View-Invariant Human Action Recognition Based on a 3D Bio-Constrained Skeleton Model

Qiang Nie, Jiangliu Wang, Xin Wang, and Yunhui Liu, *Fellow, IEEE*

***Abstract*— Skeleton-based human action recognition has been a hot topic in recent years. Most existing studies are based on the skeleton data obtained from Kinect, which is noisy and unstable, in particular, in the case of occlusions. To cope with the noisy skeleton data and variation of viewpoints, this paper presents a view-invariant method for human action recognition by recovering the corrupted skeletons based on a 3D bio-constrained skeleton model and visualizing those body-level motion features obtained during the recovery process with images. The bio-constrained skeleton model is defined with two types of constraints: 1) constant bone lengths and 2) motion limits of joints. Based on the bio-constrained model, an effective method is proposed for skeleton recovery. Two types of new motion features, the Euclidean distance matrix between joints (JEDM), which contains the global structure information of the body, and the local dynamic variation of the joint Euler angles (JEAs) are used in describing human action. These two types of features are encoded into different motion images, which are fed into a two-stream convolutional neural network for learning different action patterns. The experiments on three benchmark datasets achieve better accuracy than the state-of-the-art approaches, which demonstrates the effectiveness of the proposed method.**

***Index Terms*— Human action recognition, view-invariant, CNN, skeleton recovery, bio-constrained skeleton model.**

I. Introduction

HUMAN action recognition has attracted lots of academic efforts for many years because of its applications in human-computer interaction, gaming, video surveillance, etc. However, it is still a challenging task for two reasons: 1) the complex spatial-temporal process of human behaviors; 2) variation of the environment and recording settings [1], including the background of image, occlusion, and viewpoint. Most early works concentrated on analyzing human actions based on RGB images. However, features extracted from color sequences are susceptible to illumination and the appearance of the human body, as well as lack of motion information in depth direction. With the development of 3D sensors, such as Kinect, 3D information of the human body can be captured at a low cost and human skeleton can be estimated in real time, which boosts the research on human action recognition significantly. Depth images simplify the segmentation of the human body from the background but suffer from problems of noisy data and varying presentations in the images when observing from different directions. Therefore, human action recognition based on skeleton data is gaining increasing attention in recent years. As Johansson [2] mentioned, the skeleton is one of the most effective ways to represent human actions.

Existing skeleton-based action recognition methods usually extract some view-invariant features, such as the displacement of joints within one frame or between frames [3], the histogram of joint orientations [4], and some higher level features like Lie group [5] and the covariant matrix of joints [6]. Despite the variety of features, most methods use skeleton data captured by the Kinetic sensor, which is usually noisy and unstable particularly in the case of occlusions. As shown in Fig. 2, varying length of bones and violating joint motion limits are two common problems. In Kinect, the joint positions are determined according to the pixel features in a single depth image [7] without applying rigid structure constraints compulsively. As a consequence, the estimated skeletons are often corrupted, which will further affect the human action recognition. In order to consider joint angle limits in the 3D pose reconstruction, Akhter *et al.* [8] collected a large motion capture dataset of stretching poses to study how joint limits change with different poses and generated an over-complete pose dictionary. The study based on a large dataset is high cost and low efficiency. In the proposed method, the joint angle limits are considered based on some simple medical data from neutral zero method [9].

Besides noisy skeleton and feature extraction, the representation of spatio-temporal information of human action is also an open problem. Early works recognized human actions by using hand-crafted features and temporal models [4], [5]. Since the success of deep learning methods in image processing, a lot of methods have been proposed based on the recurrent neural networks (RNNs) and the convolutional neural networks (CNNs) for human action recognition recently [10]–[23]. RNN has advantages in modeling the sequential objects, while CNN is better at extracting high-level spatial features. Combination of RNN and CNN networks for extracting high-level spatial features as well as learning temporal patterns is also being explored [22], [24]. According to the results reported in the state-of-the-art methods, the deep learning-based methods have achieved better performance than the traditional hand-crafted feature-based methods in

Manuscript received July 24, 2018; revised December 23, 2018 and February 10, 2019; accepted March 6, 2019. Date of publication March 22, 2019; date of current version June 20, 2019. This work was supported in part by Huawei Technologies Co., Ltd., through the Childcare Robot Project, under Grant 7010358, in part by the Hong Kong Research Grants Council (HK RGC) under Grant T42-409/18-R, and in part by the T Stone Robotics Institute of The Chinese University of Hong Kong (CUHK) under Project 4930745. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Liang Wang. *(Corresponding author: Qiang Nie.)*

The authors are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: qnie@mae.cuhk.edu.hk).

Digital Object Identifier 10.1109/TIP.2019.2907048

1057-7149 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

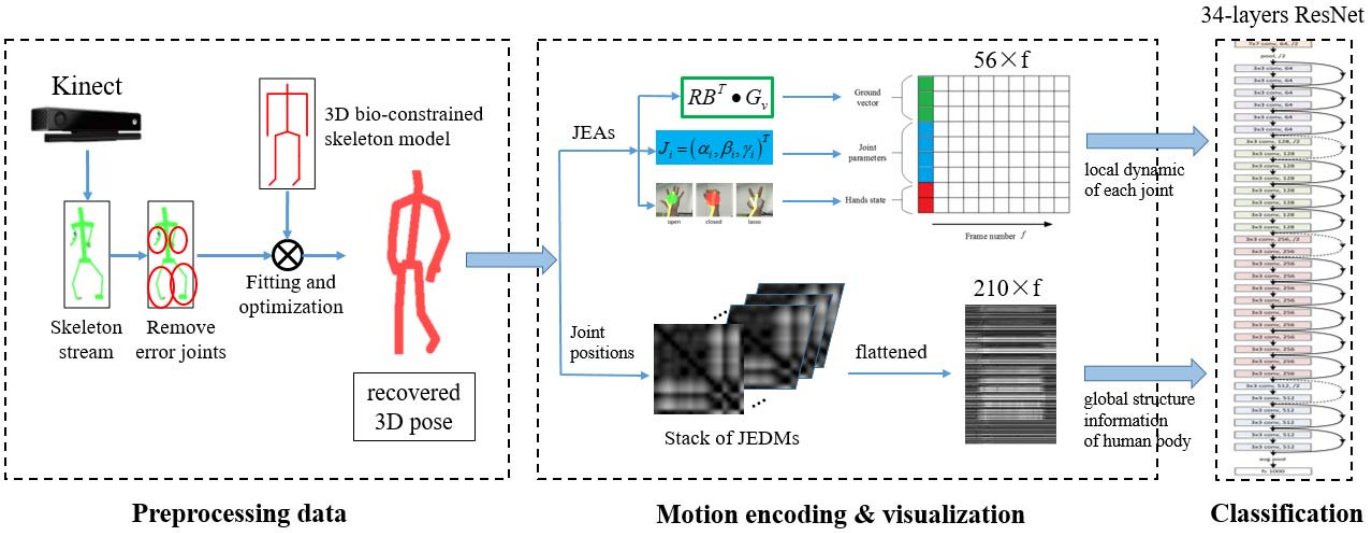


Fig. 1. Overview of the proposed method which includes data preprocessing, motion visualization and classification three steps. In the data preprocessing step, error joints or corrupt skeletons are detected and recovered based on a predefined 3D bio-constrained skeleton model. Based on the recovered skeleton, some body-level features are obtained, such as the JEAs and the JEDMs. These two view-invariant features are encoded into color and gray images respectively as representations of human actions in the step of motion visualization. Finally, the generated two types of motion images are fed into a two-stream CNN separately for action recognition and the results of the two branches are fused at decision level.

human action recognition. Nevertheless, the existing deep learning-based methods are not effective enough to represent both the spatial dependency and temporal distribution of human actions.

In this paper, we propose a 3D bio-constrained skeleton model for recovering corrupted skeletons and estimating joint Euler angles, and present a view-invariant CNN-based method for human action recognition using the recovered skeletons. The bio-constrained skeleton model is defined to satisfy two types of constraints: 1) constant bone lengths and 2) motion limits of joints. Based on these two types of constraints and motion continuity, an approach for recovering corrupt skeletons is proposed and the joint Euler angles (JEAs) are calculated during the recovery process. The body pose can be uniquely determined by a series of JEAs given the structure of the human body. Thus, to consider both the global structure and local joint variations, we use the Euclidean distance matrix between joints (JEDM) which is defined as a matrix of the pairwise Euclidean distances between joints to describe the structure of human body and relationships between joints besides JEAs. By visualizing the JEAs and the JEDM with images, human actions are represented by images which contain both the global (JEDM) and local (JEAs) pose information. In this way, the problem of action recognition is transformed into a problem of classifying these images. Hence, the advantages of CNN can be made use of in action recognition task. To summarize, our method consists of three stages: 1) Preprocess the skeleton data by using the skeleton recovery algorithm to rectify those noisy and unreasonable skeleton data; 2) Encode the local (JEAs) and the global (JEDM) pose features of the recovered skeleton into images; 3) Feed the encoded motion images, each of which contains both the spatial and the temporal information of an action, into a two-stream CNN for action recognition. The two CNN branches are fused in the decision level.

The contribution of our work can be summarized as the following aspects:

- A skeleton recovery method for tackling noisy skeleton data and estimation of the JEAs based on our 3D bio-constrained model is proposed. Different from the Kinect skeleton, the 3D bio-constrained skeleton model is a rigid structure.
- Two new view-invariant motion features, the JEAs and the JEDM, are proposed to represent the motion of each joint and the structure relationship between joints, respectively. Results of the extensive experiments verify the efficiency of the JEAs and JEDM in describing human action together. To our best knowledge, it's the first time that the JEAs and JEDM are used in the human action recognition.
- A framework from the skeleton data preprocessing to the action recognition is established, as shown in Fig. 1. Experiments on three benchmark datasets achieved a largest improvement of 4% compared to the state-of-the-art methods, which demonstrates the effectiveness of our proposed method for view-invariant action recognition.

The rest part of this paper is organized as: Section 2 gives a review of related works in view-invariant and skeleton-based action recognition. Section 3 introduces the defined 3D bio-constrained skeleton model and presents the approach for skeleton recovery and estimation of joint Euler angles. Section 4 proposes a new method for motion visualization and recognition by considering both the global and the local pose information. Section 5 provides detailed evaluations of our proposed method on three different datasets. Section 6 concludes this work.

II. Related Works

Human action recognition is still a challenging task due to the difficulty in representing the complex spatial-temporal

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

process of human actions, as well as coping with the environment settings including the background of images, occlusions, and viewpoints [1]. In this section, the review is focused on the view-invariant methods and the spatio-temporal representations of human actions.

A. View-Invariant Action Recognition Based on the Color or Depth Images

Most of the early works recognized human action based on the color image sequences. These methods not only have to deal with problems like illumination changes and segmentation of the human body from the background but also face a more thorny problem to make their algorithms robust to different views. Shen and Foroosh [25] defined a Fundamental Ratio which refers to the ratios among elements in upper left 2×2 submatrix of a fundamental matrix induced from plane motion as a view-invariant feature. This geometric transformation-based method assumes the correspondence of points in two RGB images can be accurately obtained, which is still an under exploring problem. Junejo *et al.* [26] proposed a temporal self-similarities matrix (SSM) that calculates the change of features between all pairs of frames as a descriptor in the cross-view action recognition. But this descriptor is not strictly view-invariant and mainly considers the temporal evolution. Rahmani and Mian [27] transferred unknown view data to a canonical view by using a learned Non-linear Knowledge Transfer Model (NKTM) network in cross-view action recognition. Compared to the color image, the depth image is affected less by illumination and more convenient for segmentation, but it also suffers from the varying action presentations in images caused by the observation angle. Oreifej and Liu [28] described the depth sequences using a histogram of the distribution of surface normal orientation in a 4D space of time, depth, and spatial coordinates. However, this method performs badly on the cross-view dataset. Rahmani *et al.* [29] introduced the histogram of oriented principle components (HOPC) as a descriptor for cross-view action recognition. But the overall computational time on a 3.4 GHz machine with 24GB RAM using Matlab is about 2 seconds per frame, which hinders the application of this method. Yang and Tian [30] proposed a general scheme of using super normal vector (SNV) for action recognition based on the depth images. Though it achieves a high performance on single view dataset, the SNV is not view-invariant. Unlike the color and depth data, view-invariant features can be extracted easier from skeleton data.

B. Extracting View-Invariant Features From Skeleton Data

Features based on the displacements of joints are widely used in skeleton-based action recognition for their simpleness in the calculation. Joint displacement can be divided into the spatial displacement and the temporal displacement [31]. The spatial displacements which code the structure information of human body are calculated between all pairs of joints [3], [17], [32], [33] or between some reference joints and the other joints [18], [19], [34]. The temporal displacements that depict the movement of joints are calculated between motion frames [3], [33]. Orientation is another commonly used feature to describe human pose. Xia *et al.* [4] used the histograms of 3D joint locations (HOJ3D) as a compact representation of postures for action recognition. They put the origin of a spherical coordinate system at human hip center and divided the spherical space into n bins. Each bin represents a joint orientation relative to the hip center. By casting all joints into the spherical space a HOJ3D is obtained. Some higher level features, such as Lie Group [5], [35] and the covariant matrix of joints [6], are also extracted from skeleton sequences. Reference [5] modeled the 3D geometric relationship between various body parts using the rotation matrix and translation matrix, and mapped the human actions to the curves in Lie Group. In [6], the covariant matrix that encodes the shape of the joint probability distribution of the set of random variables was used. Liu *et al.* [10] proposed a sequence-based view-invariant transformation on joint coordinates by establishing a principal coordinate system of torso frames and used the transformed joint coordinates for action recognition.

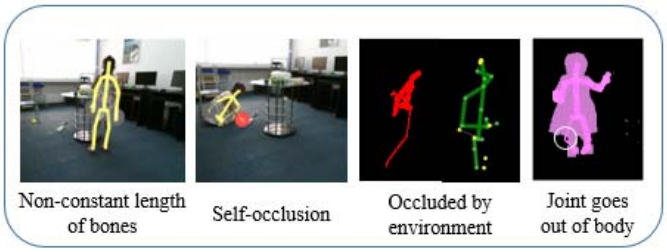


Fig. 2. Examples of corrupt skeletons obtained from kinect.

All the aforementioned methods build the body-level features directly from the raw skeleton data that is mainly obtained by Kinect sensor. However, as shown in Fig. 2, many problems exist in the Kinect skeleton, especially in the cases of cross-view action recognition where occlusions happen frequently. The corrupt skeletons will distort the extracted features and cause the mismatch between the features and the human actions. Instead of extracting motion features directly, the proposed method recovers those corrupt skeletons or error joints of the raw skeleton based on a 3D bio-constrained model first, which makes the proposed method more robust to noisy skeletons. Furthermore, joint Euler angles are calculated during this process and introduced as a view-invariant feature to describe human actions considering the articulated structure of the human body.

C. Modeling of Spatio-Temporal Information

Human action is a complex spatio-temporal process, which requires a balanced consideration between spatial changes and temporal evolution of human pose for recognition. Generally, existing representation methods of the spatio-temporal information can be divided into two categories: traditional hand-crafted feature-based models [3]–[6], [28], [32], [36], [37] and deep learning-based methods [10]–[22], [35]. The traditional models rely on hand-crafted features which are often dataset-dependent and ineffective in modeling the complex spatio-temporal process of human actions. The deep

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

learning based methods mainly employ two learning architectures: recurrent neural network (RNN) with the Long-Short Term Memory (LSTM) neurons and convolutional neural network (CNN). Du *et al.* [13] divided the whole human body into five parts and fed their motion features into five bidirectional recurrent neural networks (BRNNs) respectively. Lee *et al.* [14] proposed a novel ensemble Temporal Sliding LSTM (TS-LSTM) network which is composed of multiple parts including short-term, medium-term and long-term TS-LSTM. Liu *et al.* [15] developed a spatio-temporal LSTM with a new gating mechanism to handle the noises and occlusion cases. Song *et al.* [16] proposed a method to select discriminative joints based on LSTM. Wang and Wang [20] used a two-stream RNN architecture to model both the temporal dynamics and the spatial configurations. While the RNN architecture is suitable for modeling sequential data or temporal information, it lacks of the capability of extracting high-level patterns from the spatial information.

Different from RNN, CNN is better at learning high-level spatial features in the images, which has been verified by its success in image recognition [38]–[40]. To make use of the advantages of CNN, most CNN-based methods encode the motion features extracted from the skeleton data into images. Thereby, transfer the action recognition into an image classification problem. Liu *et al.* [10] proposed an enhanced motion visualization method that encodes the transformed coordinates into motion images in a 5D space (3 dimensions for coordinates, 1 dimension for frame number and 1 dimension for joint order). Ke *et al.* [18] transformed the cosine distance (CD) and the normalized magnitude (NM) which represent the spatial structure information of skeleton in each frame into images, and fed the encoded images of different body parts (trunk, right arm, left arm, right leg, left leg) into CNN separately. In another work of Ke *et al.* [19], displacements between four reference joints and the rest joints were encoded into gray images. Wang *et al.* [11] encoded the joint trajectories and their dynamics as the color distribution in images, which is referred to as Joint Trajectory Maps (JTM). Motion encoding has achieved promising performance in action recognition, but how to generate a more descriptive image is still underexplored. In our method, encoded JEAs and JEDM images are exploited together to consider the body structure and the variation of articulated joints.

III. Skeleton Recovery Method

As shown in Fig. 1, our method is comprised of three stages: skeleton preprocessing, motion visualization and classification. In the preprocessing step, we detect and rectify those unreasonable skeletons or joints. The recovered skeleton pose has fixed bone lengths and all joints are located within their motion limits. The consistency of human skeleton structure can be guaranteed across different datasets by recovering those human poses on the defined rigid 3D bio-constrained skeleton model.

Currently, most datasets for skeleton-based human action recognition are collected using Kinect. However, as shown in Fig. 2, there are a lot of errors in the skeletons estimated by the Kinect sensor, in particular in the case of occlusions.

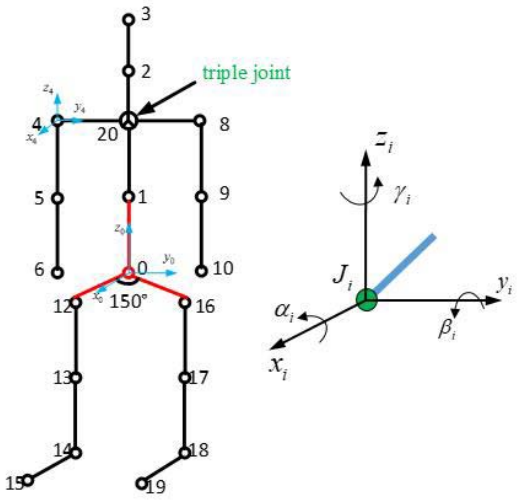


Fig. 3. 3D bio-constrained skeleton model and the definition of joint Euler angles in the local coordinate frame of J_i.

Since the algorithm inside the Kinect sensor estimates the skeleton based on pixel features in the depth image [7], the extracted skeleton often suffers from two problems: varying bone lengths and violation of the motion limits of the human body. If recognizing human actions using the raw skeleton data directly, the classification accuracy will be significantly affected by the noisy and corrupted skeletons.

A. 3D Bio-Constrained Skeleton Model

In order to recover corrupted skeletons and error positions of joints existing in the raw skeleton data, a 3D bio-constrained skeleton model is proposed as shown in Fig. 3. The "bio-constrained" means that the skeleton structure is constrained by inherent features of the human body, which refer to joint motion limits and fixed bone lengths. All the poses generated from this bio-constrained skeleton model must satisfy these two types of constraints. Besides the constraints, there are two main differences between our bio-constrained model and the Kinect skeleton model in the structure: 1) two more degrees of freedom are added to the joint SpineShoulder(20), which is denoted as a triple joint in Fig. 3; 2) the pelvis frame, formed by joints SpineBase(0), SpineMid(1), HipLeft(12) and HipRight(16), is fixed with a 150° angle between the bones of the left hip and the right hip in the horizontal plane. In the pelvis frame, the bone between joints SpineBase(0) and SpineMid(1) is perpendicular to the plane formed by bones of the left and right hip. The pelvis frame can only rotate and translate as a whole. The reason for adding two more degrees of freedom to SpineShoulder(20) is to separate motions of the left shoulder, the right shoulder, and the neck. It's a common sense that the movement of the left shoulder will not affect the position of right shoulder and neck, and vice versa. Design of a 150° angle in pelvis frame is for distinguishing the front side and the back side of the human body, which is an unsolved problem in the Kinect skeleton model. The angle value is an empirical selection after some trials. As the estimation of coordinates of hand joints is quite unstable and noisy, hand

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

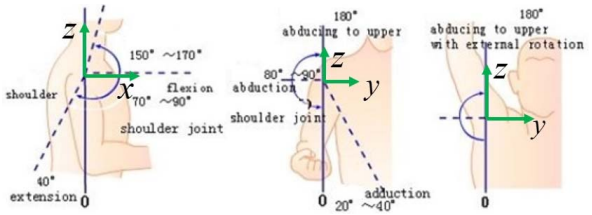


Fig. 4. The recommended motion range of right shoulder by the neutral zero method [9].

joints are not considered in our skeleton model. Actions of hands are represented by the changes of hand state instead, which is discussed in the section 5. The ID number of each joint is assigned in a similar way with the Kinect model for convenience.

Mathematically, a human skeleton can be defined as a set of joints and the bones between neighboring joints. In our model, as denoted in the equation 1, a joint is described with 5 parameters: 3D position in the camera space $PJ_i = (x_i, y_i, z_i)$, joint orientation $OJ_i = (\alpha_i, \beta_i, \gamma_i)$, initial orientation $IJ_i = (n_{xi}, n_{yi}, n_{zi})$, motion limits $LJ_i = (\alpha_{li}, \beta_{li}, \gamma_{li})$, and JointType which is an ID of the joint and refers to elbow, knee, shoulder, etc. In equation 1, LB_{ij} means the length of the bone between joints i and j. The joint orientation is defined with three Euler angles α, β, and γ. As shown in Fig. 3, in the local coordinate frame of joint i, α_i denotes the joint rotation angle about x_i axis, β_i denotes rotation angle about y_i axis and γ_i is the angle rotated about z_i axis. The sign of an angle is determined by the right hand rule. For the sake of simplifying the forward kinematics calculation, initial orientations of all joints are unified as the initial orientation of frame 0. Joint motion limits are determined according to the neutral zero method recommended by the American Association of Orthopedic Surgeons (AAOS) [9]. A joint is illegal if any of its three Euler angels is out of the corresponding motion range.

$$SK : \Omega_k = \{J_i(PJ_i, OJ_i, IJ_i, LJ_i, JointType) : OJ_i \in LJ_i, \|PJ_i - PJ_j\|_2 = LB_{ij}, i \in [0, 22]\} \quad (1)$$

Fig. 4 shows the motion limits of right shoulder recommended by the neutral zero method. In Fig. 4, the neutral zero position of the shoulder joint is shown as the solid blue line, where the arm points down to the ground naturally. The left picture illustrates the shoulder joint can rotate about y axis within a range of $[-170°, 40°]$ with respect to the neutral zero position. Similarly, from the right two pictures, we know that the rotation range of the right shoulder joint about x axis is $[-180°, 40°]$. Therefore, the motion limits of right shoulder are $\alpha \in [-180°, 40°]$ and $\beta \in [-170°, 40°]$.

B. Skeleton Recovery and Estimation of Joint Euler Angles

A novel method for skeleton recovery based on the 3D bio-constrained skeleton model is proposed. First, the constraints of the fixed bone lengths and joint motion limits are applied to the raw skeleton data to detect error joints. A parameter ω_i is used to record the detection results. ω_i equals to 1 when joint i is valid and 0 for an invalid joint. For those valid joints, we preserve their position information in the 3D bio-constrained skeleton. For those invalid joints, their positions will be relocated in the 3D bio-constrained skeleton model based on two principles: 1) positions of illegal joints are less important and serve for guaranteeing the positions of the correct joints in the kinematic chain; 2) considering the motion continuity, the true positions of the error joints in the current frame should be near their positions in the previous frame. By keeping the position information of the correct joints and relocating those error joints, the effective information of the observed pose can be recovered in the defined 3D bio-constrained skeleton. The whole skeleton recovery process can be formulated as equation 2:

$$arg \min_{\alpha,\beta,\gamma} D(SK_t^0 - SK_t^{ob}) = \sum_{i=0}^{22} [\lambda_r \omega_i \| PJ_{i,t}^0 - PJ_{i,t}^{ob} \|^2 + \lambda_w (1 - \omega_i) \| PJ_{i,t}^0 - PJ_{i,t-1}^0 \|^2] \quad (2)$$

where SK_t^0 means the state of the predefined 3D bio-constrained skeleton at instant t and SK_t^{ob} represents the observed skeleton state at instant t. α, β and γ are the JEAs, which determine the pose of a skeleton. $PJ_{i,t}^0$ is the ith joint position of the 3D bio-constrained skeleton in the frame t, $PJ_{i,t}^{ob}$ is the position of the observed joint i in the frame t and $PJ_{i,t-1}^0$ is the position of the bio-constrained joint in the frame $t-1$. $\omega_i = 1$ if joint i is valid, otherwise $\omega_i = 0$. λ_r is the weight for valid joints and λ_w is the weight for error joints. Usually, the λ_r is much larger than λ_w, which means the position of the valid joint is much more important than the estimated position of the invalid joint. By minimizing the value of equation 2, the pose of the observed skeleton can be recovered in the bio-constrained skeleton model and those invalid joint positions are rectified.

At instance k, the position of a joint $i+1$ can be calculated recursively based on its father joint i.

$$\begin{aligned} PJ_{i+1}(k) &= R_B(k) R_i(k) \overrightarrow{PJ_i PJ_{i+1}(0)} + PJ_i(k) \\ PJ_0(k) &= PJ_0^{ob}(k) \\ R_i(k) &= \prod_i {}_1^0R \; {}_2^1R \; {}_3^2R \cdots {}_i^{i-1}R \end{aligned} \quad (3)$$

where PJ_{i+1} and PJ_i are the positions of bio-constrained joints $i+1$ and i in the camera space, $\overrightarrow{PJ_i PJ_{i+1}(0)}$ is the bone vector from joint i to joint $i+1$ at instant 0, namely the bone vector in the initial skeleton pose as shown in Fig. 3. Joint 0 (SpineBase), the root joint, is aligned with the observed joint 0. $R_B(k)$ is the global rotation matrix at instant k of the whole skeleton, which rotates the whole skeleton to match the observed skeleton. $R_i(k)$ is the rotation matrix from the local coordinate frame i to the coordinate frame 0 in the bio-constrained skeleton. ${}_i^{i-1}R$ denotes the rotation of joint i, which is defined as:

$${}_i^{i-1}R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_i & -\sin\alpha_i \\ 0 & \sin\alpha_i & \cos\alpha_i \end{bmatrix} \begin{bmatrix} \cos\beta_i & 0 & \sin\beta_i \\ 0 & 1 & 0 \\ -\sin\beta_i & 0 & \cos\beta_i \end{bmatrix} \begin{bmatrix} \cos\gamma_i & -\sin\gamma_i & 0 \\ \sin\gamma_i & \cos\gamma_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

To make the calculation more efficient for online application, the joint position in the $(k+1)_{th}$ frame is calculated based on the joint position in the k_{th} frame.

$$\begin{aligned} PJ_{i+1}(k+1) &= R_B(k+1)\Delta R_i(k+1) \\ &\quad \cdot \overrightarrow{PJ_iPJ_{i+1}(k)} + PJ_i(k+1) \\ \Delta R_i(k) &= \prod_i {}^{i-1}_{i}\Delta R \cdots {}^{2}_{3}\Delta R\, {}^{1}_{2}\Delta R\, {}^{0}_{1}\Delta R \end{aligned} \tag{5}$$

where $R_B(k+1)$ denotes the global rotation matrix of the whole human body in frame $k+1$ and $\Delta R_i(k+1)$ denotes the local rotation matrix from the k_{th} frame to the $(k+1)_{th}$ frame. ${}^{i-1}_{i}\Delta R$ can be calculated by:

$$\begin{aligned} {}^{i-1}_{i}\Delta R = &\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\Delta\alpha_i & -\sin\Delta\alpha_i \\ 0 & \sin\Delta\alpha_i & \cos\Delta\alpha_i \end{bmatrix} \cdot \begin{bmatrix} \cos\Delta\beta_i & 0 & \sin\Delta\beta_i \\ 0 & 1 & 0 \\ -\sin\Delta\beta_i & 0 & \cos\Delta\beta_i \end{bmatrix} \\ &\begin{bmatrix} \cos\Delta\gamma_i & -\sin\Delta\gamma_i & 0 \\ \sin\Delta\gamma_i & \cos\Delta\gamma_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{6}$$

where $\Delta\alpha_i$, $\Delta\beta_i$ and $\Delta\gamma_i$ are changes of Euler angles at joint i between two sequential frames. By using the Rodrigues rotation matrix, ${}^{i-1}_{i}\Delta R$ can be formulated as

$$\begin{aligned} {}^{i-1}_{i}\Delta R &= (I + \Delta\alpha_i K_{xi})(I + \Delta\beta_i K_{yi})(I + \Delta\gamma_i K_{zi}) \\ &\approx I + \Delta\alpha_i K_{xi} + \Delta\beta_i K_{yi} + \Delta\gamma_i K_{zi} \end{aligned} \tag{7}$$

where I is a 3×3 identity matrix, K_{xi}, K_{yi} and K_{zi} are the elements of Lie algebra $SO(3)$ generating the rotation group $SO(3)$ of $\mathbb{R}^3$. Assuming the current direction vector of axis x is $e_x = (k_{x1}, k_{x2}, k_{x3})$, K_x is calculated as the equation 8. K_y and K_z can be obtained similar with K_x.

$$K_x = \begin{bmatrix} 0 & -k_{x3} & k_{x2} \\ k_{x3} & 0 & -k_{x1} \\ -k_{x2} & k_{x1} & 0 \end{bmatrix} \tag{8}$$

The skeleton recovery process is implemented with two steps based on equations 2∼8. In the first step, those joints located in the torso are selected to calculate the global rotation matrix R_B and the translation $\overrightarrow{t}$ of the human body. In the second step, the remain joints are divided into 5 different kinematic chains according to their relationships with each other, as shown in Fig. 5. The recovery algorithm is implemented in parallel on the 5 kinematic chains and can run at a speed of around 15 frames per second on a computer with an Intel i7 CPU and 8GB RAM, which shows a better potential for online application compared to [29].

IV. Motion Visualization and Recognition

A rigid recovered skeleton and some body-level motion features, such as the JEAs which contain the local pose information of each joint, are obtained in the former section. These body-level features not only describe the pose of the human body but also are view-invariant.

In the following, a motion visualization technology which encodes the extracted motion features in each frame into color or gray images will be introduced. Each column in the encoded image corresponds to an action frame. Thus, an action sequence with f frames is converted to an image with f columns after encoding. The generated motion images contain both the global and local pose information in each frame, as well as their variations along with time. Finally, these encoded motion images are fed into CNNs to extract high-level spatio-temporal features for action recognition.

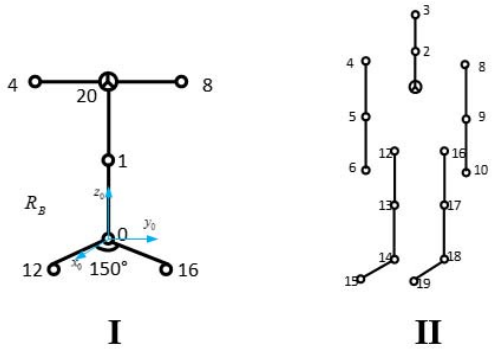


Fig. 5. The skeleton recovery algorithm is implemented with two steps: I) the joints located in torso are recovered first and the global rotation matrix is estimated; II) the remaining joints are divided into 5 kinematic chains according to their physical relationships and different chains are recovered parallelly.

A. Motion Visualization

Motion visualization is a technique to encode motion features as some visible graphics. There are many advantages by visualizing the human motion features with images. Firstly, such a kind of image contains both of the spatial information and the temporal information. By means of visualization, the motion pattern is converted into the image pattern presented on the encoded image. Therewith, the problem of action recognition is transformed into an image recognition problem. The capability of CNN in image recognition can be made full use of. Secondly, it is helpful to tackle the problem of different action durations and different action starting time in motion clips. The action duration will cause the pattern presented on the encoded image to become wider or narrower. The action starting time will cause the translation of motion pattern in the encoded image. CNN is robust to these kinds of image variations, which helps improve the robustness of action recognition algorithm to different subjects.

Considering the reasons mentioned above, we adopt motion visualization technology and propose a new approach to encode the human actions. Human action, in essence, can be defined as a continuous variation of body state during a certain time interval. A good body state descriptor should contain both the global pose of the whole body and local pose of each body part. Under this principle, a concatenate state vector $\Gamma(P_G, P_E, P_H)$ is introduced to represent human body state, where P_G is the global orientation of human body that can be derived from the global rotation matrix R_B, P_E is the local pose of human body which consists of all the JEAs, P_H is the pose state of hands. The rotation matrix R_B describes the general orientation of the human body with respect to the environment. To unify the representation and make it view-invariant, the transpose of R_B is multiplied with the gravity vector G_v in current camera space and get a pose

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

Fig. 6. 3 states of hand defined in the kinect model.

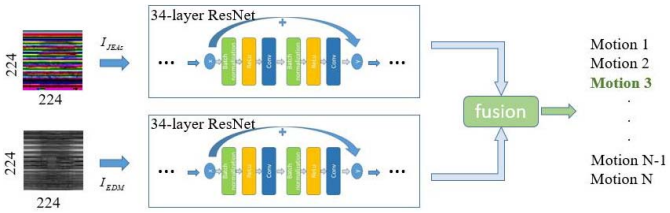


Fig. 7. The two-stream neural network based on ResNet.

vector $P_G = R_B^T \cdot G_v = (Pg_x, Pg_y, Pg_z)^T$. Hence, P_G can be seen as the gravity vector expressed in human body coordinate frame and depicts the pose of human body relative to the ground plane. For example, P_G can tell us whether a person is standing straight up, leaning or lying down, which can not be inferred from the JEAs. In our 3D bio-constrained skeleton model, joints of hands are not considered because of their noisy coordinates estimated by the Kinect. To describe the movements on hand, hand state which is more robust to noisy data and independent to the camera position is used instead of the Euler angles of hand joints. Three different states of a tracked hand are defined in Kinect: lasso, open and closed, as shown in Fig. 6. These states are capable to represent some basic actions done by hands, such as grasp, push, cup hands, point to, etc.

However, the motion images generated above emphasize more on the temporal variation of each joint while lack of the structure relationships between joints. As a complementary to JEAs feature, we use the Euclidean Distance Matrix between joints (JEDM) to describe the human body structure and the spatial relationships between joints. JEDM is defined as a matrix of the pairwise Euclidean distances between joints as denoted in equation 9, where λ is used to normalize the distance value within the range [0, 1], and $i, j \in [0, 20]$ according to the number of joints in our 3D bio-constrained skeleton model. EDM has been widely used in the modal analysis, structure representation, and recovering 3D human pose from single image [41]. It has been verified that the EDM not only encodes the underlying structure of vector representations but also can capture richer information about pairwise correlations between body joints. Besides, it is coordinate-free, invariant to rotation, translation, and reflection. As JEDM is symmetric, only the left lower half of each JEDM is encoded to a column of a gray image for action recognition as shown in Fig. 7.

$$JEDM_{i,j} = \frac{1}{\lambda} \left\| p_i - p_j \right\|_2 \tag{9}$$

Based on the definition of body state vector Γ, a human action with f frames can be defined as a series of human body states $\mathcal{A} = (\Gamma_1, \Gamma_2, \ldots, \Gamma_f)$, and motion visualization is to map these human body states to an image $\mathcal{M}$, as formulated by equation 10:

$$\mathcal{M} = \mathcal{F}(\mathcal{A}) = \mathcal{F}(\Gamma_1, \Gamma_2, \ldots, \Gamma_f) \tag{10}$$

where, $\mathcal{F}(\cdot)$ is the mapping function that maps an action from the motion feature space to the image space. Each column of the image $\mathcal{M}$ corresponds to a state Γ of the human body in a frame. Column number is ordered as the frame number. Thus, the spatial pose states and their temporal evolutions are encoded into a 2D image. The mapping function of JEAs is formulated as equation 11. The global pose P_G is normalized by its L2 length. Different joint motion limits are considered in calculating the mapped color values of JEAs to equalize the role of different joints in describing an action. For hand pose P_H, different colors are assigned to different states directly: green for the open state, red for the closed state, blue for lasso state, purple for untracked, and black for unknown. In JEDM encoding, the λ is selected as the height of person to eliminate the influence of human size and the corresponding pixel value is $p_{i,j} = 255 JEDM_{i,j}$.

In total, rotations of 17 joints must be considered in our bio-constrained skeleton model for JEAs encoding and 21 joints have to be considered for JEDM encoding. Therefore, the body state Γ_{JEAs} has a dimension of 56, which comprises the gravity pose vector $P_G(1 \times 3)$, the JEAs $P_E = (\alpha, \beta, \gamma)$ (17×3) and the hand states $P_H(2 \times 1)$. The three elements of P_G vector or each $JEAs$ are mapped to B, G, R color respectively and are put in three consecutive rows of an image column. The body state Γ_{JEDM} has a dimension of 210 since JEDM is a 21×21 matrix and only half of the elements are considered. As a result, an action $\mathcal{A}$ with f frames is encoded into two motion images, $\mathcal{M}_{JEAs}$ with a size of $56 \times f$ and $\mathcal{M}_{JEDM}$ with a size of $210 \times f$.

$$\begin{aligned} &\mathcal{F}(\Gamma_{56\times 1}) \\ &= (\mathcal{C}_{P_G}, \mathcal{C}_{P_E}, \mathcal{C}_{P_H})^T \\ &= \left(\frac{255}{\|P_G\|_2}\begin{bmatrix} Pg_x \\ Pg_y \\ Pg_z \end{bmatrix}, 255 \begin{bmatrix} \frac{\alpha - \alpha_{lmin}}{\alpha_{lmax} - \alpha_{lmin}} \\ \frac{\beta - \beta_{lmin}}{\beta_{lmax} - \beta_{lmin}} \\ \frac{\gamma - \gamma_{lmin}}{\gamma_{lmax} - \gamma_{lmin}} \end{bmatrix}_{\times 17}, \begin{bmatrix} \mathcal{C}_{HL} \\ \mathcal{C}_{HR} \end{bmatrix}\right)^T \end{aligned} \tag{11}$$

B. Motion Recognition

ResNet [38] which has been proved to be effective in the ImageNet Large-ScaleVisual Recognition Challenge (ILSVRC) is selected as the backbone of our convolutional neural network for extracting high-level features from JEAs images and JEDM images. The basic module of ResNet is shown in Fig. 7. Given an input feature x, the output of this module at layer t can be formulated as

$$y = W_2^T \sigma(BN(W_1^T \sigma(BN(x)))) + x = h_t(x) + x \tag{12}$$

where $BN(\cdot)$ refers to batch normalization, σ is a nonlinear activation function, W_1 and W_2 are the weights of the convolutional layer. Due to the recursive relationship between layers, the output of $t+1$ layer is $y_{t+1} = h_{t+1}(y_t) + y_t$. Thus, for a T

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

layers ResNet, the output at T-th layer is $y_T = \sum_{t=1}^{T} h_t(x_t)$, where x_t is the features input to the tth layer and $x_1 = \mathcal{M}$. The final output of T-layer ResNet with the input of motion image $\mathcal{M}$ is

$$\hat{y} = \phi(\mathbf{W}_c^T \mathcal{H}(\mathcal{M})) = \phi\left(\mathbf{W}_c^T \sum_{t=1}^{T} h_t(x_t)\right) \tag{13}$$

where $\mathbf{W}_c$ is the weights of the final fully connected layer for classification, $\hat{y}$ is the predicted label corresponding to ground truth y which is often encoded as a one-hot vector. $\phi(\cdot)$ denotes the mapping relationship from the high-level representation of motion image to the label space $\phi(R) : R \rightarrow \mathcal{Y}$. The training loss function is defined as a cross-entropy loss together with a regularization penalty $R(\mathbf{W})$.

$$L = -\frac{1}{N}(\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i)) + R(\mathbf{W}) \tag{14}$$

where N is the size of the training set. $\mathbf{W}$ are the trainable weights of the whole network and $R(\mathbf{W}) = \sum_i \mathbf{W}_i^2$.

The two CNN streams are trained separately and they are fused together at the decision level by averaging their prediction scores, which is formulated as:

$$s(l|\mathcal{I}_{JEAs}, \mathcal{I}_{JEDM}) = \frac{1}{2}(prob(l|\mathcal{I}_{JEAs}) + prob(l|\mathcal{I}_{JEDM})) \tag{15}$$

Actually, we also explored some other fusion strategies, such as maximum and linear weighted average, and tried training the two branches jointly by adding a fully connected layer at the top of the two branches. But the average strategy denoted in equation 15 achieves the best results. How to fuse the two branches of such a kind of two-stream network together without training it separately and achieve a better result than the average fusion strategy is one of our future work.

V. Experiments

We evaluate our method on three benchmark datasets: North-westernUCLA [42], MSRC-12 [43] and NTU RGB+D dataset [23]. The MSRC-12 is a single-view dataset, while the Northwestern-UCLA and the NTU RGB+D dataset are collected from different views. We apply the cross-subject training protocol on the MSRC-12 dataset and the cross-view training protocol on the Northwestern-UCLA dataset. For the NTU RGB+D dataset, which is the largest 3D skeleton dataset for action recognition so far, verification of our method is based on both the cross-subject and the cross-view protocol.

A. Implementation Details

In our experiments, we select a pre-trained 34-layer ResNet as the backbone network and implement the two-stream CNN with TensorFlow. There are several numbers of layers that are commonly used in ResNet: 18, 34, 50, 101, and 152. Here the choice of 34 is just a compromise between the performance and model size. Batch normalization and ReLU layers are utilized in this network. Batch size is set to 64 and training images are shuffled randomly. Weights are updated using the stochastic gradient descent (SGD) method. Momentum optimizer with a momentum value of 0.9 and a weight decay of 0.0001 are used. Initial learning rate is set to 0.015. Prediction accuracy on the test data is evaluated after every training epoch. The training is performed on a computer with 4 NVIDIA Titan XP graphics cards, 64 GB RAM and an Intel Xeon(R) processor E5-2640.

Considering the varying length of action sequences done by different subjects, the encoded motion image of each action sample is resized to 224×224 to make use of the pre-trained network. To some extent, resizing the image will cause loss of the original physical meaning of pixel values, especially resizing along the height direction. However, after encoding those features into an image, what we truly care about is the pattern presented in the image. Different patterns represent different actions. In image space, resizing operation will not change the pattern existing in the encoded image. Therewithal, resizing operation will not influence the recognition result. For those actions performed by two subjects in the NTU RGB+D dataset, their motion images are synthetic images by averaging the corresponding pixel values in the motion images of the two engaged subjects.

B. MSRC-12 Dataset

This single-view dataset comprises 594 sequences and totally 71 359 frames (approx. 6h40m) are collected by the Kinect sensor. About 6244 instances of 12 gestures are generated by 30 subjects. The skeleton data records the position of 20 joints with ∼2cm accuracy. Compared to the proposed 3D bio-constrained skeleton model, there is no neck joint in the skeleton of this dataset. So a neck joint is added to the skeleton, which is located between the head joint and the shoulder center with a distance of 1/4 length of the neck bone to the shoulder center joint. The skeleton recovery and the calculation of the JEAs and JEDM are performed as mentioned in section 3 and 4. Since no hand state is recorded in this dataset, we set all the hand states to unknown. The 12 gestures in this dataset include “start system”, “duck”, “push right”, “goggles”, “wind it up”, “shoot”, “bow”, “throw”, “had enough”, “change weapon”, “beat both” and “kick”. Even without motion information of hands, these 12 gestures have enough differences in the movements of other body parts. Therefore, the influence of ignoring hand states is negligible, which is supported by the action recognition result shown in Table I. Fig. 8 is an encoded motion image of the “duck” motion. In Fig. 8, we can clearly see that the “duck” motion is performed 11 times during the recording time. The presented patterns in the image of these 11 times of motions are quite clear and similar to each other, which indicates the robustness and effectiveness of the proposed motion visualization method.

Before feeding the encoded images into CNN, these long motion images with multiple times of motions are segmented into shorter clips that only contain single times of motion. Similar to [10], [11], we apply the cross-subject protocol on this dataset, which means motion instances performed by odd subjects are used as training set and instances acted by

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

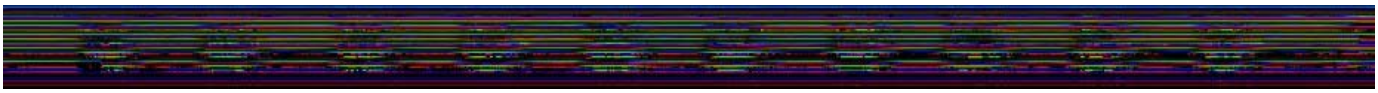


Fig. 8. Encoded image of the "duck" motion from MSRC-12 dataset.

TABLE I

COMPARISON OF RESULTS ON MSRC-12 DATASET(CROSS-SUBJECT PROTOCOL [10], [11])

Feature	Method	Accuracy(%)
Hand-crafted	ELC-KSVD 2014 [36]	90.22
	Cov3DJ 2013 [6]	91.70
CNN	JTM 2016 [11]	93.12
	ESV(average) 2017 [10]	94.60
	Ours (only JEDM)	**92.93**
	Ours (only JEAs)	**88.91**
	Ours Two-stream(JEDM + JEAs)	**94.20**

even subjects are used for testing. A final evaluation accuracy of 94.20% is achieved on this dataset.

Comparisons between the proposed method and other methods are shown in Table I. Our method achieves the-state-of-art evaluation accuracy, which outperforms the best hand-craft feature-based methods by 2.5%. Compared to the CNN-based method JTM [11], our method performs 1% better. Though our method performs as good as the ESV [10] that also visualizes actions with images, our method is more efficient considering the fact that the result of ESV fused 10 CNN streams and a number of synthesized samples are used for training. The ESV method reported a better result based on the weighted fusion strategy which only considers some selected branches, but they didn't explain explicitly how to select those streams. Hence, to evaluate the efficiency of the main method other than fusion strategies, we only compared with their results based on the same average fusion strategy in all experiments. Actually, our method can achieve a quite good result with only the JEDM stream. Meanwhile, the JEDM is much simpler to calculate compared with features used in the JTM and ESV method. Different from the proposed method, JTM method visualizes joint trajectories and ESV encodes the transformed joint coordinates. Performance of our method on this dataset verifies the efficiency of the proposed method in recognizing motions of different subjects. The recognition confusion matrix is shown in Fig. 9.

C. Northwestern-UCLA Dataset

In order to evaluate the performance of the proposed method on different views, Northwestern-UCLA dataset [42] is chosen as another test dataset. This dataset includes 10 action categories: "pick up with one hand", "pick up with two hands", "drop trash", "walk around", "sit down", "stand up", "donning", "doffing", "throw", and "carry". Each action is performed by 10 actors and totally 1494 motion sequences are collected from 3 different viewpoints. Following [42] and [10], we use the cross-view training protocol, that is, using data from the first two views as the training set and samples from the third view as the test data.

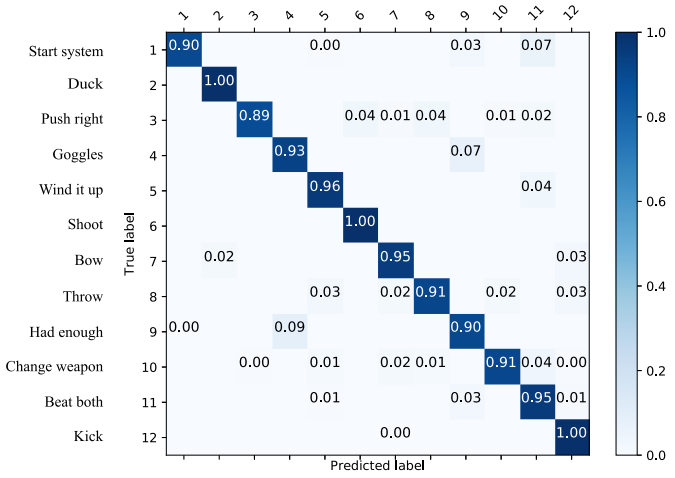


Fig. 9. Confusion matrix of MSRC12 dataset.

TABLE II

RESULTS ON NORTHWESTERN-UCLA DATASET (CROSS-VIEW PROTOCOL [10], [42])

Feature	Method	Accuracy(%)
Hand-crafted	HOJ3D 2012 [4]	54.50
	LARP 2014 [5]	74.20
	TLDS 2018 [44]	74.6
RNN	HBRNN-L 2015 [13]	78.52
	TS-LSTM 2017 [14]	89.22
	Denoised-LSTM 2018 [45]	80.25
	Multi-task RNN 2018 [46]	87.3
CNN	ESV(average) 2017 [10]	90.44
	Ours (only JEDM)	**91.47**
	Ours (only JEAs)	**86.40**
	Ours Two-stream(JEDM + JEAs)	**94.40**

Table II shows recognition results on the Northwestern-UCLA dataset. According to the results, the CNN-based methods outperform the traditional methods based on hand-crafted features significantly. Our method reaches an accuracy of 94.40%, which is the highest among all reviewed methods. Compared with the best hand-crafted feature-based method TLDS [44], our method achieves a better performance by 20%. Compared with the RNN-based methods, the proposed method outperforms the latest Multi-task RNN [46] method by 7% and TS-LSTM [14] method by 5%.

Among those skeleton-based methods, HOJ3D [4] extracts the histogram of joint orientation and LARP [5] depicts the action with Lie group. ESV [10] is the one that our method is most similar to. ESV uses the transformed joint 3D coordinates for motion visualization. While our method estimates the Euler angles of each joint and the EDM of recovered skeleton during the skeleton recovery process considering the articulated structure of human body, which are turned out to be more discriminative view-invariant features. Compared with these three methods, our method improves the recognition accuracy

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

by 40%, 20%, and 4% respectively. In detail, we obtain an accuracy of 86.40% with only the branch of JEAs and an accuracy of 91.47% with only the JEDM stream. The performance of the proposed method achieved on this dataset proves the effectiveness of the proposed features and method in different views.

However, both the MSRC-12 and the Northwestern-UCLA dataset are not big enough. Therefore, the proposed method is tested on the NTU-RGB+D dataset which is one of the largest 3D skeleton datasets in the following subsection.

D. NTU RGB+D Dataset

The NTU RGB+D dataset is one of the largest 3D skeleton datasets for action recognition. It contains 60 motion categories including "drink water", "eat meal/snack", "brushing teeth", "brushing hair", "drop", "pickup", and so on. These actions are performed by 40 subjects and more than 56K motion samples from various views are generated. Different subjects and viewpoints bring big challenges in discriminating intra- and inter-class variations. Considering the dataset size, the similarity between actions, and large noises in the dataset, NTU RGB+D dataset is a quite challenging dataset for action recognition. According to [23], we test our method utilizing both the cross-subject and the cross-view protocol. In the cross-subject protocol, the motion of subjects with ID 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38 are selected as training data which have 40 320 samples and actions of the remaining 20 subjects which comprises 16 560 samples are used for testing. In cross-view evaluation, 18 960 sequences from camera 1 are used for testing and 37 920 sequences from camera 2 and 3 are selected for training.

Performance of our method and comparison with state-of-the-art methods are shown in Table. III. The cross-subject accuracy and the cross-view accuracy of our method are 86.68% and 91.79% respectively, which are the best among the reviewed methods. In the cross-subject evaluation, our method achieves a 25% improvement compared to the best hand-crafted feature-based method LieGroups [35] and a nearly 10% improvement compared to the best RNN-based method two-stream GCA-LSTM [16]. Among the CNN-based methods, our method increases the cross-subject evaluation accuracy by 3.4% compared with CNN+Motion+Trans method [21] and the cross-view evaluation accuracy by 1.7% compared with LSTM+CNN method [22]. SkeletonNet [18] and ESV [10] encode the joint coordinates in motion visualization, which are different from our method that based on the JEAs and JEDM. Reference [19] proposed a novel skeleton transformer module to select the important skeleton joints automatically in data processing, while our method recovers those corrupt skeletons based on the proposed 3D bio-constrained skeleton model. Performance of our method which is much better than [10], [18] and [19] should attribute to the effectiveness of the view-invariant features (JEAs and JEDM) and the skeleton recovery process.

The confusion matrix of the cross-view predictions is shown in Fig. 10. For those actions with high recognition error, either they have higher inter-class similarity with other actions or mainly consist of movements conducted by fingers which are recorded noisily. For example, about 10% of "wipe face" action is wrongly recognized as "brushing teeth". Both of these two actions are slight motions of the hand moving around the face. The action of "walk towards each other" has a 15% confusion rate with "walk apart from each other" because they have the same leg motions. Some other errors like confusing "take off a shoe" with "wear a shoe" also have a high incidence due to noisy hand skeleton and similarity between these two actions. To reduce these errors requires more consideration of interactions between human or between human and surroundings. Better observation and description of hand motion are also needed for recognizing those actions with tiny movements on hands. Although limited by the inaccurate data, the proposed method still reveals its effectiveness in human action recognition with a best performance on this dataset.

TABLE III
RESULTS ON NTU RGB+D DATASET [23]

Feature	Method	CS(%)	CV(%)
Hand-crafted (depth)	HON4D 2013 [28]	30.56	7.26
	SNV 2017 [30]	31.82	13.61
Hand-crafted (skeleton)	LARP 2014 [5]	50.08	52.76
	LieGroups 2017 [35]	61.37	66.95
RNN	HBRNN-L 2015 [13]	59.07	63.97
	Part-aware LSTM 2016 [23]	62.93	70.27
	ST-LSTM+TG 2016 [15]	69.20	77.70
	STA-LSTM 2017 [16]	73.40	81.20
	Two-stream RNN 2017 [20]	71.30	79.50
	Two-stream GCA-LSTM 2018 [47]	77.10	85.10
CNN	SkeletonNet 2017 [18]	75.94	81.16
	S-Mul-Score fusion 2017 [17]	76.20	82.30
	JTM 2016 [11]	76.32	81.08
	Clips + CNN +MTLN 2017 [19]	79.57	84.83
	ESV(average) 2017 [10]	79.81	86.70
	CNN+Motion+Trans 2017 [21]	83.20	89.30
	LSTM+CNN 2017 [22]	82.89	90.10
	DPRL+GCNN 2018 [24]	83.5	89.8
	Ours (only JEDM)	**78.55**	**84.54**
	Ours (only JEAs)	**75.89**	**81.75**
	Ours Two-stream(JEDM + JEAs)	**86.68**	**91.79**

E. Evaluation of the Skeleton Recovery Process

To further evaluate the contribution of the skeleton recovery module, we compare the action recognition results based on the original skeleton data and the recovered skeleton data using JEDM feature. As the feature of JEAs can only be obtained after recovering, such a comparison is not able to be implemented based on the JEAs. The comparison results are shown in Table. IV. The results listed in the Table. IV demonstrate that skeleton recovery process does help improve the action recognition accuracy. Under the cross-subject training protocol, the performance based on the recovered skeleton data achieves an improvement of 1.77% and 0.52% respectively on MSRC-12 and NTU RGB+D dataset compared with the accuracy based on the original skeleton data. Under the cross-view protocol, using the recovered skeleton can increase the accuracy by 1.07% on Northwestern-UCLA dataset and 0.62% on NTU RGB+D dataset.

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

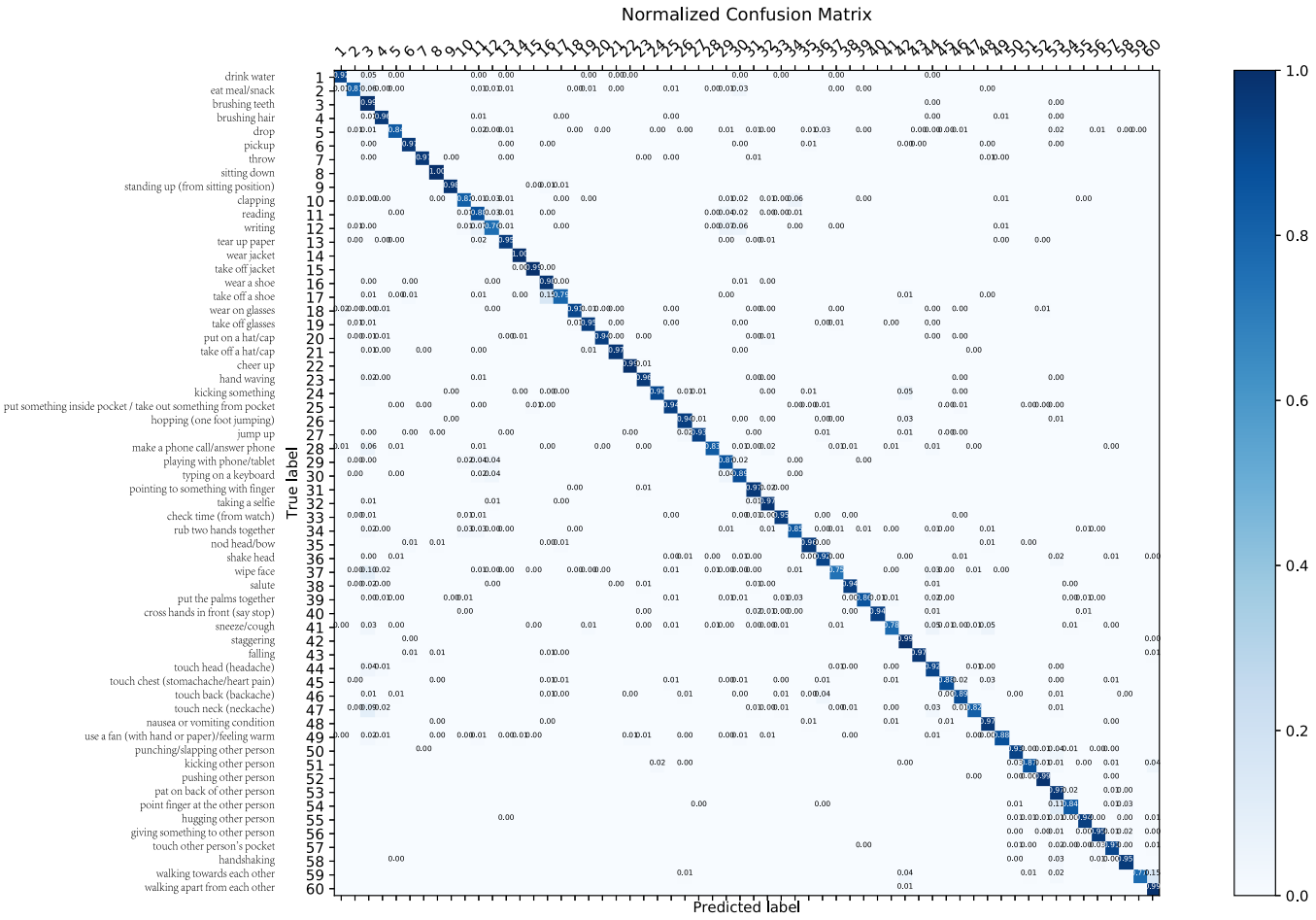


Fig. 10. Confusion matrix of the NTU RGB+D dataset under cross-view protocol (overall accuracy=91.79%).

TABLE IV
EVALUATION OF THE SKELETON RECOVERY MODULE BASED ON JEDM FEATURE

Dataset	Training Protocal	Accuracy(%)	
		original	recovered
MSRC-12	CS	91.16	92.93
Northwestern-UCLA	CV	90.40	91.47
NTU RGB+D	CS	78.03	78.55
	CV	83.92	84.54

In order to visualize the contribution of the skeleton recovery process in our method for action recognition, contrasts between the original skeleton and the recovered skeleton are made in Fig. 11. In the motion of "pick up with one hand", the original skeletons (the first row) are much noisier than the recovered skeletons (the second row), especially when people squatting down and self-occlusion happening. The recovered skeleton which is constrained with fixed bone lengths and joint motion limits has a more stable and clearer skeleton structure. Clear skeletons are crucial in skeleton-based action recognition for distinguishing similar actions. Similar results can also be found in the comparison of motions of "pickup with two hands", "sit down", "doffing" and "carry", where occlusions happen frequently. These skeleton contrasts indicate that pose recovery based on bio-constrained skeleton model is effective in data denoising and improving the recognition performance.

From another perspective, the skeleton recovery process is also a feature extracting process where we calculate the joint Euler angles. The JEAs is an important view-invariant feature in our proposed method. By combining the JEAs with the JEDM feature, recognition accuracy can be improved by 2∼8% in our experiments. As a consequence, the skeleton recovery module is necessary and contributes to the performance of the proposed method from many different points.

F. Evaluation of the JEAs and JEDM Features

Experiments on three benchmark datasets show that both the JEAs and JEDM features have a good and stable performance in action recognition. The JEAs focuses more locally on each body part, while the JEDM encodes the whole body structure information and pairwise relationships among joints. Generally, the JEDM stream performs 3∼4% better than the JEAs stream, which indicates that in the action recognition the structure variation of the whole body is more discriminative than the dynamic variations of joints. Furthermore, it also verified that a better result can be obtained by fusing the predictions of the two streams compared to each of them, which means the motion information contained by the JEAs

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

Fig. 11. Comparison between the original skeletons and the recovered skeletons in the actions of Northwestern-UCLA dataset: the first two rows are motion sequences of "pick up with one hand", among them the upper images are the original skeletons and the lower images are the recovered skeletons; the second two rows are some comparisons in motions of "pick up with two hands", "sit down", "doffing" and "carry".

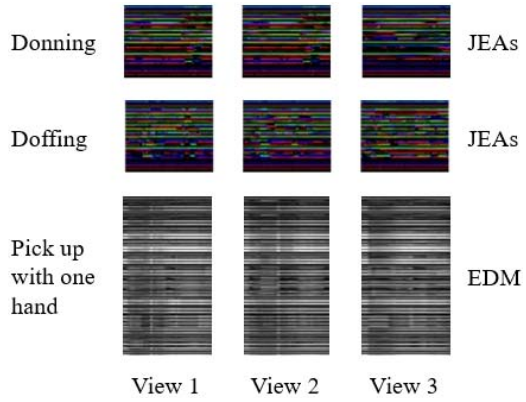


Fig. 12. Motion images from 3 different views of the "donning", "doffing", and "pick up with one hand" actions.

and the JEDM is complementary to each other. In the dataset of MSRC-12 and Northwester-UCLA which have fewer action categories, the fusion of these two streams improves the result by 2∼3% compared with the single JEDM stream. While in the NTU RGB+D dataset, the fusion result achieved nearly 8% improvement. As NTU RGB+D dataset has more action categories and more similar actions than the other two datasets, the local information of each joint plays a more important role in increasing the recognition accuracy compared to experiments on the other two datasets. Different from the ESV method where the different streams are different presentation forms of the same information, the two branches of our two-stream network have different information, which is a critical difference between the proposed method and the ESV method. Inspired by the experimental results, combining the body structure information and the joint information to design better action representations and learning network for human action recognition is one of our following works.

In Fig. 12, motion images of the same action in three different views show high similarity to each other. The consistency of motion images in different viewpoints also proves the efficiency of the JEAs and the JEDM as view-invariant action descriptors.

VI. Conclusion

This paper proposed a method for view-invariant skeleton-based action recognition based on a proposed 3D bio-constrained skeleton model. In the 3D bio-constrained skeleton model, lengths of bones are fixed and joint motion limits recommended by the American Association of Orthopedic Surgeons (AAOS) are utilized. Skeleton recovery with the defined bio-constrained skeleton model is introduced to deal with corrupted skeletons or error joints. Based on the recovered skeleton, two new kinds of view-invariant motion features, the JEAs and the JEDM, are extracted as descriptors of human actions. The two types of features represent the local joint dynamic variations and the global structure information of the human body, respectively. These two descriptors are visualized with images and those encoded motion images are fed into a two-stream CNN for action recognition. Predictions of the two branches are fused together with average strategy

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

at the decision level. Tests on some benchmark datasets, such as MSRC-12 dataset, Northwester-UCLA dataset, and NTU RGB+D dataset, verify the effectiveness of the proposed method. Extensive experiments also demonstrated that JEAs and JEDM are more efficient view-invariant features in describing human actions compared with joint coordinates. Currently, we train the JEAs branch and the JEDM branch separately. Further research will be conducted on combining the local joint information with the body structure information to design better spatio-temporal representations of human action and learning network.

ACKNOWLEDGMENT

The authors would like to thank the help of their colleagues at the CUHK T Stone Robotics Institute of the Chinese University of Hong Kong.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

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

[3] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Naïve-Bayes-nearest-neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 14–19.

[4] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.

[5] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[6] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. IJCAI*, vol. 13, 2013, pp. 2466–2472.

[7] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Jun. 2011, pp. 1297–1304.

[8] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1446–1455.

[9] C. R. M. D and A. Weymann, "The neutral zero method—A principle of measuring joint function," *Injury*, vol. 26, pp. 1–11, Sep. 1995.

[10] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

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

[12] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

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

[14] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1012–1020.

[15] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, Sep. 2016, pp. 816–833.

[16] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI*, Feb. 2017, pp. 1–8.

[17] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[18] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining deep part features for 3-D action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, Jun. 2017.

[19] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[20] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 499–508.

[21] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.

[22] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using LSTM and CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 585–590.

[23] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. (Apr. 2016). "NTU RGB+D: A large scale dataset for 3D human activity analysis." [Online]. Available: https://arxiv.org/abs/1604.02808

[24] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5323–5332.

[25] Y. Shen and H. Foroosh, "View-invariant action recognition using fundamental ratios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–6.

[26] I. N. Junejo, E. Dexter, P. Laptev, and I. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[27] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2458–2466.

[28] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.

[29] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2430–2443, Dec. 2016.

[30] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.

[31] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.

[32] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, 2014.

[33] X. Yang and Y. Tian, "Effective 3D action recognition using EigenJoints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, Jan. 2014.

[34] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 626–633.

[35] Z. Huang, C. Wan, T. Probst, and L. van Gool, "Deep learning on Lie groups for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6099–6108.

[36] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, "Discriminative key pose extraction using extended LC-KSVD for action recognition," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Nov. 2014, pp. 1–8.

[37] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 465–470.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[39] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

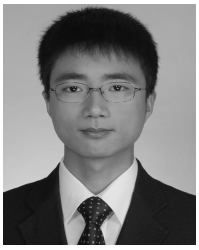
[40] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[41] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2823–2832.

[42] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2649–2656.

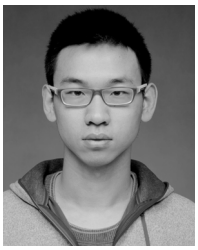
Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.

[43] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2012, pp. 1737–1746.
[44] W. Ding, K. Liu, E. Belyaev, and F. Cheng, "Tensor-based linear dynamical systems for action recognition from 3D skeletons," *Pattern Recognit.*, vol. 77, pp. 75–86, May 2018.
[45] G. G. Demisse, K. Papadopoulos, D. Aouada, and B. Ottersten, "Pose encoding for robust skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, Jun. 188–194.
[46] H. Wang and L. Wang, "Learning content and style: Joint action recognition and person identification from human skeletons," *Pattern Recognit.*, vol. 81, pp. 23–35, Sep. 2018.
[47] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.



Qiang Nie was born in 1990. He received the B.E. degree in mechanical engineering from Nanchang University in 2012 and the M.S. degree in mechanical engineering from Shanghai Jiao Tong University in 2015. He is currently pursuing the Ph.D. degree with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong. His research interests include robot design, artificial intelligence, and deep learning. His current research interests are on human action recognition and human behavior understanding.

Jiangliu Wang was born in 1994. She received the B.E. degree in automation from Nanjing University in 2015. She is currently pursuing the Ph.D. degree with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong. Her research interests include human–robot interaction, action recognition, and computer vision.



Xin Wang was born in 1992. He received the B.E. and M.Sc. degrees in control science and engineering from the Harbin Institute of Technology. He is currently pursuing the Ph.D. degree in mechanical and automation engineering, The Chinese University of Hong Kong. His research interests include objection detection and grasp pose detection.

Yunhui Liu (M'89–F'09) received the B.Eng. degree from the Beijing Institute of Technology, the M.Eng. degree from Osaka University, and the Ph.D. degree from The University of Tokyo in 1992. After working with the Electrotechnical Laboratory of Japan as a Research Scientist, he joined The Chinese University of Hong Kong (CUHK) in 1995, where he is currently a Choh-Ming Li Professor of mechanical and automation engineering and the Director of the CUHK T Stone Robotics Institute. He is also an Adjunct Professor with the State Key Laboratory of Robotics Technology and System, Harbin Institute of Technology, China. He has published over 300 papers in refereed journals and refereed conference proceedings, and was listed in the Highly Cited Authors (Engineering) by Thomson Reuters in 2013. His research interests include visual servoing, medical robotics, multifingered grasping, mobile robots, and machine intelligence. He has received numerous research awards in international journals on robotics and automation from international conferences and government agencies. He was the General Chair of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. He has served as an Associate Editor for the IEEE TRANSACTION ON ROBOTICS AND AUTOMATION. He is the Editor-in-Chief of *Robotics and Biomimetics*.

Authorized licensed use limited to: University of Warwick. Downloaded on February 26,2020 at 10:51:42 UTC from IEEE Xplore. Restrictions apply.