Joint Distance Maps Based Action Recognition With Convolutional Neural Networks

Chuankun Li, Yonghong Hou, *Member, IEEE*, Pichao Wang, *Student Member, IEEE*, and Wanqing Li, *Senior Member, IEEE*

***Abstract*—Motivated by the promising performance achieved by deep learning, an effective yet simple method is proposed to encode the spatio-temporal information of skeleton sequences into color texture images, referred to as joint distance maps (JDMs), and convolutional neural networks are employed to exploit the discriminative features from the JDMs for human action and interaction recognition. The pair-wise distances between joints over a sequence of single or multiple person skeletons are encoded into color variations to capture temporal information. The efficacy of the proposed method has been verified by the state-of-the-art results on the large NTU RGB+D Dataset and small UTD-MHAD Dataset in both single-view and cross-view settings.**

***Index Terms*—Action recognition, convolutional neural networks (ConvNets), joint distance maps (JDM).**

I. Introduction

RECOGNITION of human actions from Red, Green, Blue, and Depth (RGB-D) data has attracted increasing attention [1], [2] in computer vision in recent years due to the recent advance of easy-to-use and low-cost depth sensors such as Kinect sensors. Compared to conventional RGB data, the depth modality has the advantages of being insensitive to illumination changes and reliable to estimate body silhouettes and skeletons [3]. In the past few years, many handcrafted skeleton features [4]–[14] have been proposed for action recognition. Skeleton-based features are often less computationally intensive compared to the features extracted from depth and RGB sequences. However, most existing skeleton-based action recognition methods primarily focus on single subject scenarios and representing spatial information based on hand-designed features. The temporal information is often modeled using dynamic time warpings or hidden Markov models. Recently, several methods based on recurrent neural networks (RNNs) [15]–[19] have also been proposed for skeleton-based action recognition. RNNs tend to overemphasize the temporal information especially when the training data are not sufficient, leading to overfitting. To date, it remains unclear how skeleton sequences of both single and multiple subjects could be effectively represented and fed to deep neural networks for recognition. For example, one can conventionally consider a skeleton sequence as a set of individual frames with some form of temporal smoothness, or as a subspace of poses or pose features, or as the output of a neural network encoder. Which one among these and other possibilities would result in the most effective representation for action recognition is not well understood.

This letter presents an effective yet simple method to encode the pair-wise distances of skeleton joints of single or multiple subjects into texture images, referred to as joint distance maps (JDM), as the input of convolutional neural networks (ConvNets) for action recognition. The proposed method follows a similar approach to that in the works [20], [21]. In [20], the joint coordinates of a skeleton sequence are organized in a matrix, where the three Cartesian components (x, y, z) of joints are seen, respectively, as the three channels (R,G,B) of a color image. Due to the small size of the resultant images, it is impossible to tune an existing ConvNets. Wang *et al.* [21] proposed to encode joint trajectories into texture images and capture temporal information by mapping the trajectories into a hue, saturation, value space. However, this method is view dependent. The method proposed in this letter overcomes the drawbacks of the previous works [20], [21]. First, the size of JDMs is comparable to the image size used to train existing ConvNets and consequently JDMs can be used for fine-tuning a pretrained model. Second, because the three-dimensional (3-D) space distances between joints are view independent, the proposed JDMs are suitable for cross-view action recognition. The proposed method was evaluated on two popular benchmark datasets, NTU RGB+D Dataset [18] which includes actions performed by single and two subjects and UTD-MHAD Dataset [22]. State-of-the-art performance was achieved.

The rest of letter is organized as follows: Section II describes the proposed method. The experimental results on two popular public datasets are presented in Section III. Section IV concludes the letter with remarks and future work.

II. Proposed Method

The proposed method consists of three major components, as illustrated in Fig. 1, the construction of four JDMs as the input of four ConvNets, ConvNets training and the late score fusion. Unlike the previous methods [21], [23]–[25] in which 3-D points are created and projected onto three orthogonal planes of the real-world coordinate system centered on the

Manuscript received January 31, 2017; accepted March 3, 2017. Date of publication March 6, 2017; date of current version March 30, 2017. This work supported in part by the National Natural Science Foundation of China under Grant 61571325 and in part by the Key Projects in the Tianjin Science and Technology Pillar Program under Grant 15ZCZD GX001900. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. *(Corresponding author: Pichao Wang.)*

C. Li and Y. Hou are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: chuankunli@tju.edu.cn; houroy@tju.edu.cn).

P. Wang and W. Li are with the Advanced Multimedia Research Lab, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: pw212@uowmail.edu.au; wanqing@uow.edu.au).

Color versions of one or more of the figrues in this letter are available online at http//ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2678539

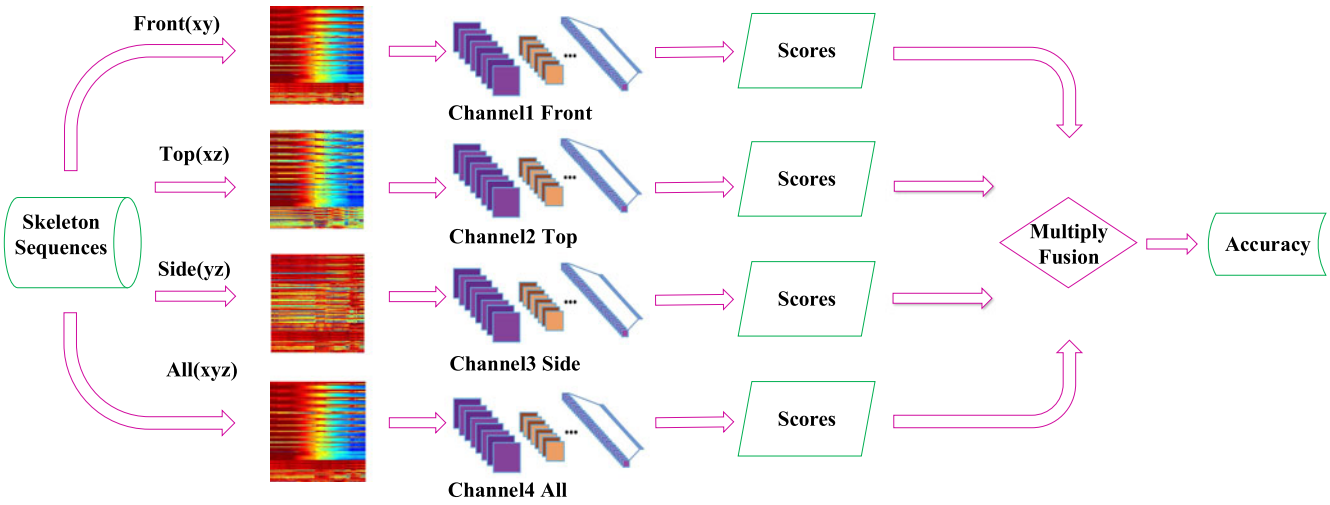

Fig. 1. Framework of the proposed method.

camera to encode a sequence of skeletons (i.e., locations of joints) into three images using the projected joint locations, the proposed method

1) employs pair-wise distances of joints;
2) encodes the distances in the three orthogonal 2-D planes into three JDMs; and also
3) encodes the distances in the 3-D space into the forth JDM.

The information captured by the three JDMs in the three orthogonal 2-D planes is complementary to each other. Importantly, the adoption of pair-wise joint distances and the forth JDM has significantly improved the robustness of the proposed method to viewpoint variations.

Assume a skeleton sequence has t frames and each frame consists of m joints representing the skeletons of single or multiple subjects performing actions or interactions, thus, there are $m \times (m-1)/2$ pair-wise joint distances for each frame. These distances are arranged in one column and a $(m \times (m-1)/2) \times t$ matrix will be generated from the sequence. In order to generate fixed length t' images, bilinear interpolation is applied to scale the number of columns of the distance matrix from t to t'. The distance values in the matrix are then converted into colors to generate a $(m \times (m-1)/2) \times t'$ JDM image so as to encode the spatial-temporal information into colors and texture. Fig. 3 illustrates the process of mapping joint distances to colors. For single subject performing actions, the first person joints are copied to the second person. In the following, the process to generate the four JDMs from the pair-wise distances of joints is detailed.

A. Mapping of Joint Distance (From Skeleton Sequences to Images)

Let $p_j = (px, py, pz)$ be the coordinates of the jth joint in each frame, where $j \in \{1, \ldots, m\}$. For each subject, 12 joints (hip center, spine, shoulder center, head, elbow left, wrist left, hand left, elbow right, wrist right, hand right, ankle left, and ankle right) that have relatively less noise are used in this letter. The m joints of all subjects in each frame can be represented as: $p = \{p_1, p_2, \ldots, p_m\}$ and the numbering of joints follows a fixed order to maintain the correspondence between frames. In this letter, the numbering ranges from subject to subject and within a subject it is from the trunk to right leg, passing through the left arm, right arm, and left leg. An action or interaction instance G of t frames can then be expressed as $G = \{p^1, p^2, \ldots, p^t\}$. The Euclidean distance D^i_{jk} between joints j and k at frame i is denoted as

$$D^i_{jk} = \left\| p^i_j - p^i_k \right\|_2, \quad j, k \in m; j \neq k. \tag{1}$$

The joint distances of all frames in a skeleton sequence are arranged in their temporal order and expressed as follows:

$$H_{jk} = \{D^1_{jk}, D^2_{jk}, \ldots, D^t_{jk}\} \tag{2}$$

where each column represents the pair-wise joint distances of one frame. In such a way, the spatial configuration of joints is described implicitly through the pair-wise distances while the temporal information is represented explicitly. The number of columns of H_{jk} is then scaled from t to t' through bilinear interpolation, i.e.,

$$BL_{jk} = f(H_{jk}) = \{D'^1_{jk}, D'^2_{jk}, \ldots, D'^{t'}_{jk}\} \tag{3}$$

where $f(H_{jk})$ is the interpolation function.

B. Encoding Joint Distance

Inspired by the works [21], [23], [24], where the authors used color to encode the temporal information, Hue is adopted in this letter to encode the variations of joint distances. Specifically, the jet colormap ranging from blue to red is employed though other color schemes may be equally effective as well. Notice that subjects may be of different heights, which lead to the joint distances vary significantly for subjects to subjects. Instead of normalizing the skeleton as in the work [7], the joint distances ranging from zero to the maximum within a sample are mapped to the colormap. Such mapping serves an analog function to normalization

$$H(j,k,i) = \text{floor}\left(\frac{D'^i_{jk}}{\max(BL_{jk})} \times (h_{\max} - h_{\min}) \right), \quad i \in t' \tag{4}$$

where $H(j,k,i)$ denotes the Hue that is assigned to the jkth joint distance in ith frame; $h_{\max}$ and $h_{\min}$ are the range of the Hue.

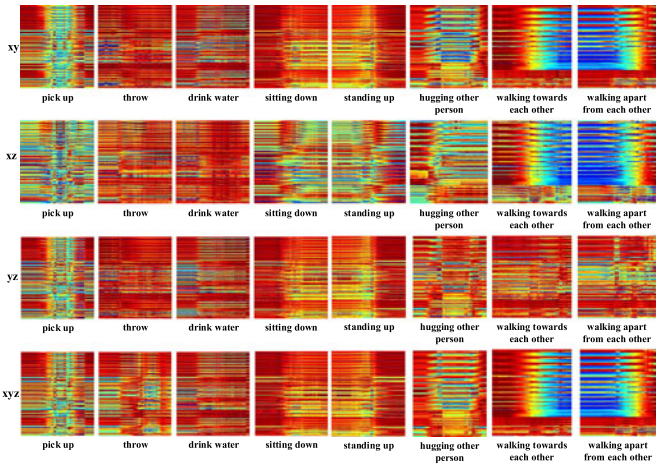

Fig. 2. Sample JDMs generated by the proposed method on the NTU RGB+D dataset.

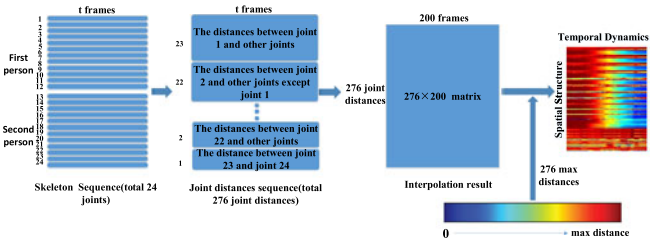

Fig. 3. Process of mapping joint distances to colors.

Four JDMs are generated using the above process, three are generated by calculating the pair-wise distances in the three orthogonal planes (*xy*, *yz,* and *xz*) in the real-world coordinates centered at the camera and the forth JDM is generated in the 3-D space (*xyz*). Fig. 2 shows sample JDMs calculated from sample actions of the NTU RGB+D dataset.

C. Network Training and Class Score Fusion

The original AlexNet is a classic neural network with five convolutional layers and three full-connected layers. It performs very well in image classification. In this letter, the network is fine-tuned on a small dataset and trained from scratch on a large dataset. In fine-tuning, the pretrained models on Large Scale Visual Recognition Challenge 2012, a version of ImageNet (ILSVRC-2012) are used for initialization. As observed, the fine-tuning model performs better (i.e., with higher recognition accuracy) when the training data are not sufficient. However, training from scratch is expected to perform better than fine-tuning on large datasets, especially when the constructed images are very different in texture from the images of ImageNet.

1) Training: In the experiments, the same layer configuration as the one in [26] was adopted for the four ConvNets. The network weights are learned using the mini-batch stochastic gradient descent with the momentum being set to 0.9 and weight decay being set to 0.0005. The batch size was set to 256. All JDMs are resized to 256×256. The learning rate was initially set to 0.001 for fine-tuning and 0.01 for the training from scratch, and then, it is decreased according to a fixed schedule.

TABLE I
COMPARISONS OF THE DIFFERENT SCHEMES ON THE NTU RGB+D AND UTD-MHAD DATASET

Scheme	Dataset		
	NTU RGB+D		UTD-MHAD
	Cross-subject	Cross-view	Cross-subject
F-Front (*xy*)	58.2%	60.2%	79.3%
F-Top (*xz*)	58.7%	57.3%	74.7%
F-Side (*yz*)	62.1%	60.4%	68.8%
F-All (*xyz*)	60.0%	67.9%	80.0%
F-Max-Score fusion	68.5%	73.8%	85.1%
F-Ave-Score fusion	69.3%	74.1%	86.2%
F-Mul-Score fusion	71.5%	77.1%	**88.1%**
S-Front (*xy*)	65.1%	67.4%	68.6%
S-Top (*xz*)	63.5%	63.3%	67.9%
S-Side (*yz*)	62.8%	63.5%	65.4%
S-All (*xyz*)	66.8%	74.9%	73.2%
S-Max-Score fusion	72.9%	79.1%	80.5%
S-Ave-Score fusion	74.4%	80.6%	80.9%
S-Mul-Score fusion	**76.2%**	**82.3%**	85.8%

For NTU RGB+D dataset, the learning was stopped after 35000 iterations and the learning rate decreases every 15 000 iterations. The training was stopped after 900 iterations and the learning rate decreases every 400 iterations on UTD-MHAD dataset. For all experiments, the dropout regularization ratio was set to 0.5 in order to reduce complex coadaptations of neurons in nets.

2) Multiply-Score Fusion: Given a testing skeleton sequence (sample), four JDMs are generated and fed into four different trained ConvNets. Multiply-score fusion was used, that is, the score vectors outputted by the four ConvNets are multiplied in an element-wise way and the max score in the resultant vector is assigned as the probability of the test sequence being the recognized class. The index of this max score corresponds to the recognized class label and expressed as follows:

$$\text{label} = \text{Fin}(\max(v_1 \circ v_2 \circ v_3 \circ v_4)) \tag{5}$$

where v is a score vector, $\circ$ refers to element-wise multiplication, and $Fin(\cdot)$ is a function to find the index of the element having the maximum score.

III. EXPERIMENTAL RESULTS

The proposed method was evaluated on two public benchmark datasets: NTU RGB+D Dataset and UTD-MHAD Dataset. The NTU RGB+D Dataset has 60 actions, 11 of which are interactions between two subjects. Experiments were conducted to evaluate the effectiveness of individual JDMs in the proposed method, three types of score fusion methods and two training models. In all experiments, $h_{\min} = 0$, $h_{\max} = 255$, $t' = 200$, $m = 24$.

A. Evaluation of Individual of JDMs, Fusion Methods, and Training Models

The results of individual JDMs, three fusion methods and two training models are listed in Table I, where F and S represent the models of fine-tuning and training from scratch, respectively. For example, F-Front (*xy*) stands for training the front channel JDM by fine-tuning.

TABLE II
EXPERIMENTAL RESULTS (ACCURACIES) ON NTU RGB+D DATASET

Method	Cross-subject	Cross-view
Lie Group [12]	50.1%	52.8%
Dynamic Skeletons [13]	60.2%	65.2%
HBRNN [15]	59.1%	64.0%
Deep RNN [18]	56.3%	64.1%
Part-aware LSTM [18]	62.9%	70.3%
Deep LSTM [18]	60.7%	67.3%
ST-LSTM [19]	65.2%	76.1%
ST-LSTM + Trust Gate [19]	69.2%	77.7%
JTM [21]	73.4%	75.2%
Proposed Method	**76.2%**	**82.3%**

From Table I, it can be seen that the four JDMs are complimentary to each other, and the multiply-score fusion method improves the final accuracy substantially compared with average and max fusion methods. The fine-tuning performed better on the UTD-MHAD dataset as expected due to its relatively small size. Training the model from scratch works much better on the large NTU RGB+D dataset.

B. NTU RGB+D Dataset

To the best knowledge of authors, NTU RGB+D Dataset [18] is currently the largest action recognition dataset. It was captured by multiple Kincet v2 cameras. The dataset has more than 56 thousand sequences and 4 million frames, containing 60 actions and interactions (11 in total) by 40 subjects aged between 10 and 35. In addition, it has front view, two side views and left, right 45 degree views of the actions. This dataset is challenging due to the large intraclass and viewpoint variations.

We follow the protocols used in [18], namely cross-subject evaluation and cross-view evaluation. In cross-subject evaluation, the IDs of training subjects in this evaluation are: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38; remaining subjects are reserved for testing. In cross-view evaluation, testing set includes the samples of camera 1, and samples of cameras 2 and 3 were used for training. Table II lists the performance of the proposed method and those reported to date.

From this table we can see that our proposed method achieved the state-of-the-art results compared with both hand-crafted features based methods and deep learning methods. The work [12] focused only on single person action and could not model well multiperson interactions. The dynamic skeletons method [13] performed better than some RNN-based methods, implying the weakness of general RNNs [15], [18] that only mines the short-term dynamics and tends to overemphasize the temporal information even on large training data. LSTM and its variants [18], [19] performed better due to their ability to utilize long-term context compared to RNNs but it is still weak in exploiting spatial information. The JTM method [21] encoding the joint trajectories into texture images performed better than the spatio-temporal long short-term memory (ST-LSTM) and Trust Gate methods [19] in cross subject setting but lower in the cross view setting due to the fact that JTMs are sensitive to viewpoints. The proposed method demonstrates its robustness to view variations.

TABLE III
EXPERIMENTAL RESULTS (ACCURACIES) ON UTD-MHAD DATASET

Method	Accuracy(%)
EIC-KSVD [27]	76.19%
Kinect and Inertial [22]	79.10%
Cov3DJ [28]	85.58%
JTM [21]	85.81%
Proposed Method	**88.10%**

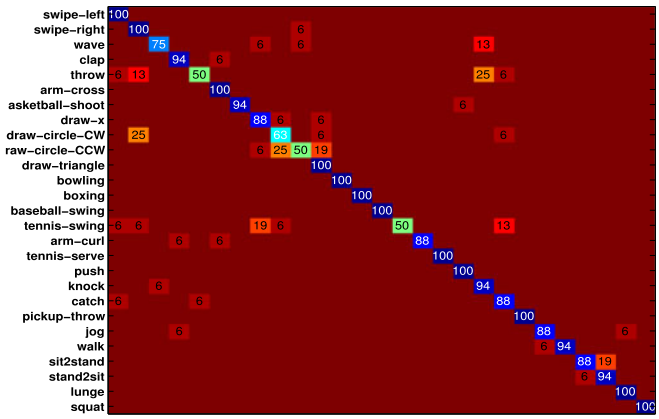

Fig. 4. Confusion matrix of proposed method for UTD-MHAD.

C. UTD-MHAD Dataset

UTD-MHAD [22] is a multimodal action dataset, captured by one Microsoft Kinect camera and one wearable inertial sensor. This dataset contains 27 actions performed by eight subjects (four females and four males) with each subject performing each action four times. In the evaluations, a cross-subject protocol is adopted, that is, odd subjects were used for training and even subjects for testing. Table III compared the performance of the proposed method with the state-of-the-art methods. Notice that both depth and inertial sensor data were used in [22].

The confusion matrix is shown in Fig 4. This dataset is very challenging, because of most actions based on arm motion. From the confusion matrix, we can see that the proposed method still needs improvement in distinguishing some actions of similar pair-wise joint distances, such as "draw−circle−clockwise (CW)" and "draw−circle−counter clockwise (CCW)," which are performed in opposite directions.

IV. CONCLUSION

This letter addresses the problem of skeleton-based human action recognition with ConvNets. A simple yet effective method is proposed that encodes the pair-wise joint distances to color variations, where the spatial-temporal information is converted into texture patterns. This new representation well captures the spatio-temporal information in skeleton sequences, and is suitable for both single-view and cross-view action recognition. State-of-the-art results were obtained on two widely-used datasets and have verified the effectiveness of the proposed method.

REFERENCES

[1] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 9–14.

[2] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–10.

[3] 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.

[4] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 14–19.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.

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

[7] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2752–2759.

[8] Z. Shao and Y. Li, "A new descriptor for multiple 3D motion trajectories recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 4749–4754.

[9] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1351–1357.

[10] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3D discriminative skeletal features for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2013, pp. 471–478.

[11] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2013, pp. 486–491.

[12] 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.*, 2014, pp. 588–595.

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

[14] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *Proc. IEEE Int. Conf. Digital lmage Comput., Tech. Appl.*, 2014, pp. 1–8.

[15] 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.

[16] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4041–4049.

[17] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3697–3704.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1010–1019.

[19] 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.*, 2016, pp. 816–833.

[20] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. Asian Conf. Pattern Recognit.*, 2016, pp. 579–583.

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

[22] 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.

[23] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. O. Ogunbona, "ConvNets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. ACM Multimedia Conf.*, 2015, pp. 1119–1122.

[24] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 4, pp. 498–509, Aug. 2016.

[25] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[27] 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. IEEE Int. Conf. Digital lmage Comput., Tech. Appl.*, 2014, pp. 1–8.

[28] 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. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.