Research Article

3D Skeletal Human Action Recognition Using a CNN Fusion Model

Meng Li¹ and Qiumei Sun²

¹*School of Mathematics and Statistics, Hebei University of Economics and Business, Shijiazhuang, China*

²*Department of Student Affairs Management, Hebei University of Economics and Business, Shijiazhuang, China*

Correspondence should be addressed to Meng Li; mli269-c@my.cityu.edu.hk

Received 13 October 2020; Revised 30 March 2021; Accepted 8 April 2021; Published 22 April 2021

Academic Editor: Luigi Rodino

Copyright © 2021 Meng Li and Qiumei Sun. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Smart homes have become central in the sustainability of buildings. Recognizing human activity in smart homes is the key tool to achieve home automation. Recently, two-stream Convolutional Neural Networks (CNNs) have shown promising performance for video-based human action recognition. However, such models cannot act directly on the 3D skeletal sequences due to its limitation to the 2D image video inputs. Considering the powerful effect of 3D skeletal data for describing human activity, in this study, we present a novel method to recognize the skeletal human activity in sustainable smart homes using a CNN fusion model. Our proposed method can represent the spatiotemporal information of each 3D skeletal sequence into three images and three image sequences through gray value encoding, referred to as skeletal trajectory shape images (STSIs) and skeletal pose image (SPI) sequences, and build a CNNs' fusion model with three STSIs and three SPI sequences as input for skeletal activity recognition. Such three STSIs and three SPI sequences are, respectively, generated in three orthogonal planes as complementary to each other. The proposed CNN fusion model allows the hierarchical learning of spatiotemporal features, offering better action recognition performance. Experimental results on three public datasets show that our method outperforms the state-of-the-art methods.

1. Introduction

Smart homes are the most important technology of sustainable building. In the development of the next generation of smart homes, vision-based action analyzing methods are of key importance since they can make it possible for human occupants to interface with household appliances using just their physical actions instead of mouse, keyboard, touchscreen, or remote control devices. With the rapid development of the RGB-D sensors (e.g., Microsoft Kinect) and real-time full-body tracking with low cost [1, 2], 3D skeletal action analysis has drawn great attentions [3–5]. Compared to the conventional RGB cameras, the RGB-D sensors have the advantage of being reliable to estimate 3D skeletal joint positions of the human body. Different from RGB data, skeletal data is robust to different indoor environment settings. Hence, 3D skeletal action analysis methods are more suited to smart home technology than RGB-based methods. However, it is still a challenging task to efficiently learn proper features of 3D skeletal sequences in

3D human action analysis due to the large spatiotemporal variation in human actions. Recently, two-stream Convolutional Neural Networks (CNNs) have arisen as an advanced technique to recognize human actions in videos and achieve promising recognition performance [6]. Nevertheless, CNNs were proposed for 2D image video input, and it cannot work directly on the 3D skeletal action sequences.

At the test stage, a skeletal sequence is represented by three STSIs and three SPI sequences (multiframe optical flow) that are passed through CNN_F to extract spatial features and temporal features, respectively. These are then pooled across spatial streams and temporal streams and passed through CNN_R to get compact descriptors. Finally, the skeletal action recognition accuracy is obtained by a class score fusion.

In this paper, we propose a novel method consisting of 3D skeletal sequence mapping and CNN fusion model for skeletal action recognition. Using gray value encoding, the proposed skeletal sequence mapping projects each 3D skeletal action sequence into three image sequences and

three images, referred to as skeletal trajectory shape images (STSIs) and skeletal pose image (SPI) sequences. Specifically, given a skeletal sequence, the 3D joint positions of each frame of the skeletal action sequence and the whole skeletal action sequence are encoded into sparse scatter points with gray values and specific size in each of the three orthogonal planes to generate three STSIs and three SPI sequences, respectively. Then, the three STSIs and three SPI sequences are fed into the proposed CNN fusion model for skeletal action recognition.

Our proposed CNN fusion model is built based on the two-stream CNNs [6]. The two-stream CNNs employ two separate CNN streams, a spatial CNN stream, which extracts spatial features from 2D images, and a temporal CNN stream, which operates on optical flow information. Final action recognition accuracy is obtained by a late class score fusion. We extend the two-stream CNNs and introduce a new CNN fusion model that allows three STSIs and three SPI sequences as input to get spatiotemporal features of skeletal trajectory shape and skeletal pose sequence for skeletal action recognition. The overall flowchart of our approach is illustrated in Figure 1. Our main contributions include three aspects.

- (1) We propose a novel CNN fusion model built based on a two-stream architecture with three SPI sequences and three STSIs as input to allow the hierarchical learning of spatiotemporal features of skeletal trajectory shape and skeletal pose sequence for skeletal human action recognition.
- (2) We propose a 3D skeletal sequence mapping method to represent the spatiotemporal information carried in 3D skeletal sequences into three SPI sequences and three STSIs on three orthogonal planes by gray value encoding. The gray value encoding is introduced to project the 3D joint positions of each frame of the skeletal action sequence and the whole skeletal action sequence into sparse scatter points with gray value and specific size in each of the three orthogonal planes.
- (3) The spatial pooling layer and temporal pooling layer are designed to obtain compact spatial features and temporal features in the two-stream architecture.

2. Background

Recently, deep learning methods have been applied on skeletal action recognition [7–9]. In the work by Du et al. [10], the skeletal joints are divided into five sets corresponding to five body parts. They are fed into five LSTMs for feature fusion and classification. In the Zhu et al. [11], the skeletal joints are fed to a deep LSTM at each time slot to learn the inherent co-occurrence features of skeletal joints. In the study by Shahroudy et al. [12], the long-term context representations of body parts are learned with a part-aware LSTM. In the study by Song et al. [13], both the spatial and temporal information of skeletal sequences are learned with a spatial-temporal LSTM. A Trust Gate is also proposed to remove noisy joints. However, RNNs tend to overemphasize

the temporal information especially when the training data is insufficient, thus leading to overfitting [14].

CNNs have also been applied to this problem [14, 15]. Hou et al. [14] proposed the Joint Trajectory Map (JTM), which represents both the spatial configuration and dynamics of joint trajectories into three texture images through color encoding, and then fed these texture images to CNNs for classification. However, each JTM just remains the 2D position information of each frame of the 3D skeletal sequence, which may lose some important information. Li et al. [15] proposed to encode the spatiotemporal information of skeletal sequences into joint distance maps (JDMs), and CNNs are employed to exploit the features from the JDMs for human action recognition. However, this method destroys the spatial structure of each skeletal pose. Pham et al. [16] proposed an Enhanced-SPMF map from skeletal data for the action recognition method; however, this method also partly lost the skeletal structure information of human poses.

Different from all above methods, we propose a novel CNN fusion model that extends two-stream architecture [6] for 3D skeletal action recognition. Our CNN fusion model allows three STSIs and three SPI sequences as input to obtain spatiotemporal features of skeletal trajectory shape and skeletal pose sequence through hierarchical learning. Besides, our gray value encoding can capture the depth information of 3D skeletal joint positions. Hence, our model retains the 3D spatial structure of each human action instead of 2D position information.

3. Proposed Approach

Our proposed approach consists of 2 parts. (1) 3D skeletal sequence mapping. (2) A CNN fusion model for skeletal action classification.

3.1. 3D Skeletal Sequence Mapping. Suppose that each skeleton contains joints, and the illustration of an example skeleton with 15 joints is shown in Figure 2. The 3D skeletal action $A = \{A_1, A_2, \dots, A_T\}$ is performed over T frames, where $A_i = \{P_1, P_2, \dots, P_N\}$ indicates the skeleton of the i th frame and p_j^i represents the 3D coordinates of the j th joint in A_i .

For each skeletal action A , the 3D joint positions of the whole skeletal action sequence A are encoded into sparse scatter points with specific size in each of three orthogonal planes to generate three STSIs, where the three orthogonal planes are defined by the real-world coordinate of the depth sensor. In addition, we propose to use the gray value to represent the depth information of sparse scatter points so that the sparse scatter points can reflect the 3D position information of the skeletal joint. For each skeletal joint p_j^i , the gray value $G(i, j)$ according to its depth information is calculated as follows:

$$G(i, j) = 255 \frac{D(i, j)}{\max(D)}, \quad (1)$$

FIGURE 1: CNN fusion model for skeletal action classification.

FIGURE 2: Skeletal joints captured by the depth sensor.

where $D(i, j)$ denotes the depth value of p_j^i and $\max(D)$ denotes the maximum depth value of A . The depth value represents the distance from the skeletal joint to the corresponding projection plane. Through this process, the 3D position information of each p_j^i is encoded, and the STSIs capture the 3D spatial distribution of joints of the skeletal action which are illustrated in Figure 3.

Similarly, for each frame of the skeletal action, the 3D joint positions of A_i are represented by SPI using gray value encoding. The multiframe optical flows of SPI sequences are illustrated in Figure 4.

3.2. CNN Fusion Model. We design the CNN fusion model (Figure 1) for skeletal action recognition. Each three STSIs

are passed through the first part of the spatial stream CNN_F separately, aggregated at a spatial-pooling layer, and then sent through the remaining part of the network CNN_R . Similarly, the multiframe optical flows of each three SPI sequences are passed through the first part of the temporal stream CNN_F separately, aggregated at a temporal-pooling layer, and then sent through the remaining part of the network CNN_R . All branches in CNN_R share the same parameters. For the spatial CNN stream and temporal CNN stream, we use element-wise maximum operation across the spatial CNN stream in the spatial-pooling layer and temporal CNN stream in the temporal-pooling layer. These two pooling layers should be placed close to the last convolutional layer for optimal action classification. For each branch, our CNN fusion model consists of mainly five convolutional layers followed by two fully connected layers and a softmax classification layer. The model is pretrained on ImageNet images and then fine-tuned on all skeletal action sequences in the training dataset.

4. Experiments

4.1. Experimental Datasets. Experiments were conducted on three public benchmark datasets: Florence3D-Action dataset [17], Toyota Smarthome dataset [18], and NTU RGB + D dataset [12].

Florence3D-Action dataset has been captured using a Kinect camera. It includes 9 actions: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. During acquisition, 10 subjects were asked to perform the above actions. This resulted in a total of

FIGURE 3: STSIs in the three Cartesian planes for action “wave” in Florence3D-actions’ dataset: (a) the front (xy) plane; (b) the top (xz) plane; (c) the side (yz) plane.

FIGURE 4: Multiframe optical flows of SPI sequences in the three Cartesian planes for action “wave” in Florence3D-actions’ dataset: (a) the front (xy) plane; (b) the top (xz) plane; (c) the side (yz) plane.

FIGURE 5: Sample frames of the Florence3D-Action dataset. The red points in each frame were plotted according to the 3D positions of skeletal joints obtained using the method of [1]. (a) Wave. (b) Drink from a bottle. (c) Bow. (d) Read watch. (e) Sit down. (f) Answer phone.

215 action samples. The 3D positions of 15 skeletal joints are provided with the dataset. The 3D skeletal joints were obtained using the method of [1]. Sample frames of the Florence3D-Action dataset are shown in Figure 5.

Toyota Smarthome dataset is a very challenging large dataset. It has been captured using 7 Kinect cameras. It includes 31 daily living actions: walk, drink from cup, sit down, read book, get up, watch TV, eat at table, stir, use

FIGURE 6: Sample frames of the Toyota Smarthome dataset. The red points in each frame were plotted according to the 3D positions of skeletal joints obtained using the method of [18]. (a) Walk. (b) Lay down. (c) Read book. (d) Drink from cup. (e) Stir. (f) Get up.

FIGURE 7: Sample frames of the NTU RGB + D dataset. The red points in each frame were plotted according to the 3D positions of skeletal joints obtained using the method of [12]. We just select 13 skeletal joints of each pose for convenient viewing. (a) Brushing hair. (b) Dropping. (c) Wearing jacket. (d) Wearing jacket. (e) Patting others' back. (f) Handshaking.

telephone, enter, leave, use laptop, clean up, clean dishes, take pills, drink from bottle, pour from bottle, eat snack, lay down, cut, pour from kettle, use stove/oven, pour water, drink from glass, pour grains, boil water, pour from can, from can, insert teabag, use tablet, and cut bread. The subjects are senior people in the age range 60–80 years old.

During acquisition, 18 subjects were aware of the recording but they were unaware of the purpose of the study. The videos were clipped per action, resulting in a total of 16,115 action samples. The 3D positions of 13 skeletal joints are provided with the dataset. The 3D skeletal joints were obtained using the method of [18]. Sample frames of the Toyota

TABLE 1: Comparisons with other skeletal action recognition methods on the Florence 3D-Action dataset.

Approach	Accuracy (%)
Multipart bag-of-poses [17] (2013)	82.0
Motion trajectories [20] (2015)	87.0
Lie group [21] (2014)	90.9
Bi-LSTM [22] (2018)	93.0
MIMTL [23] (2017)	95.3
Optical spectra [14] (2018)	95.6
JDM [15] (2017)	96.1
Proposed (spatial)	97.3
Proposed (temporal)	95.1
Proposed (fusion)	98.2

TABLE 2: Comparisons with other skeletal action recognition methods on the Toyota Smarthome dataset.

Approach	CS	CV1	CV2
HMM [24] (2016)	25.7	10.2	13.4
DT [25] (2011)	41.9	20.9	23.7
LSTM [26] (2016)	42.5	13.4	17.2
I3D [27] (2017)	53.4	34.9	45.1
I3D + NL [28] (2018)	53.6	34.3	43.9
Optical spectra [14] (2018)	53.7	31.2	43.5
JDM [15] (2017)	54.0	34.5	44.6
Separable STA [18] (2019)	54.2	35.2	50.3
DPI [29] (2019)	54.5	36.0	52.8
Proposed (spatial)	54.4	35.7	52.5
Proposed (temporal)	51.1	30.2	47.9
Proposed (fusion)	55.0	36.3	53.2

TABLE 3: Comparisons with other skeletal action recognition methods on the NTU RGB + D Dataset.

Approach	CS	CV
Lie group [21] (2014)	50.1	52.8
Part-aware LSTM [12] (2016)	62.9	70.3
ST-LSTM + Trust Gate [30] (2016)	69.2	77.7
STA-LSTM [31] (2017)	73.4	81.2
GCA-LSTM [32] (2017)	74.4	82.8
VA-LSTM [33] (2017)	79.4	87.6
SR-TSL [33] (2017)	84.8	92.4
HCN [34] (2018)	86.5	91.1
2s-AGCN [35] (2019)	88.5	95.1
AGC-LSTM [36] (2019)	89.2	95.0
DGNN [37] (2019)	89.9	96.1
NAS-GCN [38] (2020)	89.4	95.7
Proposed (spatial)	88.3	93.6
Proposed (temporal)	83.1	86.8
Proposed (fusion)	90.0	95.2

Smarthome dataset are shown in Figure 6. The NTU RGB + D dataset is a large-scale dataset for human action recognition (Figure 7). It includes 60 actions: drinking, eating, brushing teeth, brushing hair, dropping, picking up, throwing, sitting down, standing up, clapping, reading, writing, tearing up paper, wearing jacket, taking off jacket, wearing a shoe, taking off a shoe, wearing on glasses, taking off glasses, putting on a hat/cap, taking off a hat/cap,

cheering up, hand waving, kicking something, reaching into self pocket, hopping, jumping up, making/answering a phone call, playing with phone, typing, pointing to something, taking selfie, checking time, rubbing two hands together, bowing, shaking head, wiping face, saluting, putting palms together, crossing hands in front, sneezing/coughing, staggering, falling down, touching head, touching chest, touching back, touching neck, vomiting, fanning self,

FIGURE 8: Confusion matrix for the Florence3D-Action dataset.

FIGURE 9: Confusion matrix for the Toyota Smarthome dataset in cross-subject (CS) protocol.

FIGURE 10: Confusion matrix for the Toyota Smarthome dataset in cross-view 1 (CV1) protocol.

punching/slapping other person, kicking other person, pushing other person, patting others back, pointing to the other person, hugging, giving something to other person, touching other persons pocket, handshaking, walking towards each other, and walking apart from each other. There are two protocols in this action dataset, i.e., Cross-Subject (CS) and Cross-View (CV). The 3D skeletal joints were obtained using the method of [12].

4.2. Implementation Details. In our CNN fusion model, the locations of spatial and temporal pooling layers are placed to the fourth convolutional layers of both spatial and temporal streams. As in the study by Wang et al. [19], both the spatial stream and temporal streams of our CNN fusion model are pretrained on ILSVRC-2012. After obtaining the STSIs and SPI sequences, our CNN fusion model is fine-tuned on all skeletal human activity training sets. The images according to the input STSI and SPI sequences are resized to 256×256 . The weights of our CNN fusion model are learned using the mini-batch stochastic gradient descent with the momentum being set to 0.7 and weight decay being set to 0.0003. The learning rate is initially set to 0.001, and then, it is set to 0.0001 for every 5,000 iterations and stops at 20,000 iterations.

4.3. Experimental Results. As shown in Tables 1 and 2, for the Florence3D-Action dataset and the Toyota Smarthome

dataset, the proposed approach achieves higher recognition accuracies than other existing methods since our proposed CNN fusion model can allow hierarchical learning to get spatiotemporal features of skeletal trajectory shape and skeletal pose sequence for skeletal action recognition. To allow a fair comparison, for the Florence3D-Action dataset, we followed the cross-subject test setting of [18], for the Toyota Smarthome dataset, we followed the cross-subject (CS) and two cross-view (CV1 and CV2) protocols of [18], and for the NTU RGB + D dataset, we followed the cross-subject (CS) and cross-view (CV) protocols of [12]. On the NTU RGB + D dataset, our method works very well although our method performs less well than DGNN for cross-view protocol. However, DGNN needs intensive computation, which may restrict its real application (Table 3).

Figure 8 shows the confusion matrix for the Florence3D-Action dataset. We can see that our approach works very well. The classification confusions occur when the two actions are highly similar to each other like “drink from a bottle” and “answer phone” in the case of the Florence3D-Action dataset.

Figures 9–11 show the confusion matrices for the Toyota Smarthome dataset. The Toyota Smarthome dataset contains 31 daily living actions. According to [18], we use 31 daily living actions for CS protocol and 19 daily living actions for CV1 and CV2 protocols. For each action sample in the Toyota Smarthome dataset, the subject performs actions

FIGURE 11: Confusion matrix for the Toyota Smarthome dataset in cross-view 2 (CV2) protocol.

without any information about how to perform it. Hence, the dataset consists of a rich variety of actions with huge intraclass differences. Although the dataset is very challenging, the confusion matrices show that our method can still get good effect for many types of actions.

5. Conclusions and Future Work

In this paper, we have proposed a novel CNN fusion model for 3D human skeletal action recognition in smart homes. First, we propose the 3D skeletal sequence mapping to represent the spatiotemporal information of each 3D skeletal sequence into three images and three image sequences through gray value encoding, referred to as skeletal trajectory shape images (STSIs) and skeletal pose image (SPI) sequences. Then, we construct a CNNs fusion model with three STSIs and three SPI sequences as input for skeletal action recognition. Our experimental results have shown that our proposed approach has superior performance in comparison with several state-of-the-art methods on two public action datasets.

In future work, we will consider introducing a new mechanism to deeply mine the human-object patterns of human-object actions in smart homes. It can improve the effectiveness of this model in real-world applications. In addition, the features surrounding 3D skeletal data should be captured for recognizing human-object actions. This will be another focus of our future research.

Data Availability

The data used to support the findings of this study are included within this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Foundation of Hebei Department of Human Resources and Social Security (no. C201810 “河北省引进留学人员资助项目(课题)” (in Chinese)), Natural Science Foundation of Hebei Province (no. F2019207118), Foundation of Hebei Educational Department (no. ZD2021319), and Foundation of Hebei University of Economics and Business (no. 2019PY01).

References

- [1] J. Shotton, A. Fitzgibbon, M. Cook et al., “Realtime human pose recognition in parts from single depth images,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, Colorado Springs, CO, USA, June 2011.
- [2] G. Rogez, P. Weinzaepfel, and C. Schmid, “LCR-Net++: multi-person 2D and 3D pose detection in natural images,”

- IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, 2019.
- [3] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: a survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.
 - [4] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging*, pp. 149–187, Springer, Berlin, Germany, 2013.
 - [5] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: a review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.
 - [6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 568–576, 2014.
 - [7] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: a survey," *Image and Vision Computing*, vol. 60, pp. 4–21, 2017.
 - [8] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio et al., "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proceedings of 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 476–483, Washington, DC, USA, May 2017.
 - [9] P. Wang, W. Lia, P. Ogunbona et al., "RGB-D-based human motion recognition with deep learning: a survey," *Computer Vision & Image Understanding*, vol. 118–139, 2017.
 - [10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Boston, MA, USA, June 2015.
 - [11] W. Zhu, C. Lan, J. Xing et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proceedings of 30th AAAI Conference on Artificial Intelligence (AAAI-16)*, pp. 3697–3703, Phoenix, AZ, USA, February 2015.
 - [12] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+ D: a large scale dataset for 3D human activity analysis," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, Las Vegas, NV, USA, June 2016.
 - [13] S. Song, C. Lan, J. Xing et al., "Spatio-temporal attention based LSTM networks for 3D action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, p. 3459, 2018.
 - [14] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, 2018.
 - [15] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.
 - [16] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, S. A. Velastin, and P. Zegers, "A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera," *Sensors*, vol. 20, no. 7, p. 1825, 2020.
 - [17] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 479–485, Portland, OR, USA, June 2013.
 - [18] S. Das, R. Dai, M. Koperski et al., "Toyota Smarthome: real-world activities of daily living," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, October 2019.
 - [19] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of ECCV 2016-European Conference on Computer Vision*, Glasgow, UK, August 2016.
 - [20] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.
 - [21] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595, Columbus, OH, USA, June 2014.
 - [22] B. T. Amor, D. Hassen, and B. A. Boulbaba, "Coding kendall's shape trajectories for 3d action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
 - [23] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3d action recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 519–529, 2017.
 - [24] C. Xingquan, G. Yufeng, L. Mengxuan et al., "Infrared human posture recognition method for monitoring in smart homes based on hidden markov model," *Sustainability*, vol. 8, no. 9, pp. 892–899, 2016.
 - [25] H. Wang, K. Alexander, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, Colorado Springs, CO, USA, June 2011.
 - [26] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3054–3062, Las Vegas, NV, USA, June 2016.
 - [27] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, IEEE, Honolulu, HI, USA, July 2017.
 - [28] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.
 - [29] M. Liu, F. Meng, C. Chen et al., "Joint dynamic pose image and space time reversal for human action recognition from videos," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, HI, USA, January 2019.
 - [30] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proceedings of European Conference on Computer Vision ECCV 2016*, pp. 816–833, Springer, Amsterdam, Netherlands, October 2016.
 - [31] 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 *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4263–4270, San Francisco, CA, USA, February 2017.
 - [32] J. Liu, G. Wang, P. Hu, L. Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1647–1656, Honolulu, HI, USA, July 2017.

- [33] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2117–2126, Venice, Italy, October 2017.
- [34] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," arXiv preprint arXiv: 1804.06055, 2018.
- [35] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, Long Beach, CF, USA, June 2019.
- [36] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1227–1236, Long Beach, CF, USA, June 2019.
- [37] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7912–7921, Long Beach, CF, USA, June 2019.
- [38] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 3, pp. 2669–2676, 2020.