SKELETON-BASED HUMAN ACTION RECOGNITION USING SPATIAL TEMPORAL 3D CONVOLUTIONAL NEURAL NETWORKS

Juanhui Tu[1], *Mengyuan Liu*[2], *Hong Liu*[1,*]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
{juanhuitu@pku.edu.cn, liumengyuan@ntu.edu.sg, hongliu@pku.edu.cn}

ABSTRACT

It remains a challenge to extract spatial-temporal information from skeleton sequences for 3D human action recognition. Although most recent action recognition methods based on Recurrent Neural Networks (RNN) have achieved outstanding performance, one of the shortcomings of these methods is the tendency to overemphasize the temporal information. Since 3D Convolutional Neural Networks(3D CNN) can simultaneously learn features from both spatial and temporal dimensions through capturing correlations among three-dimensional signals, this paper proposes a novel two-stream model using 3D CNN. To our best knowledge, this is the first attempt to use 3D CNN in the field of skeleton-based action recognition. Our method consists of three stages. First, skeleton joints are mapped into a 3D coordinate space to encode the spatial and temporal information. Second, 3D CNN models are separately employed to extract deep features from both spatial and temporal stream. Third, to enhance the ability of discriminative features to capture global relationships, we extend each stream into multi-temporal version. Extensive experiments on the large-scale NTU RGB-D dataset and the public SmartHome dataset demonstrate that our method outperforms most of RNN-based methods, which verify the complementary property between spatial and temporal information and the robustness to noise.

Index Terms— 3D Human Action Recognition, Skeleton Sequences, 3D Convolutional Neural Networks

1. INTRODUCTION

Human action recognition has been widely applied in various applications, including intelligent surveillance, human-computer interaction and video analysis [1, 2, 3, 4]. 3D representation of human action provides more comprehensive and discriminative information than 2D RGB videos. During recent years, the skeleton-based 3D action recognition has been attracting increasing attention due to its high level representation and robustness to appearances and surrounding distractions.

Recurrent Neural Networks (RNN) [5], [6] have been used to model temporal evolutions of skeleton sequences [7]. These RNN-based methods tend to overstress the temporal information [8]. However, for a given skeleton sequence, there are two important factors to recognize action classes: one is the description of the spatial structure of skeleton joints, and the other is to extract temporal information among multiple frames of the sequence. Hence, the combination of spatial and temporal information is the most effective representation. Considering that 3D Convolutional Neural Networks can extract correlations among high-dimensional signals by performing 3D convolutions [9], we present a two-stream 3D CNN model for skeleton-based action recognition.

To extract correlations by 3D convolutions, the well-designed spatial-temporal encodings of skeleton joints through mapping into a 3D coordinate space to encoded into volume as input. Therefore, the spatial and temporal information can be effectively learned by 3D CNN simultaneously. In addition, the two-stream consists of spatial and temporal streams, which compensate for each other to enhance the representation of spatial and temporal information.

2. RELATED WORK

2.1. RNN-based Methods

Most recent action recognition methods are based on Recurrent Neural Networks. Du *et al.* [7] proposed an end-to-end hierarchical RNN to encode the relative motion between skeleton joints. Skeletons were split into anatomically-relevant parts, which were fed into each independent subnet to extract local features. Since LSTM can learn long-term and short-term dependencies in the input sequences using special gating schemes, many works chose LSTM to learn features. Shahroudy *et al.* [10] proposed a part-aware LSTM which has part-based memory sub-cells and a new gating mechanis-

This work is supported by National Natural Science Foundation of China (NSFC, No.U1613209,61340046,61673030), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004), Specialized Research Fund for Strategic and Prospective Industrial Development of Shenzhen City (No.ZLZBCXLJZI20160729020003), Scientific Research Project of Shenzhen City (No.JCYJ20170306164738129), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No.ZDSYS201703031405467). Hong Liu* is the Corresponding author.

978-1-5386-1737-3/18/$31.00

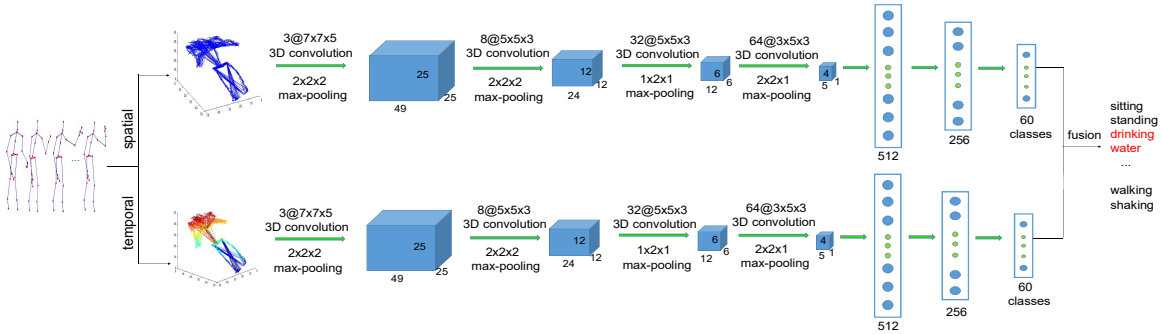


Fig. 1: Overall pipeline of the proposed two-stream 3D CNN

m, showing that LSTM outperforms some hand-crafted features and RNN. However, RNN-based methods tend to focus on the representation of temporal information [8].

2.2. 3D CNN-based Methods

3D CNN has been established as a natural and suitable choice for action recognition, object recognition [11], vehicle detection [12] and human pose estimation [13] to receive a 3-dimensional input. It was proposed for RGB sequence-based action recognition [9], [14] for the first time. 3D convolutional layer takes a volume as input and outputs a volume. Both spatial information and temporal information are abstracted layer by layer from the low-level features to high-level features. Tran *et al.* [15] proposed a simple, yet effective approach to spatial-temporal feature learning using 3-dimensional convolutional neural network, which verified that 3D CNN achieves faster and higher accuracy. Especially, the features used in [15] have four properties for an effective video descriptor: generic, compact, efficient and simple. Cao *et al.* [16] provided a more effective and robust joints-pooled 3D deep convolutional descriptor (JDD), generating promising results on real-world datasets. In general, 3D CNN can capture correlations among three-dimensional signals thereby exploring distinctive spatial-temporal information.

Our main contribution lies in three aspects: (1) Well-designed spatial-temporal encodings of skeleton sequences are quite effective for 3D CNN to learn. (2) We originally propose a novel two-stream 3D CNN model, which is mutually compensated and robust to noise. Especially, it can effectively avoids overfitting. (3) To the best of our knowledge, this is the first attempt to use 3D CNN for skeleton-based action recognition, which achieves competitive performances on challenging datasets.

3. TWO-STREAM 3D CNN

This section illustrates the pipeline (see Fig. 1) of the proposed two-stream 3D CNN for skeleton-based action recognition. First, a sequence-based transform method is used, which eliminates the effect of view variations. Second, the spatial information and temporal information among multiple frames are encoded into spatial volume and temporal volume respectively, which capture the spatial structure of body and emphasize the chronological order. Third, the 3D CNN is capable to learn spatial and temporal features. In contrast, RNN-based methods provide good temporal modeling but lack the combination of spatial-temporal. Finally, original skeleton sequences are convert into multi-temporal sequences to capture large scale of temporal information.

3.1. Spatial and Temporal Volume

Since different views impact the appearance of skeletons, a spatial transform proposed by Liu *et al.* [17] is adopted as a preprocessing step to solve the problem of viewpoint variations. Assuming there are F frames in an action and each skeleton consists of M joints, the *m-th* skeleton joint on the *f-th* frame is formulated as $p_m^f = (x_m^f, y_m^f, z_m^f)^T$, where $f \in (1, ..., F)$, $m \in (1, ..., M)$. For there are limited marked joints in each action sequence, the interpolation operation between the consecutive joints is applied to enrich joint information. Next, skeleton joints from each action sequence are mapped into a 3D coordinate space S and then encoded into spatial volume and temporal volume separately. Particularly, it is not only effective for 3D CNN to capture correlations but also to solve the problem of inconsistent frames for each skeleton sequence to retain complete moving information. Let $F_s(x_m^f, y_m^f, z_m^f)$ be spatial value in spatial volume indicating regions of motion, which represents the encoded spatial information.

$$F_s\left(x_m^f, y_m^f, z_m^f\right) = \begin{cases} 1, & if\ p_m^f \in S \\ 0, & otherwise \end{cases} \tag{1}$$

Considering that it is difficult to recognize two actions with similar motion regions but reverse chronological order, such as actions *"standing up"* and *"sitting down"*. Therefore, let $F_t(x_m^f, y_m^f, z_m^f)$ represent time value by adding corresponding frame number of action to time volume. For $F_t(x_m^f, y_m^f, z_m^f)$ reflects the order of time, it can distinguish them successfully. For the results presented here, a simple replacement for use is defined as:

$$F_t(x_m^f, y_m^f, z_m^f) = norm(f), \qquad f \in (1, 2, ..., F) \tag{2}$$

where function *norm* indicates that $F_t(x_m^f, y_m^f, z_m^f)$ is normalized to [0,1]. Compared with $F_s(x_m^f, y_m^f, z_m^f)$,

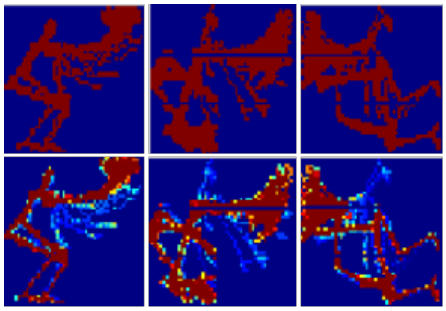


Fig. 2: Illustration of spatial volume (upper) and temporal volume (lower) of action *"hand waving"* from three orthogonal planes

$F_t(x_m^f, y_m^f, z_m^f)$ encodes temporal information of action sequences. As shown in Fig. 2, it illustrates the difference between spatial encoding and temporal encoding of action *"hand waving"*. It can be seen that temporal encoding captures the temporal variations. The deeper the color, the more backward the time sequence. Importantly, the adoption of fusing the spatial and temporal features reinforces each other to achieve better performance.

3.2. Two-Stream 3D CNN Model

A simple yet effective network for 3D convnet proposed in this paper. As shown in Fig. 1, to compare the performance of spatial encoding and temporal encoding, the architecture of spatial stream is the same as that of temporal stream to be consistent. For individual stream, the 3D CNN network is comprised of four layers of 3D convolution, each followed by a max-pooling and two fully connected layers. As same as [18], the filter numbers for each convolution layer are 3, 8, 32 and 64, respectively and 512, 256 neurons for fully connected layers separately. And the kernel size of filters are $7 \times 7 \times 5$, $5 \times 5 \times 3$, $5 \times 5 \times 3$, $3 \times 5 \times 3$ respectively for convolutional layers. Particularly, to reduce overfitting and improve the generalization of classifier, dropout layer [19] is added between the convolution layer and the max-pooling layer to eliminate overfitting. In addition, the padding is used after the first three convolution layer in order to make sure that the input size is equal to the output size of the convolution operation to guarantee the number of convolution layer. Most importantly, the full model only has 910K parameters.

The spatial stream and temporal stream are trained separately and fused during the forward propagation stage for decision making. For each network, with weight parameters W_S and W_T respectively, the class-membership probabilities for classes C given the action's observation x are represented as $(P(C|x, W_S), P(C|x, W_T))$. To compute the final class-membership probabilities for the action recognition classier, the class-membership probabilities are multiplied element-wisely from the two-stream network:

$$P(C|x) = P(C|x, W_S) * P(C|x, W_T) \tag{3}$$

Then, the class label of c^* is calculated as $c^* = argmax\ P(C|x)$. All classification outputs are softmax layer as Eq.(4) and trained with cross-entropy loss as Eq.(5). Furthermore, for activation functions, all the layers in the networks use the rectified linear unit(ReLU), represented as $f(x) = max(0, x)$.

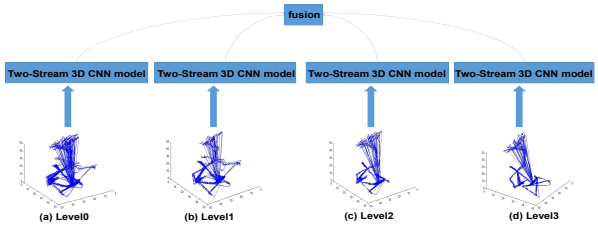


Fig. 3: Multi-Temporal structure for action *"hand waving"*

$$P(C|W) = \frac{exp(x_C)}{\sum_k exp(x_k)} \tag{4}$$

where x_k is the output of the neuron k.

$$\begin{aligned} L(W, D) = &- \frac{1}{|D|} \sum_{i=0}^{|D|} [ylog(P(C^{(i)}|x^{(i)}, W)] - \\ &\frac{1}{|D|} \sum_{i=0}^{|D|} [(1-y)log(1 - P(C^{(i)})|x^{(i)}, W)] \end{aligned} \tag{5}$$

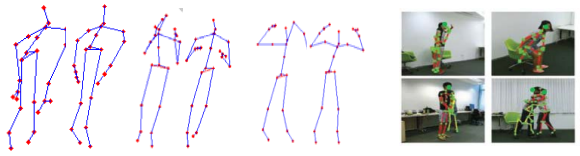
where D represents the trained dataset and y indicates the true class label of each action sequence.

3.3. Multi-Temporal Structure

The method of implementing 3D CNN model with different scale of convolutional filters can extract more discriminative information and capture large scale of temporal information. However, this way adds complexity of the 3D CNN model. This paper converts original skeleton sequences into multi-temporal sequences, and then uses two-stream 3D CNN model to extract deep features, respectively. As shown in Fig. 3, 3D volume represents the volume of encoding spatial and temporal information. Then multi-temporal 3D volumes are trained by the two-stream 3D CNN model respectively and fused to get final the result. Specifically, given a skeleton sequence with F frames, "Level 0" represents the entire skeleton sequences; "Level 1" represents the subsequence from the beginning to the $\lfloor F/2 \rfloor - th$ frame; "Level 2" represents the subsequence from $\lfloor F/4 \rfloor - th$ frame to $\lfloor 3F/4 \rfloor - th$ frame; "Level 3" represents the subsequence from $\lfloor F/2 \rfloor - th$ frame to the end. 3D volumes extracted from different temporal levels not only capture the multi-scale specific local patterns, but also enhance the global relationships.

4. EXPERIMENTS AND ANALYSIS

The proposed method is evaluated on two public benchmark datasets: NTU RGB+D Dataset and SmartHome Dataset. Ex-



(a) Reverse time series action pairs (b) Noisy skeletons

Fig. 4: Snaps from the NTU RGB+D dataset [10]

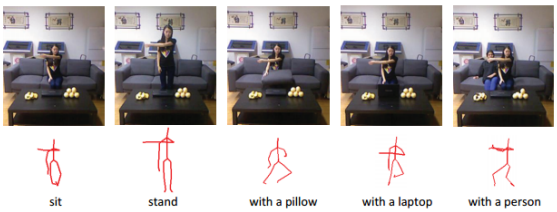


Fig. 5: Skeletons of action *"wave"* in SmartHome dataset [17]

periments are conducted to evaluate the effectiveness of the two-stream 3D CNN model.

4.1. Datasets and Settings

NTU RGB+D dataset contains 56880 sequences of 60 classes performed by 40 subjects and captured by three cameras. It is a very challenging dataset due to different sequence length, reverse time series action pairs and noisy skeleton joints. Some snaps are shown in Fig. 4. To ensure a fair comparison, we follow the two standard protocols used in the [10]. In cross-subject evaluation, the 40 subjects are split into training and testing groups. Each group contains 20 subjects. For cross-view evaluation, all samples of camera 1 are picked for testing and samples of cameras 2 and 3 are picked for training. SmartHome dataset[1] [17] is collected by our lab, which contains six types of actions: "box", "high wave", "horizontal wave", "curl", "circle", "hand up". Each action is performed 6 times (three times for each hand) by 9 subjects in 5 situations: "sit", "stand", "with a pillow", "with a laptop", "with a person", generating 1620 depth sequences. Skeleton joints in SmartHome dataset contain much noises, due to the effect of occlusions and the unconstrained poses of action performers. Skeletons of action *"wave"* are shown in Fig. 5. For evaluation, subjects #1, 3, 5, 7, 9 are used for training and subjects #2, 4, 6, 8 are used for testing.

Normalization step applied on the joint coordinates by translating them to a body centered coordinate system with the "middle of the hip" joint as the origin. For the mapped 3D coordinate space S, the width height is set to be 50. The network weights are learned using the mini-batch stochastic gradient descent with learning rate set to 0.0005, momentum value set to 0.9 and weight decay set to 1.0e-6. The size of minibatches is 32 and the probability of dropout is 0.3. We randomly sample 10% of the initial training set as a validation set for hyper-parameter optimization.

[1] It is provided in https://github.com/NewDataset/dataset.git.

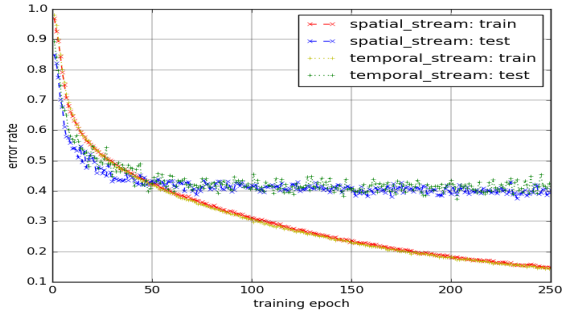


Fig. 6: Convergence curves on the NTU RGB+D dataset [10]

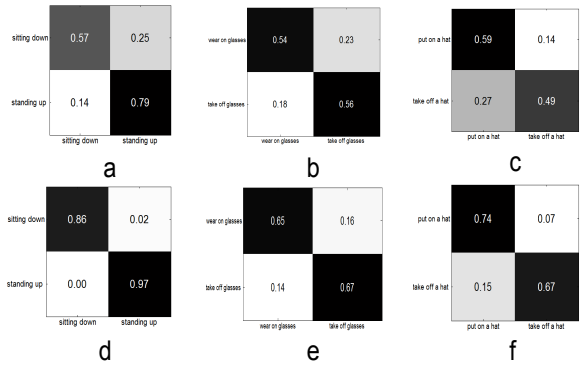


Fig. 7: Confusion matrix of some representative actions for a to c using spatial-stream 3DCNN network and d to f using two stream 3DCNN network

4.2. Evaluation of Two-Stream 3D CNN

Fig. 6 shows the convergence curves on the NTU RGB+D dataset for spatial stream and temporal stream, where the error rate tends to converge when the training epoch grows to 250. The result verifies the effectiveness of the 3D CNN architecture. Table I evaluates two-stream 3D CNN model method. By fusing the spatial stream and temporal stream, it has no obvious effect on SmartHome dataset for cross subject evaluation. Because SmartHome dataset does not contain similar action pairs that has opposite chronological order. On the contray, two-stream 3D CNN respectively achieves 5.46% and 5.90% higher than individual stream on NTU RGB+D dataset for cross-view evaluation. These improvements verify that the two-stream can mutually reinforce. Furthermore, three pairs of representative actionss confusion matrix as shown in Fig. 7. Like action pairs "sitting down" and "standing up", for spatial stream, the probability of classifying sitting down to standing up is 0.25. While for two-stream, the probability drops to 0.02. It can be seen that the error rates of mutual recognition has a reduction relatively.

Table 1: Evaluation of two-stream 3D CNN model

Method	Dataset		
	SmartHome (CS)(%)	NTU RGB+D (CS)(%)	NTU RGB+D (CV)(%)
Spatial Stream	78.61	56.06	62.41
Temporal Stream	71.32	56.22	61.97
Two-Stream	**79.38**	**62.13**	**67.87**

Table 2: Results of multi-temporal scheme on NTU RGB+D dataset

Method	CS(%)	CV(%)
Level 0	62.13	67.87
Level 1	52.30	56.53
Level 2	53.42	58.49
Level 3	52.87	57.68
Level 0+1+2+3	**66.85**	**72.58**

4.3. Evaluation of Multi-Temporal Structure

The level l of our two-stream 3D CNN model is considered to have notable impact on the performance. Table II show the recognition accuracies with different values of l from 0 to 3. It can be observed that our method achieves the best performance on the NTU RGB+D dataset when fusing all levels.

Compared to state-of-the-art methods on the NTU RGB+D dataset for cross-subject and cross-view evaluation, the results are reported in Table III. Since this dataset provides rich samples for training deep models, e.g., HBRNN-L [7], achieved higher accuracy than most of hand-crafted based methods. It verifies the effectiveness of the RNN-based methods. Besides, our method performs better than methods such as "Deep RNN" [10], "Deep LSTM" [10], Part-aware LSTM [10] for both cross-subject and cross-view protocols. And it also outperforms the ST-LSTM [20] for cross-subject evaluation and obtains competitive results for cross-view evaluation.

Compared to other methods on the SmartHome dataset, as shown in Table IV, the proposed two-stream 3D CNN model achieves the best performance, with the accuracy of 79.38%, which is better than Synthesized+Pre-trained [17]. Compared to ConvNets [25] and JTM [8], the improvements are 8.27% and 12.16% respectively. These improvements verify that our method can work well against noisy data.

Our method outperforms these methods mainly due to the following reasons. First, 3D convolutional neural network can sufficiently capture correlations , thereby learning spatial and temporal information simultaneously. Particularly, the well-designed form of spatial volume and temporal volume is useful for 3D CNN to represent information; Second, the network of two-stream enhances the spatial-temporal information and compensates for each other; Third, the multi-temporal structure learns multi-scale information including local patterns and global relationships.

Table 3: Comparisons on the NTU RGB+D dataset

Methods	Year	CS(%)	CV(%)
HOG2 [21]	2013	32.24	22.27
Lie Group [22]	2014	50.08	52.76
Skeletal Quads [23]	2014	38.62	41.36
FTP Dynamic Skeletons [24]	2015	60.23	65.22
HBRNN-L [7]	2015	59.07	63.97
Deep RNN [10]	2016	56.29	56.29
Deep LSTM [10]	2016	60.69	67.29
Part-aware LSTM [10]	2016	62.93	70.27
ST-LSTM [20]	2016	61.70	**75.50**
Multi-temporal 3D CNN(ours)	2017	**66.85**	**72.58**

Table 4: Comparisons on the SmartHome dataset

Methods	Year	Cross Subject(%)
ConvNets [25]	2015	67.22
JTM [8]	2016	71.11
SM+MM [26]	2017	77.92
Skeleton Visualization [17]	2017	78.61
Two-stream 3D CNN(ours)	2017	**79.38**

5. CONCLUSION AND FUTURE WORK

This paper proposes a novel two-stream 3D CNN model for action recognition based on skeleton sequences. The proposed spatial-temporal stream can learn more motion details of local and global by individual stream's mutual enhancement. Meanwhile, the simple yet effective 3D CNN architecture overcomes the overfitting problem. Experimental results show that our method outperforms most of state-of-the-art RNN-based approaches and verify the effectiveness of using 3D CNN learn the processed skeleton data. And the multi-temporal version do increase the ability of 3D CNN model to capture multi-scale information. In the future, in order to train 3D CNN more effectively, we will focus on different ways of encoding skeleton data.

6. REFERENCES

[1] Chen Chen, Baochang Zhang, Zhenjie Hou, Junjun Jiang, Mengyuan Liu, and Yun Yang, "Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4651–4669, 2017.

[2] Mengyuan Liu, Hong Liu, and Chen Chen, "Robust 3d action recognition through sampling local appearances and global distributions," *IEEE Transactions on Multimedia*, 2017.

[3] Mengyuan Liu, Hong Liu, and Chen Chen, "3d action recognition using multi-scale energy-based global ternary image," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[4] Mengyuan Liu and Hong Liu, "Depth context: A new descriptor for human activity recognition by using sole

depth sequences," *Neurocomputing*, vol. 175, pp. 747–758, 2016.

[5] Alex Graves, Ed., *Supervised Sequence Labelling with Recurrent Neural Networks*, 2012.

[6] Alex Graves, Abdel Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (ICASSP)*, 2013, pp. 6645–6649.

[7] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[8] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM on Multimedia Conference (ACMMM)*, 2016, pp. 102–106.

[9] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 221–231, 2013.

[10] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.

[11] Daniel Maturana and Sebastian Scherer, "Voxnet: A 3D convolutional neural network for real-time object recognition," *IEEE Intelligent Robots and Systems (IROS)*, pp. 922–928, 2015.

[12] Bo Li, "3D fully convolutional network for vehicle detection in point cloud," *arXiv preprint arXiv:1611.08069*, 2016.

[13] Agne Grinciunaite, Amogh Gudi, Emrah Tasli, and Marten den Uyl, "Human pose estimation in space and time using 3D CNN," in *European Conference on Computer Vision Workshops (ECCVW)*, 2016, pp. 32–39.

[14] Graham Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler, "Convolutional learning of spatio-temporal features," *IEEE European Conference on Computer Vision (ECCV)*, pp. 140–153, 2010.

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4489–4497.

[16] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu, "Body joint guided 3D deep convolutional descriptors for action recognition," *arXiv preprint arXiv:1704.07160*, 2017.

[17] Mengyuan Liu, Hong Liu, and Chen Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition (PR)*, pp. 346–362, 2017.

[18] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 1–7.

[19] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[20] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *IEEE European Conference on Computer Vision (ECCV)*, 2016, pp. 816–833.

[21] Eshed Ohn Bar and Mohan Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 465–470.

[22] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 588–595.

[23] Georgios Evangelidis, Gurkirt Singh, and Radu Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2014, pp. 4513–4518.

[24] Jian Fang Hu, Wei Shi Zheng, Jianhuang Lai, and Jianguo Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5344–5352.

[25] Yong Du, Yun Fu, and Liang Wang, "Skeleton based action recognition with convolutional neural network," in *Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 579–583.

[26] Mengyuan Liu, Qinqin He, and Hong Liu, "Fusing shape and motion matrices for view invariant action recognition using 3d skeletons," in *IEEE International Conference on Image Processing (ICIP)*, 2017.