents the movement of the points along the axes X, Y, and Z

$$\mathbf{TF} = \begin{bmatrix} & & & t_x \\ & \boldsymbol{R} & & t_y \\ & & & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7}$$

where TF is the transformation matrix of the joints movement from Set_A to Set_B

$$\theta_x, \theta_y, \theta_z = \text{Euler}(R) \tag{8}$$

where Euler is a function calculates the rotation angles θ_x, θ_y, and θ_z around the axes X, Y, and Z, respectively, from the rotation matrix R.

2) Local Motion (Local Transformation Matrix): We define the local motion as the movement of each joint according to its parent joint. The calculation of the local motion metrics is based on taking the parent joint as the origin instead of the hip center joint. Fig. 1(b) illustrates the difference between the global and the local joint metrics. For example, the local motion of the joint $j4$ in the figure is calculated according to its parent joint $j3$ with the coordinate system X_{j3}, Y_{j3}, and Z_{j3}. However, the global motion of the joint $j4$ is calculated according to the joint $j1$ (hip center) with the coordinate system X_{j1}, Y_{j1}, and Z_{j1}. Same case for the joint $j3$ with its local motion according to its parent joint $j2$ and the global motion according to the joint $j1$. The case of the joint $j2$ is a little different because its parent joint is the same as the origin joint, which makes the local motion metrics are the same as the global metrics.

The global transformation matrix of a joint can be written as the product of the local transformations of its parent joints in the hierarchy sequence until the hip center joint. In order to obtain the local transformation matrix of a joint, we multiply the global transformation (calculated previously) and the inverse of the product of the local transformations of its parent joints. After calculating the local transformation matrix, we get the local rotation $(\theta_x, \theta_y, \theta_z)$ and translation metrics (t_x, t_y, t_z) using (7) and (8)

$$M_{\text{global}(j)} = \prod M_{\text{local}(j)} \cdot M_{\text{local}(j-1)} \cdot \cdots \cdot M_{\text{local}(j_1)} \tag{9}$$

$$M_{\text{local}(j)} = M_{\text{global}(j)} \cdot \left[M_{\text{local}(j-1)} \cdot \cdots \cdot M_{\text{local}(j_1)}\right]^{-1} \tag{10}$$

where $M_{\text{global}(j)}$ is the global transformation matrix of the joint j, and $M_{\text{local}(j)}$ is the local transformation matrix of the joint j. $M_{\text{local}(j-1)} \cdot \cdots \cdot M_{\text{local}(j_1)}$ are the local transformation matrices of the parent joints from the joint $j-1$ to the origin j_1 (hip center) in the hierarchy.

3) Angles Between Limbs: The angles between limbs determine the direction of the joints by checking the size of the angles in each frame, which we consider as an essential parameter in the motion quantification that provides information cannot be extracted from the transformation matrix. Fig. 1(b) shows the eight angles $(a1, a2, \ldots, a8)$ selected to evaluate the motion between limbs. For example, the change in the angle $a3$ (left leg) provide clues about the direction of the joint $j4$. When the angle $a3$ getting larger, the direction of the joint $j4$ is toward the bottom, and when the angle getting smaller, the joint direction is toward the top.

We use the law of cosines to calculate the angle of a triangle formed by the limbs based on the coordinates of its vertices. In our case, each of the eight angles is considered as one of the vertices of the triangle formed with the two adjacent joints. Fig. 1(b) shows the triangle formed by the joints $j2$–$j4$ to calculate the angle of the joint $j3$. Unlike the transformation matrix of the joints that are calculated between frames to model the displacement from a frame i to a frame $i+1$, the angles between limbs are calculated in each frame i

$$d_{(j_a,j_b)} = \|j(x_a, y_a, z_a) - j(x_b, y_b, z_b)\| \tag{11}$$

$$a3 = \text{Ang}(j_3) = \cos^{-1}\left(\frac{d^2_{(j_2,j_3)} + d^2_{(j_4,j_3)} - d^2_{(j_4,j_2)}}{2 * d_{(j_2,j_3)} * d_{(j_4,j_3)}}\right) \tag{12}$$

where $d_{(j_a,j_b)}$ is the Euclidian distance between two joints. $\text{Ang}(j_3)$ is the angle of the joint j_3 calculated from its two adjacent joints j_2 and j_4 [a_3 in Fig. 1(b)].

C. Motion Comparison

We propose a comparison algorithm between two skeletons motions to evaluate the performance of the quantification and provide an evaluation of motion similarity, which can be used for many human–computer interaction applications that demand imitating movements and automatic similarity evaluation. The comparison algorithm calculates the distance between two motions metrics, then evaluates their similarity depending on a given threshold for each metric. There are

Algorithm 2 Motion Comparison

Input: $Skel1$, $Skel2$, N_frms
Output: $Similarity\_Percentage$
Thresholds: Gth_θ, Lth_θ, Gth_t, Lth_t, $th\_angles = [tha1, tha2 \dots tha8]$
Percentages: $G\theta_\%$, $L\theta_\%$, $Gt_\%$, $Lt_\%$, $angles_\% = [a1_\%, a2_\% \dots a8_\%]$, $GFrame_\% = 0$, $LFrame_\% = 0$, $AngFrame_\% = 0$

$[G\_Motion1, L\_Motion1, Angles1] =$ **QUANTIFY**$(Skel1, Nfrms)$
$[G\_Motion2, L\_Motion2, Angles2] =$ **QUANTIFY**$(Skel2, Nfrms)$
for $i = 1$ to N_frms **do**
  for $j = 1$ to 15 **do**
    $[G\_d\theta, G\_dt] =$ **Distance**$(G\_Motion1(i,j), G\_Motion2(i,j))$
    $[L\_d\theta, L\_dt] =$ **Distance**$(L\_Motion1(i,j), L\_Motion2(i,j))$
    if $G\_d\theta \leq Gth_\theta$ **then** $GFrame_\% = GFrame_\% + G\theta_\%$ **end**
    if $G\_dt \leq Gth_t$ **then** $GFrame_\% = GFrame_\% + Gt_\%$ **end**
    if $L\_d\theta \leq Lth_\theta$ **then** $LFrame_\% = LFrame_\% + L\theta_\%$ **end**
    if $L\_dt \leq Lth_t$ **then** $LFrame_\% = LFrame_\% + Lt_\%$ **end**
  end for
  for $k = 1$ to 8 **do**
    $[da_1, da_2, \dots, da_8] =$ **Distance**$(Angles1(i,k), Angles2(i,k))$
    if $da_1 \leq tha_1$ **then** $AngFrame_\% = AngFrame_\% + a1_\%$ **end**
    ⋮
    if $da_8 \leq tha_8$ **then** $AngFrame_\% = AngFrame_\% + a8_\%$ **end**
  end for
  if $AngFrame_\% > 65\%$ **then**
    $Frame_\% = ((GFrame_\% + LFrame_\%) + AngFrame_\%)/2$
  else
    $Frame_\% = 0$
  end if
  $Sum\_Frames = Sum\_Frames + Frame_\%$
end for
$Similarity\_Percentage = Sum\_Frames \ / \ N\_frms$
return $Similarity\_Percentage$

three levels for evaluation, joints comparison, frames comparison, then whole motion comparison. While the evaluation is based on a frame with frame comparison, it requires that the two motions have the same number of frames. In case that the number of frames is different, the algorithm compares the sequence of smaller frames number with the first subsequence of the second sequence to check the similarity, then the first sequence is slid to the next subsequence by one frame and the motion similarity is checked again, etc. The subsequence that generates high similarity result is considered as the final results of the comparison. The superiority of our algorithm lies on checking in which part of the sequence exactly the motions are similar. For example, the sequence of larger frames may contain all the skeleton position of the sequence of smaller frames but they are distributed over the sequence. By comparing the same number of frames, we guaranty that all the positions must be consecutive in order to generate high similarity. The inputs of the comparison algorithm are the metrics of two skeleton motions generated from the quantification algorithm including, local and global joints transformations, and the angles between limbs. Algorithm 2 (motion comparison) gives a detailed illustration of the comparison process.

1) Distance Between the Two Motions: In order to compare the two motions, we calculate the distance between the skeletons metrics of the two sequences. The distance determines how different are the translations, rotations, and the angles between limbs of one skeleton from the other skeleton of the same joint at the same frame number. Large distance indicates that there is a large difference between the joints motions. The diversity of the metrics used for motion quantification are efficient for motion comparison. For instance, if the distance between the global rotation joints reveals that the joints movements are similar, the local rotation distance gives more proof whether the joints were moved in similar ways. The distance between the angles of the limbs of the two skeletons at each frame improves the comparison precision by checking the joints direction.

2) Threshold Comparison: It is impossible to find two motions that have exactly the same metrics because of the sensitivity of the joints to the change in the coordinates values. We set a threshold for each metric distance to allow an error margin for the comparison. During our experiments, the thresholds values were set according to the importance of the metrics and their impact weight on the whole motion. For example, we set the threshold of the global rotation greater than the local rotation, and the threshold of some angles of the limbs are set greater than others according to their sensitivity to the change during the motion. For example, the elbow angle $a6$ is more changeable than the shoulder angle $a5$ (Fig. 1). In Section IV, we show the influence of changing the thresholds on the motion comparison.

3) Percentage Similarity: The evaluation of the joints similarity is based on assigning a percentage to each joint to indicate its similarity rate. This rate is calculated based on giving a partial percentage to each of the joints metrics, including rotation angles, translation vectors, and limbs angels. The partial percentage is assigned to a metric if it is less than a threshold. We consider the sum of the partial percentages of the rotation angles and the translation vectors as the first part of the joint evaluation, and the partial percentage of the angles between limbs as the second part. The two parts are averaged to get the final similarity percentage of the joint. After that the evaluation of the whole skeleton motion in each frame is calculated by averaging the percentages of the 15 joints. In the end, the percentage similarity of the whole motion is the average of all frames percentages calculated previously (Algorithm 2).

IV. Experimental Results

We evaluated our quantification method using the comparison algorithm on skeleton motions of actions from the UTD-MHAD dataset, and we also performed an evaluation of motions on a dataset that we recorded using Microsoft Kinect V2 [19] for motion comparison and user study. In both cases, we consider only 15 joints presented in Fig. 1(a) for experiments. Also, a comparison on skeleton alignment with existing methods was also conducted.

A. UTD-MHAD Dataset

1) UTD-MHAD Results: UTD-MHAD dataset [40] was collected using Microsoft Kinect depth sensor and inertial wearable sensors for action recognition. The dataset contains 27 actions, namely right arm swipe to the left, right arm swipe to the right, right-hand wave, two hand front clap, right arm

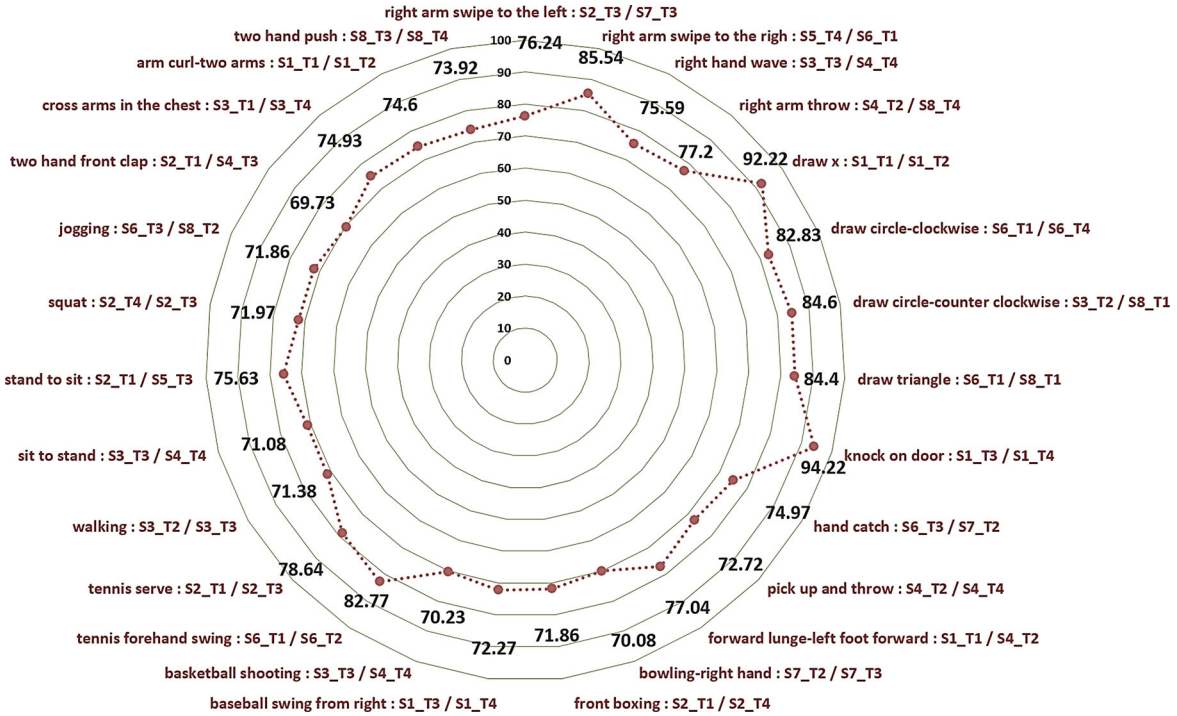


Fig. 4. Similarity comparison on pairs of similar actions types from the UTD-MHAD dataset. S: subject (user) and T: trial.

throw, cross arms in the chest, basketball shooting, right-hand draw *x*, right-hand draw circle (clockwise), right-hand draw circle (counter clockwise), draw triangle, bowling (right hand), front boxing, baseball swing from right, tennis right-hand forehand swing, arm curl (two arms), tennis serve, two hand push, right-hand knock on a door, right-hand catch an object, right-hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge (left foot forward), and squat (two arms stretch out). The actions have been performed by eight subjects, and each subject repeated the action four times.

In spite of the fact that the goal of the dataset is to be used to recognize human actions which is different from our goal, we exploit the fact that there are similar actions performed in different ways by different people to evaluate the motion similarity using our proposed method. Our evaluation method is based on choosing two similar actions performed by different subjects or by the same subject, then we apply the proposed algorithm to quantify and compare two actions. The results expected from the algorithm should generate a high similarity percentage for actions of the same type and a low percentage for different actions. Fig. 4 shows the evaluation of the comparison results of randomly selected pairs of subjects performed the same action. The evaluation results show a similarity percentage above 70% in most cases which is evidence that the proposed quantification and comparison algorithms can effectively represent and evaluate the human motion. The results vary from one pair to another depending on how the same action is performed. Many factors could influence the action performance, such as the speed of the subject, and the body parts direction during the movement. For example, the action: *hand catch* can be performed with an arm at the upper body side or at the lower body side which makes the comparison generates 74% of similarity percentage. However, the actions that are usually performed in similar ways where there is no possibility for a large difference in performance, such as: *knock on the door*, the algorithm generates a high similarity percentage (94.22%).

In order to test the performance of the algorithm when the motions are different, we selected 22 pairs of different actions types performed by different subjects. Fig. 5 shows a percentage similarity lower than 40% in all cases. This time also the algorithm shows the stability in evaluating different actions by generating a low similarity percentage. The lowest similarity percentage is generated for the actions: *walking—sit to stand* with 19.24% and the actions: *sit to stand—stand to sit* with 19.27%. We can notice that the actions are completely different during the skeleton motion which reflects the low similarity percentage. However, for the actions: *forward lunge (left foot)—squat*, the algorithm generates the highest similarity percentage (39.1%) because most of the body parts have common movement since both actions performed by bending down. The only difference is that the first action requires moving the leg forward while bending. Generally, the results generated from the proposed quantification and comparison algorithms presented in Fig. 4 and Fig. 5 shows that when the two motions performed with the same body parts in the same direction but with a little difference in the performance, the comparison algorithm generates high similarity percentage. However, when there is a difference in the movement direction even in only one body part, the comparison algorithm generates a low similarity percentage.

2) Threshold Influence on the UTD-MHAD Dataset Actions: Since the comparison between the two motions is

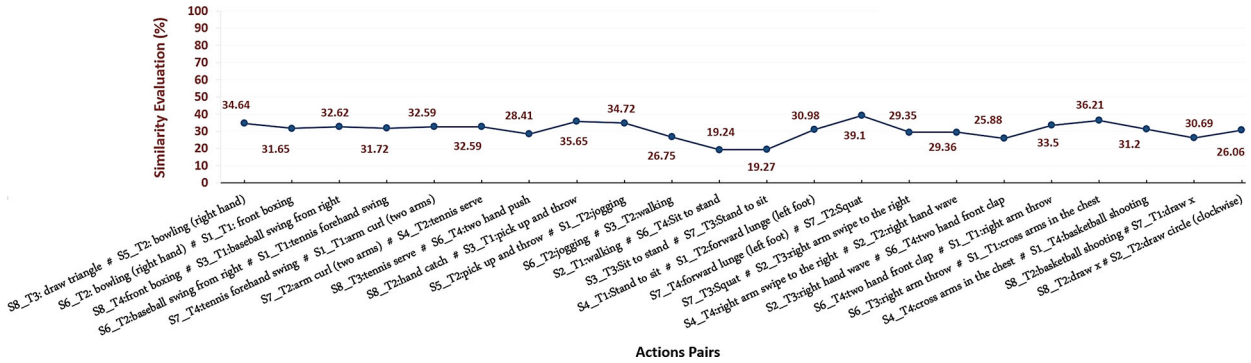


Fig. 5. Similarity comparison on pairs of different actions types from the UTD-MHAD dataset. S: subject and T: trial.

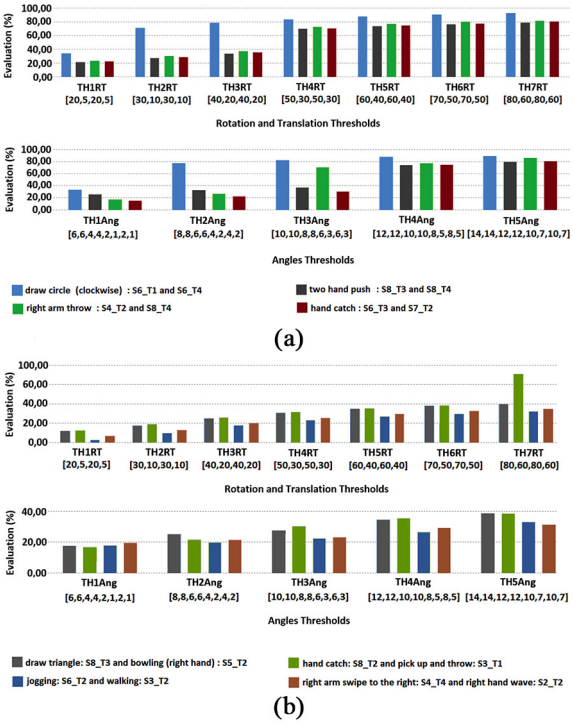


Fig. 6. Influence of the thresholds on the motion comparison using the UTD-MHAD dataset. On pairs of the (a) same actions and (b) different actions.

based on calculating the distance between the metrics, the judgment of the motion similarity is based on thresholds. In order to investigate the impact of the thresholds on the comparison, each time we change the thresholds and evaluate the comparison results. We define two lists of thresholds. The first list $THiRT = [Gr, Gt, Lr, Tt]$ represents the threshold list i of the rotations and translations. Gr, Gt, Lr, and Lt are the global and local rotation angles and translation vectors thresholds. The second list $THiAngles = [a1, a2, \ldots, a8]$ represents the thresholds list i of the angles $a1$–$a8$.

Fig. 6(a) shows the comparison results of four pairs of similar actions at different thresholds. In Fig. 6(a) top, we fix the angle thresholds at $THAngles = [12, 12, 10, 10, 8, 5, 8, 5]$ and we changed the $THiRT$ thresholds from $TH1RT$ to $TH7RT$, then we check the comparison results. We notice that from $TH4RT$ to $TH7RT$, the algorithm generates correct judgments for the four actions pairs. However, for the thresholds $TH2RT$ and $TH3RT$, only one pair of actions is judged correctly. The comparison results at $TH1RT$ are incorrect for all the four actions pairs. The reason is that the threshold is too small so that similar actions were judged as different actions when there is a small difference in motion. In Fig. 6(a) bottom, we fix the thresholds $THiRT$ at $THiRT = [40, 20, 40, 20]$ and we change the angles thresholds from $TH1Ang$ to $TH5Ang$, then we check the results again. At $TH4Ang$ and $TH5Ang$, the algorithm generates a high percentage similarity for the four actions, but at thresholds $TH2RT$ and $TH3RT$, only one pair of actions is evaluated as having similar motions. However, at $TH1Ang$, the algorithm generates a very low percentage similarity for the four actions because of the small values of the angles thresholds.

In Fig. 6(b), this time, we show the threshold impact on four pairs of different actions types. We repeat the same process using the same thresholds values as in Fig. 6(a). In Fig. 6(b) top, we fix the angles thresholds at $THiAng = [12, 12, 10, 10, 8, 5, 8, 5]$, then we change the thresholds $THiRT$. The algorithm generates correct judgment for the actions with a percentage of similarity less than 40% for all the four actions pairs from $TH1RT$ to $TH6RT$. Only one pair of actions is judged as similar at $TH7RT$ because of the large thresholds values. However, in Fig. 6(b) bottom, fixing the rotations and translations thresholds at $THiRT = [40, 20, 40, 20]$ and changing the angles threshold from $TH1Ang$ to $TH5Ang$ generates correct judgment with percentage similarity less than 40% for all the actions pairs at all the thresholds.

B. User Movements Comparison Study Using Kinect

1) Experimental Settings: Microsoft Kinect can accurately estimate 3-D human body joints from a single depth image [20], which makes it a useful device for human motion analysis. We recorded four sports movements performed with five users using Kinect V2 to evaluate the performance of the quantification algorithm and analyze its use in real situations with different users. The experiments settings are shown in Fig. 7. The user stands in front of the Kinect, then performs the movement from a predefined distance. The sports movements are basketball shooting, football kick, baseball bat swing, and tennis racquet swing. In fact, each of those sports movements can be performed in different ways. For example,

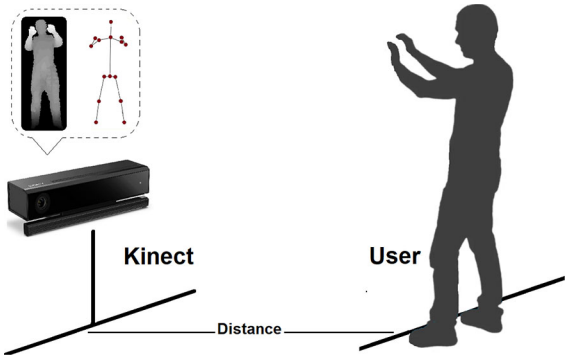


Fig. 7. Experimental settings to record users motion for user study.

the basketball shooting can be performed with jumping or without jumping. Each user performs the movement twice. One time from a distance of 1.8 m from the Kinect and another time from a distance of 3.6 m. In both cases, the user performs the movement in the same way. We asked the users to perform the movements in the way that they know, without teaching them how to perform them. We only require that the same movement must be performed with the same body parts for all the users. For example, the football kick movement must be performed with the right leg by each user. The reason behind such a restriction is that we want to test the performance of the algorithms in evaluating the motion similarity for the applications that require movements imitation, such as sports movements learning. During our analysis of the users' motions, we check factors that may influence the motion similarity besides the joints movement direction. We calculate the motion similarity from two different distances, with users of slightly different heights, and with different speed of movements. The five users are named as *User*1, *User*2, . . . , *User*5, which have the heights 1.7, 1.6, 1.78, 1.73, and 1.78 m, respectively. We define the speed of the motion as the number of frames required by the user to complete the movement. Each pair of users performs the same movement, then we compare their performance and check the influence of the speed, the distance, and the height.

2) Movement Comparison: Here, we categorize the movements by four sports types. For each category, we show the comparison results of each pair of users. The similarity generated from our algorithm in the following experiments is based on a threshold that was set according to our desired level of similarity after a set of initial trials. Generally, a similarity value between 50% and 60% is generated when the movements have common motion. When the similarity is getting larger above 60%, it means that the movements have more common body parts that move in similar directions with almost the same angles between limbs during the motion. Since the visual judgment in our method plays an important role, and in order to show the performance of the user compared to the similarity generated from the algorithm, in Fig. 8, for each user movement, we show the depth frames that indicate a clear change in motion.

1) *Basketball Shooting:* Table I (basketball shooting) shows the similarity comparison of four pairs performed the basketball movement which are *User*3–*User*5, and Fig. 8 (basketball shooting) shows the performance of the three users from 1.8 m of the camera distance. We notice that the performance of *User*3 is close to *User*4. Both users imitated shooting the ball without jumping. However, *User*5 imitated the movement with different body position and with jumping while shooting. The *Pair*1 includes the movements of *User*3 and *User*4 from the same distance with a difference in speed of nine frames, a difference in height of 5 cm, and a small difference in their positions while performing the movement. While there is no big difference between the conditions of the two performers, the similarity between them is judged as 72.08%. In *Pair*2, the same user performed the movement twice in the same way presented in Fig. 8 (basketball shooting) from short and long distances with a difference in speed of eight frames. This time the algorithm also generates a high similarity percentage of 88.81%. Although the distance from the camera is different, the similarity still accurate. A difference in speed of 19 frames between the movements of the *Pair*3 generates a similarity of 60.4% and a difference of 27 frames for *Pair*4 generates a similarity of 44.68. The reason for the low similarity generated from the algorithm is that there is a big difference between their performance. While the algorithm is based on comparing frames of the same number of order, the difference in speed means that similar frames are not in the same order, and hence, the movements are evaluated as different in performance.

TABLE I
SIMILARITY COMPARISON ON THE FOUR SPORTS MOVEMENTS

Pair	User	Height (m)	Speed (frames)	Distance (m)	Similarity (%)
Basketball Shooting					
1	User3	1.78	44	1,8	72.08
	User4	1.73	35	1,8	
2	User3	1.78	44	1,8	88.81
	User3	1.78	36	3.6	
3	User5	1.78	54	3.6	60.4
	User4	1.73	35	3.6	
4	User5	1.78	62	1,8	44.68
	User4	1.73	35	1,8	
Football Kick					
1	User4	1.73	47	1,8	75.27
	User4	1.73	43	3.6	
2	User5	1.78	39	1,8	45.75
	User2	1.6	78	1,8	
Baseball Bat Swing					
1	User1	1.7	139	1,8	65.77
	User3	1.78	115	1,8	
Tennis Racquet Swing					
1	User1	1.7	61	1,8	72.11
	User1	1.7	53	3.6	
2	User3	1.78	45	3.6	60
	User1	1.7	45	3.6	

Fig. 8. Key depth frames of the four sports movements used for user study.

2) *Football Kick:* The evaluation of the football kick movements is presented in Table I (football kick), and the performance of the two users involved in the comparison is shown in Fig. 8 (football kick). *Pair*1 represents the comparison of the same user (*User*4) performed the movement from different distances with a small difference in speed of four frames, which leads to generate a similarity of 75.27%. As we previously have seen with the basketball movements, the distance from the camera does not influence the motion similarity, we can conclude that only the performance of the movement influences the motion similarity in this case, given that there is a small difference in the speed. The *Pair*2 includes a comparison between *User*5 and *User*2 with a difference in height of 18 cm, from the same camera distance and with a large difference in speed (39 frames). By looking at how the movements are performed, we can say that the low similarity generated from the algorithms (45.74%) is due to the difference in the performance and a large difference in the speed.
3) *Baseball Bat Swing:* In Table I (baseball bat swing), we show a comparison of a pair of users performed the baseball bat movement from the same short camera distance with a relatively small difference in height. The user's key body positions are shown in Fig. 8 (baseball bat swing). By looking at the depth frames, we notice that the movements are different at the beginning and at the end of the two sequences. However, in the rest of the frames, the movements look similar according to the visual observation. The large difference in speed (24 frames) and performance leads to a similarity value of 65.77%.
4) *Tennis Racquet Swing:* Table I (tennis racquet swing) shows comparison results of two pairs performing the tennis racquet swing. *Pair*1 shows a motion similarity of 72.11% for the same user (*User*1) performed the movement from long and short distances with a difference in speed of nine frames. The key body positions of *User*3's performance are shown in Fig. 8 (tennis racquet swing). The second pair (*Pair*2) shows the results of the *User*1 and the *User*2 performed the movement from different distances with the same speed of 45 frames. The position of the bodies during the movement of the two users are different from each other. *User*1 used two hands while swinging, but *User*3 used only his right-hand while moving his body from the right side to the front of the camera. The similarity between the two users was judged as 60% since they have the same speed but different performance. The big difference lies in using one hand by *User*1 and two hands by *User*3.

Fig. 9 presents the 3-D trajectories of the users' joints motion performing the sports movements, including the boxing movement, and labeled with the speed and the camera distance. The global shape of all the joints trajectories looks similar for the same sport type, whereas due to the sensitivity of the joints to the motion change, rarely when we can have exactly similar trajectories shapes and size even when the depth frames appearance in Fig. 8 look similar.

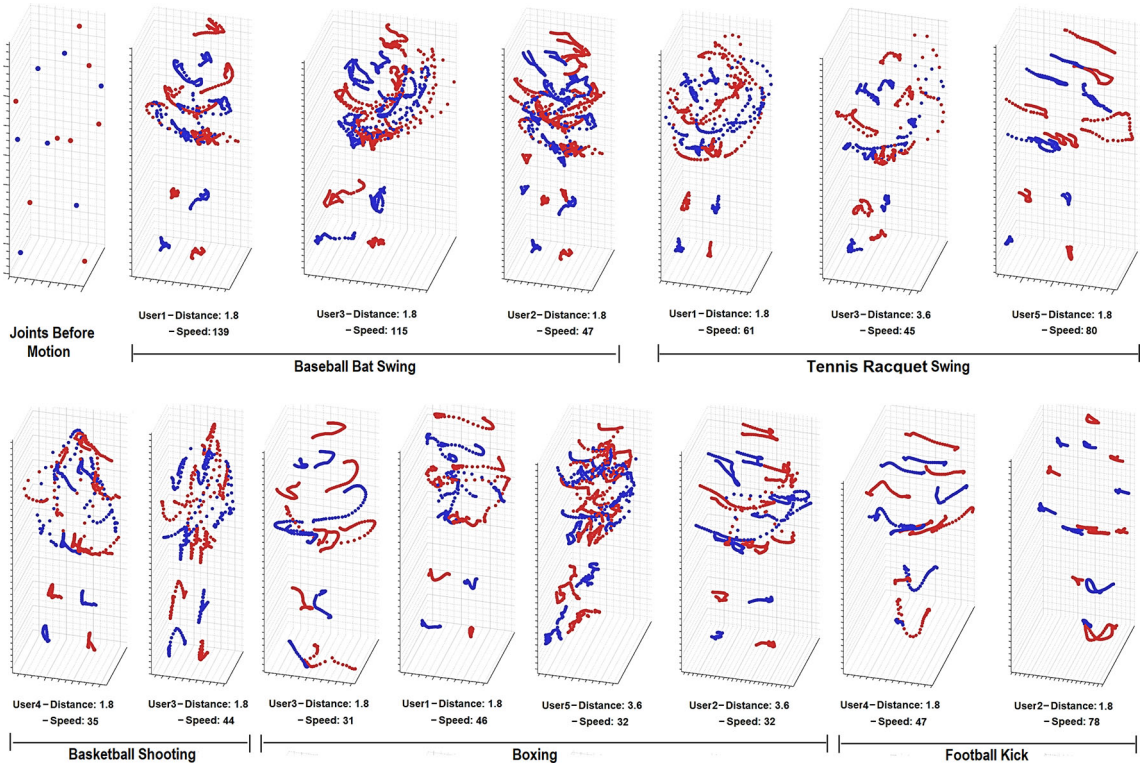


Fig. 9. Skeleton motions trajectories of some users mentioned in Table I and Fig. 8.

3) Influential Factors: In the previous analysis of the user study, we considered some factors (user height, camera distance, and speed) to check their impact on the motion comparison using the proposed quantification. By observing the results of the comparison, we conclude that the height and the distance from the camera do not have an impact on the motion comparison. As we have already seen, some of the pairs performed the movement from different distances in the same way but the comparison algorithm generates high similarity percentage. Same case for the height, where the algorithm generates a high similarity percentage regardless of the difference in height. A difference in height and distance means a difference in the joints coordinates values. While the proposed quantification algorithm focuses on how the joints move and not where the joints are, the height and the distance do not influence the comparison. However, the speed has a big influence on the comparison as it is related to user performance. Our algorithm considers the comparison between every two frames of the same rank, which generates low similarity percentage for the frames when the difference in speed is large because the same body position of the first user could be found after the next n frames of the second user movement. Overall, we notice that when the speed increases, the motion similarity decreases in both cases of similar and different motions. However, in some cases, when the motions have a very high level of similarity, the difference in speed of fewer than ten frames does not affect the comparison. Finally, the main factor that influences the motion similarity is how the movement is performed. If a user performs the movement with different body parts from the other user or in different trajectory direction, the comparison algorithm will generate a low similarity percentage even when the two users have the same movement speed.

4) Thresholds Adjustment: Because of the sensitivity of joints coordinates change in 3-D space, it is almost impossible to find two motions that have exactly the same metrics values even though when they look exactly the same according to the visual observation. A threshold is a necessity to allow an interval of error. Therefore, the adjustment of the threshold is very important for motion comparison. As we previously have seen with the evaluation of similar and different actions on the UTD-MHAD dataset, different thresholds generate different results. When we want to consider two motions as similar even though there is a small difference in the performance, large thresholds must be used. But if we want the comparison algorithm to be very strict in the evaluation, we must use small thresholds. For example, the results generated with the previous user study of the sports movements are generated by adjusting the thresholds each time, then we check the results of the algorithm with visual observation. The thresholds used for comparing the sports movements require larger values than the thresholds of UTD-MHAD dataset. Depending on the level of similarity we are looking for, with other types of movements, the algorithm may need larger or smaller threshold values. In practice, choosing the threshold depend on the experts of the field where the comparison algorithm is applied. For example, applying the algorithm to learn a user practicing some movements, such as martial art sports movement based on existing saved movement for comparison, requires the intervention of a coach to decide at what level the comparison should be strict to set the right threshold. Setting the right threshold requires initially a set of trails of different

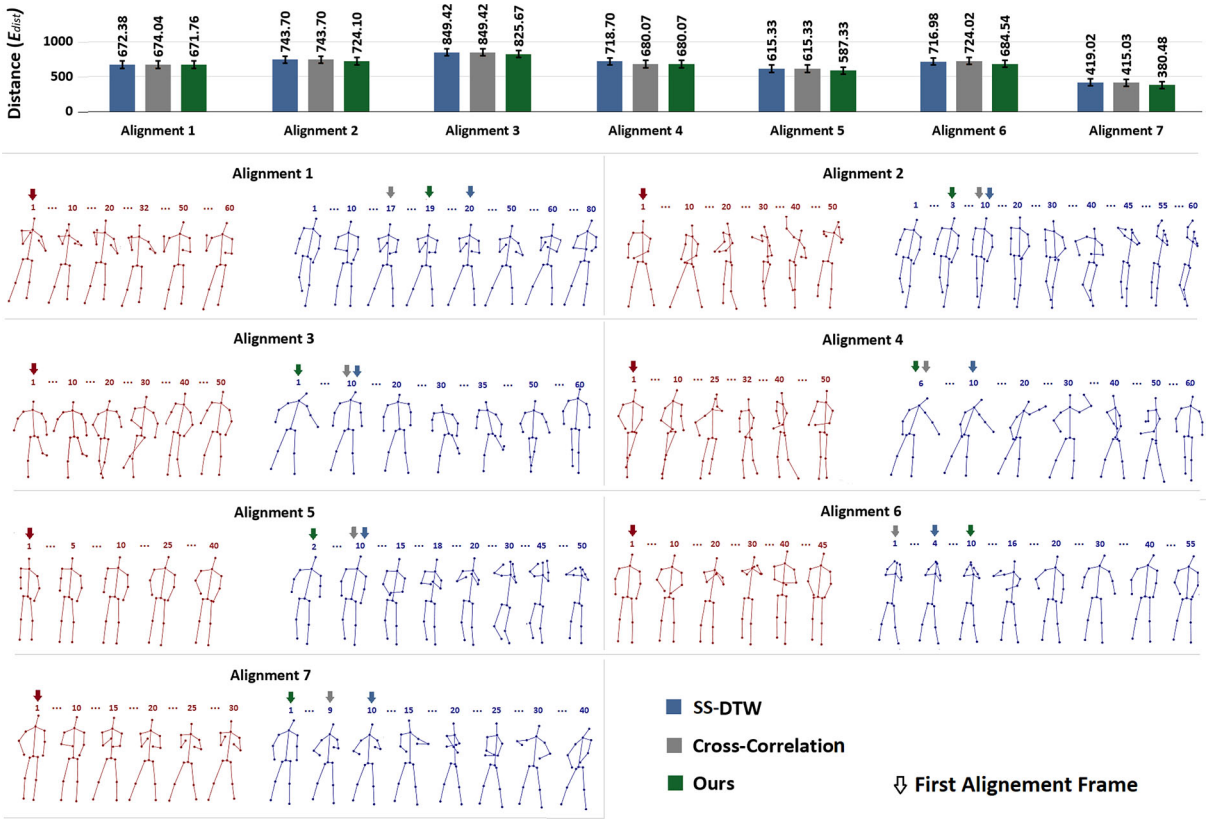


Fig. 10. Alignment results of seven pairs of skeletons motions generated by SS-DTW, cross-correlation, and our algorithm. The arrows indicate the first alignment frame of the skeleton sequence in red color with the sequence in blue color.

movements to check the level of similarity each time until getting the desired similarity results.

C. Comparison With Existing Methods

According to our knowledge, the closely related work to ours is the work proposed in [37]. They presented an algorithm to align two sequences of skeletons using subsequence dynamic time warping (SS-DTW), which is a revised version of DTW. It was used to measure the similarity between sequences vary in speed and length. The measurement is based on calculating the distance matrix between the two sequences and choose the path that corresponds to the minimum distance. Following their experimental settings, we also compare our results with other techniques by using the Euclidian distance (E_{dist}) [37] metric to measure the difference between the aligned sequences as ground truth for evaluation. Our comparison was performed with SS-DTW and cross-correlation [37] to evaluate the alignment accuracy of the seven pairs of skeletons motion presented in Fig. 10. The motions were extracted from the sports movements in a way that the large sequence (in blue color) has to contain some of the similar frames of the smaller sequence (red color). Due to the limited space, we show only the key frames. Each time we slide the smaller sequence by one frame, then we calculate the SS-DTW, cross-correlation, and the similarity between the two sequences using our comparison algorithm. With each slide, we obtain distance values of SS-DTW and cross-correlation, and also a similarity value from our algorithm. The sliding process finishes when the last frames of the two sequences are aligned. The first frame index of the best alignment for each method is saved, then the E_{dist} is measured to evaluate the three algorithms decisions by calculating the distance between the sequences based on the previously saved frame index. The algorithm which has the minimum distance is considered to have the best alignment. The results of the comparison are presented in Fig. 10. Our algorithm generates distances with less or equal values to the other methods in all cases, which means that the alignment decided by our approach is the best one. The figure also shows the first alignment frame decided by each method in the seven samples.

V. Conclusion

A method for human motion quantification and comparison has been proposed in this paper. The proposed method showed a novel algorithm based on estimating motion metrics to model the human movement based on the 3-D joints coordinates. This paper tried to solve a challenging problem in the field of human motion analysis by proposing a set of metrics that quantify the human joints movement based on the rotation and the translation of the joints, and the angles between limbs. The main advantage of the proposed method is that it can estimate the motion from only 3-D Cartesian coordinates of the body joints without prior knowledge about the movement parameters. In order to test the effectiveness of the motion quantification algorithm, we also proposed a comparison algorithm to evaluate the similarity between the two

motions. The overall motion evaluation results and user study showed that on the one hand, our motion quantification algorithm can effectively model the human movement, and on the other hand, using the quantification with the comparison algorithm is efficient to judge the similarity between two human movements. Furthermore, the flexibility of adjusting the comparison threshold allows the proposed technique in this paper to be used for many types of applications.

REFERENCES

[1] G. C. Hapsari and A. S. Prabuwono, "Human motion recognition in real-time surveillance system: A review," *J. Appl. Sci.*, vol. 10, no. 22, pp. 2793–2798, 2010.

[2] Y. Nie, C. Xiao, H. Sun, and P. Li, "Compact video synopsis via global spatiotemporal optimization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 10, pp. 1664–1676, Oct. 2013.

[3] Y. Nie, H. Sun, P. Li, C. Xiao, and K.-L. Ma, "Object movements synopsis via part assembling and stitching," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 9, pp. 1303–1315, Sep. 2014.

[4] J. Schrammel, L. Paletta, and M. Tscheligi, "Exploring the possibilities of body motion data for human computer interaction research," in *Proc. Symp. Austrian HCI Usability Eng. Group*, 2010, pp. 305–317.

[5] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[6] D. Lin, R. Zhang, Y. Ji, P. Li, and H. Huang, "SCN: Switchable context network for semantic segmentation of RGB-D images," *IEEE Trans. Cybern.*, to be published.

[7] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Büla, and P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 711–723, Jun. 2003.

[8] B. Sheng *et al.*, "Retinal vessel segmentation using minimum spanning superpixel tree detector," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2707–2719, Jul. 2019.

[9] B. Zhang, S. Jiang, K. Yan, and D. Wei, "Human walking analysis, evaluation and classification based on motion capture system," in *Health Management: Different Approaches and Solutions*, K. Smigorski, Ed. Rijeka, Croatia: IntechOpen, 2011, ch. 20, pp. 361–398.

[10] E. Gianaria, M. Grangetto, M. Lucenteforte, and N. Balossino, "Human classification using gait features," in *Proc. Int. Workshop Biometric Authentication*, 2014, pp. 16–27.

[11] N. García, J. Rosell, and R. Suárez, "Motion planning by demonstration with human-likeness evaluation for dual-arm robots," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[12] B. Brahmi, M. Saad, M. H. Rahman, and C. Ochoa-Luna, "Cartesian trajectory tracking of a 7-DOF exoskeleton robot based on human inverse kinematics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 3, pp. 600–611, Mar. 2019.

[13] Y. Tao, Y. Shen, B. Sheng, P. Li, and R. W. H. Lau, "Video decolorization using visual proximity coherence optimization," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1406–1419, May 2018.

[14] B. Huang, Z. Li, X. Wu, A. Ajoudani, A. Bicchi, and J. Liu, "Coordination control of a dual-arm exoskeleton robot using human impedance transfer skills," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 5, pp. 954–963, Mar. 2019.

[15] J. Hwang, J. Kim, A. Ahmadi, M. Choi, and J. Tani, "Dealing with large-scale spatio-temporal patterns in imitative interaction between a robot and a human by using the predictive coding framework," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[16] C.-L. Hwang and G.-H. Liao, "Real-time pose imitation by mid-size humanoid robot with servo-cradle-head RGB-D vision system," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 181–191, Jan. 2019.

[17] Z. Chen, T. Gao, B. Sheng, P. Li, and C. L. P. Chen, "Outdoor shadow estimating using multiclass geometric decomposition based on BLS," *IEEE Trans. Cybern.*, to be published.

[18] B. Sheng, P. Li, C. Gao, and K.-L. Ma, "Deep neural representation guided face sketch synthesis," *IEEE Trans. Vis. Comput. Graphics*, to be published.

[19] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[20] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1297–1304.

[21] D. Mehta *et al.*, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, p. 44, 2017.

[22] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-D posture data," *IEEE Trans. Human–Mach. Syst.*, vol. 45, no. 5, pp. 586–597, Oct. 2015.

[23] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[24] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM Multimedia*, 2016, pp. 102–106.

[25] D. Arsenault and A. Whitehead, "Wearable sensor networks for motion capture," in *Proc. Int. Conf. Intell. Technol. Interact. Entertainment*, 2015, pp. 158–167.

[26] H. Zhang and Z.-Y. Zhang, "Human motion capture system based on distributed wearable sensing technology," in *Proc. Int. Conf. Wireless Commun. Sensor Netw.*, 2014, pp. 383–390.

[27] G. Tao, S. Sun, S. Huang, Z. Huang, and J. Wu, "Human modeling and real-time motion reconstruction for micro-sensor motion capture," in *Proc. IEEE Int. Conf. Virt. Environ. Human–Comput. Interfaces Meas. Syst.*, 2011, pp. 1–5.

[28] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 755–762.

[29] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3D pose estimation from a single depth image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 731–738.

[30] A. Shafaei and J. J. Little, "Real-time human motion capture with multiple depth cameras," in *Proc. Conf. Comput. Robot Vis.*, 2016, pp. 24–31.

[31] 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.

[32] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn, "Multisensor-fusion for 3D full-body human motion capture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 663–670.

[33] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1561–1570.

[34] H. Rhodin *et al.*, "Learning monocular 3D human pose estimation from multi-view images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1–10.

[35] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.

[36] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowl. Based Syst.*, vol. 158, pp. 43–53, Oct. 2018.

[37] X. Chen and M. Koskela, "Sequence alignment for RGB-D and motion capture skeletons," in *Proc. Int. Conf. Image Anal. Recognit.*, 2013, pp. 630–639.

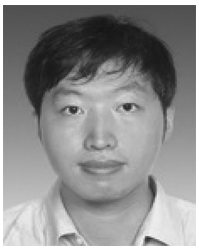
[38] D. Kulić, M. Choudry, G. Venture, K. Miura, and E. Yoshida, "Quantitative human and robot motion comparison for enabling assistive device evaluation," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, 2013, pp. 196–202.

[39] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica A Found. Adv.*, vol. 32, no. 5, pp. 922–923, 1976.

[40] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 168–172.

Aouaidjia Kamel received the M.Eng. degree in computer science from the Abbès Laghrour University of Khenchela, Khenchela, Algeria, in 2009. He is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

His current research interests include understanding human behavior, human–machine interaction, human pose estimation, machine learning, and deep neural networks.



Bin Sheng received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2011.

He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include machine learning, virtual reality, and computer graphics.

Ping Li received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2013.

He is currently with Hong Kong Polytechnic University, Hong Kong. He has one image/video processing national invention patent, and has excellent research project reported worldwide by *ACM TechNews*. His current research interests include image/video stylization, GPU acceleration, and creative media.

Jinman Kim received the B.S. (Hons.) and Ph.D. degrees in computer science from the University of Sydney, Sydney, NSW, Australia, in 2000 and 2005, respectively.

Since 2006, he has been a Research Associate with Royal Prince Alfred Hospital, Camperdown, NSW, Australia. From 2008 to 2012, he was an ARC Post-Doctoral Research Fellow, one year leave from 2009 to 2010 to join the MIRALab Research Group, Geneva, Switzerland, as a Marie Curie Senior Research Fellow. Since 2013, he has been with the School of Information Technologies, University of Sydney, where he was a Senior Lecturer, and became an Associate Professor in 2016. His current research interests include medical image analysis and visualization, computer-aided diagnosis, and telehealth technologies.

David Dagan Feng (F'03) received the M.Eng. degree in electrical engineering and computer science from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 1985 and 1988, respectively.

He is currently the Head of the School of Information Technologies, the Director of the Biomedical and Multimedia Information Technology Research Group, and the Research Director with the Institute of Biomedical Engineering and Technology, University of Sydney, Sydney, NSW, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his fields. Many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned.

Dr. Feng was a recipient of the Crump Prize for Excellence in Medical Engineering from UCLA. He has served as the Chair of the International Federation of Automatic Control Technical Committee on Biological and Medical Systems, organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions. He is a fellow of the Australian Academy of Technological Sciences and Engineering.