LEARNING EXPLICIT SHAPE AND MOTION EVOLUTION MAPS FOR SKELETON-BASED HUMAN ACTION RECOGNITION

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

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

ABSTRACT

Human action recognition based on skeleton sequences has wide applications in human-computer interaction and intelligent surveillance. Although previous methods have successfully applied Long Short-Term Memory(LSTM) networks to model shape evolution of human actions, it still remains a problem to efficiently recognize actions, especially for similar actions from sequential data due to the lack of the details of motion. To solve this problem, this paper presents an improved LSTM-based network to jointly learn explicit long-term shape evolution maps (SEM) and motion evolution maps (MEM). Firstly, human actions are represented as compact SEM and MEM, which mutually compensate. Secondly, these maps are jointly learned by deep LSTM networks to explore high-level temporal dependencies. Then, a weighted aggregate layer (WAL) is designed to aggregate outputs of LSTM networks cross different temporal stages. Finally, deep features of shape and motion are combined by decision level fusion. Experimental results on the currently largest NTU RGB+D dataset and public SmartHome dataset verify that our method significantly outperforms the state-of-the-arts.

Index Terms**—** Human Action Recognition, Skeleton Sequences, Long Short-Term Memory, Depth Sensor

1. INTRODUCTION

Human action recognition is an important branch of computer vision due to its relevance to a wide range of applications, such as intelligent surveillance [1], human computer interaction [2] and video analysis [3]. With the advent of cost-effective and easy-operation depth sensors such as Microsoft Kinect, it has become feasible to estimate skeletons in real-time [4]. A plurality of works [5–8] have been conducted on skeleton-based action analysis. Despite significant progress, how to properly model long-term human actions for sequential data still remains a challenging problem.

Relation to prior work: Recently, deep learning based methods, especially Recurrent Neural Networks (RNN) and LSTM networks, have achieved superior performance in action recognition [9–11]. Du *et al.* [9] proposed an end-to-end hierarchical RNN to model human physical structure and temporal dynamics of the skeletal joints. Shahroudy *et al.* [12] proposed a part-aware LSTM network to enforce the model towards learning the long-term contextual representations for different body parts individually, which showed that LSTM outperforms RNN and some hand-crafted methods. Observing that previous RNN-based methods only model the contextual dependency in the temporal domain, Liu *et al.* [10] introduced a spatial-temporal LSTM to jointly learn both spatial and temporal relationships among joints. Since not all joints are informative for action analysis, Liu *et al.* [13] extended LSTM network to a Global Context-Aware Attention LSTM network, which has the capability of selectively focusing on the informative joints in each frame of the skeleton sequence.

However, the aforementioned RNN/LSTM-based methods do not consider the details of motion of skeleton sequence. Our improved LSTM-based network adopts motion evolution maps (MEM) to reduce redundant information between adjacent frames, and highlight the details of movement information of skeleton joints accordingly. Further, the fusion of MEM and shape evolution maps (SEM) are complementary to depict skeleton sequence. Moreover, since the methods mentioned above merely take the output of the last frame of deep LSTM networks as the representation of skeleton sequence, some discriminative shape and motion representations can be lost. Consequently, a weighted aggregate layer (WAL) is utilized to learn and assign adaptive weights to the output of each frame of deep LSTM networks and obtain the weighted summation.

Generally, our method contains two main contributions:

(1) An improved LSTM-based network is proposed to jointly learn explicit SEM and MEM, which mutually compensate and describe skeleton sequence more comprehensively.

(2) WAL is designed to aggregate different output frames of deep LSTM networks with adaptive weights, which speeds up convergence and boosts recognition performance.

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).

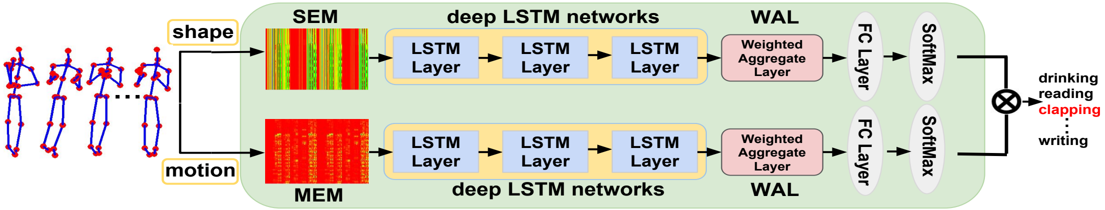

Fig. 1: Overall framework of the improved LSTM-based network, which consists of shape subnetwork and motion subnetwork.

2. THE PROPOSED METHOD

In this section, an improved LSTM-based network is illustrated in Fig.1. It consists of shape subnetwork and motion subnetwork. For each subnetwork, similar to [11, 14], we build the main LSTM network as the base model by stacking three LSTM layers called deep LSTM networks followed by one FC layer with a SoftMax classifier. Differently, WAL is added between the last LSTM layer and the FC layer to reserve more discriminative representations of input sequence. At last, deep representations of shape and motion are fused by multiplication at decision level. The remainder of this section is organized as follows: we first formally describe SEM and MEM, then introduce WAL. Finally, decision level fusion of the network is presented.

2.1. Shape and Motion Evolution Maps

The proposed SEM and MEM explicitly depict the temporal dynamics of skeletons. Assuming that a skeleton sequence has *N* frames and each frame consists of *M* joints representing the skeletons of single or multiple subjects performing actions. For single subject, the joint coordinates of the first person are copied to the second person. Let $\boldsymbol{p}^j = [x_j, y_j, z_j]$ be the 3D coordinates of the j^{th} joint in each frame, where $j \in (1, 2, ..., M)$. Then the M joints of all subjects in each frame can be represented as $\boldsymbol{f} = [\boldsymbol{p}^1, \boldsymbol{p}^2, ..., \boldsymbol{p}^M]^{\mathrm{T}}$. Note that the numbering of joints follows a fixed order to maintain the correspondence between frames.

Similar to the previous work [9], we concatenate the 3D coordinates of different joints at each time step. Then SEM matrix of all frames in a skeleton sequence is arranged in their temporal order and defined as follows:

$$\mathbf{SEM} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_N]^{\mathrm{T}}, \tag{1}$$

where each row of **SEM** represents the skeleton shape information of each frame. In order to solve the problem of the variant length of the sequence, the bilinear interpolation is applied to scale the number of rows from N to $N^{'}$:

$$\mathbf{SEM}_{\_norm} = Bilinear(\mathbf{SEM}) = [\boldsymbol{f}^{'}_1, \boldsymbol{f}^{'}_2, ..., \boldsymbol{f}^{'}_{N'}]^{\mathrm{T}}, \tag{2}$$

where $N^{'}$ denotes the fixed length of the sequence. As for MEM, the skeleton motion between time step *t* and *t*+1 is computed as:

$$\boldsymbol{q}_t = \boldsymbol{f}_{t+1} - \boldsymbol{f}_t = [\boldsymbol{f}^1_{t+1} - \boldsymbol{f}^1_t, \boldsymbol{f}^2_{t+1} - \boldsymbol{f}^2_t, ..., \boldsymbol{f}^M_{t+1} - \boldsymbol{f}^M_t], \tag{3}$$

where *t* is the frame index, $\boldsymbol{q}_t$ is the skeleton motion at time step *t*. Therefore, MEM matrix of all frames in a skeleton sequence can be represented as:

$$\mathbf{MEM} = [\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_{N-1}]^{\mathrm{T}}, \tag{4}$$

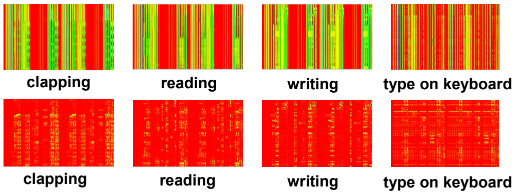

Fig. 2: Sample SEM (the upper) and MEM (the lower) generated by the proposed method on NTU RGB+D dataset.

Similarly, the bilinear interpolation is utilized to generate the fixed length $N^{'}$ of sequence:

$$\mathbf{MEM}_{\_norm} = Bilinear(\mathbf{MEM}) = [\boldsymbol{q}^{'}_1, \boldsymbol{q}^{'}_2, ..., \boldsymbol{q}^{'}_{N'}]^{\mathrm{T}}, \tag{5}$$

Fig.2 shows the compact sample SEM and MEM generated from the corresponding actions on NTU RGB+D dataset [12]. For the similar actions "clapping", "reading", "writing", their SEM are similar, capturing the shape information of each frame of action sequence. Their MEM are discriminative because they reduce redundant information, meanwhile highlight the details of motion information of skeleton joints between adjacent frames. Furthermore, MEM is relatively weak in the description of the global shape information of the skeleton sequence. Therefore, the fusion of SEM and MEM is capable of mutually compensating to represent the skeleton sequence more completely.

2.2. Weighted Aggregate Layer

RNN is a popular model for sequential data modeling and feature extraction [15]. Due to the vanishing gradient and error blowing up problems [14], the standard RNN cannot store information for long periods of time. LSTM [16] is an advanced RNN architecture which mitigates this problem. As illustrated in Fig.3(a), an LSTM neuron contains a memory cell c_t which has a self-connected recurrent edge of weight 1. At each time step *t*, the neuron can choose to write, reset, and read the memory cell governed by the input gate i_t, forget gate f_t, and output gate o_t. The output vectors of deep LSTM networks are $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{N'}]$.

For the sequence level classification, only the extracted features $\boldsymbol{h}_{N'}$ from the last frame of the output flow to the FC layer to make the final prediction. However, the ability to model long-term dependencies of LSTM is relatively weakened as the sequence length increases. Consequently, it can lose some discriminative shape and motion representations from $\boldsymbol{h}_1$ to $\boldsymbol{h}_{N'-1}$. Inspired by such insight, a brand new layer called WAL is designed to aggregate information over

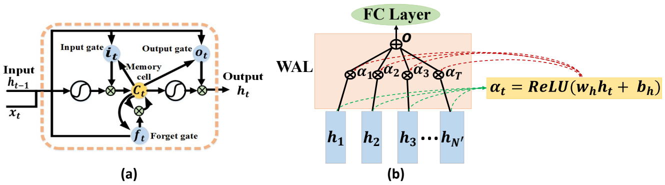

Fig. 3: (a) LSTM neuron. (b) Weighted aggregate layer mechanism.

time. Specifically, the representation $\boldsymbol{o}$ of the sequence is the weighted summation of the output vectors at all time steps:

$$\boldsymbol{o} = \sum_{t=1}^{N'} \alpha_t \cdot \boldsymbol{h}_t, \tag{6}$$

In formulation (6), WAL can be seen as a "context" vector of the input sequence by computing an adaptive weighted average of the output vectors H. The weight α_t for each output vector $\boldsymbol{h}_t$ can be calculated as:

$$\alpha_t = ReLU(\boldsymbol{w}_h \boldsymbol{h}_t + \boldsymbol{b}_h), \tag{7}$$

where $\boldsymbol{w}_h$ is the learnable parameter matrix, $\boldsymbol{b}_h$ is the bias vector. Considering that the good convergence performance of *ReLU*, we use the learned non-linear function of *ReLU* to compute a scalar importance weight for $\boldsymbol{h}_t$. Fig.3(b) illustrates the weighted aggregate layer mechanism. As the representation $\boldsymbol{o}$ of the sequence, the weighted summation is the input of FC layer.

2.3. Decision Level Fusion

For the improved LSTM-based network, SEM and MEM are directly fed into the network to model long-term temporal dynamics and generate the compact and discriminative representation for SoftMax classifier. The final predicted probability given a sequence X is obtained by multiplying the separate predicted probability from shape and motion subnetwork:

$$p(C|X, \boldsymbol{W}) = \frac{e^{\boldsymbol{s}_C}}{\sum_{j=1}^{C} e^{\boldsymbol{s}_j}} \cdot \frac{e^{\boldsymbol{m}_C}}{\sum_{j=1}^{C} e^{\boldsymbol{m}_j}}, \tag{8}$$

where $\boldsymbol{s}$, $\boldsymbol{m}$ are the scores of the sequence X for SEM and MEM over C classes respectively. $\boldsymbol{W}$ is the learnable weight matrix.

Then, the class label is calculated as $argmax(p(C|X, \boldsymbol{W}))$. The network is trained with cross-entropy loss as:

$$L = -\sum_{i=1}^{C} [\boldsymbol{y} log(P(C_i|X, \boldsymbol{W})] \tag{9}$$

where C represents the number of classes and $\boldsymbol{y}$ indicates the true class label of each action sequence.

3. EXPERIMENTS

In this section, we firstly describe the datasets and protocols, and introduce the implementation details. Then we evaluate the proposed method on public NTU RGB+D dataset [12] and SmartHome dataset [17], and finally report the experimental results and analysis.

3.1. Datasets and Protocols

NTU RGB+D dataset [12] is currently the largest skeleton-based action recognition dataset. It contains 56,880 sequences of 60 classes performed by 40 subjects from three main cameras. It is challenging due to similar actions and noisy data, as shown in Fig.4. Following the Cross-Subject (CS) protocol in [12], we split the dataset into 40,320 training samples and 16,560 testing samples. Following the Cross-View (CV) protocol in [12], the training and testing sets have 37,920 and 18,960 samples, respectively.

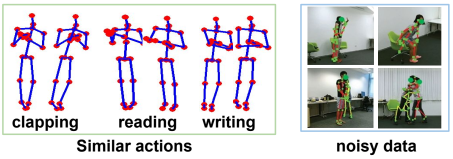

Fig. 4: Snaps from NTU RGB+D dataset [12].

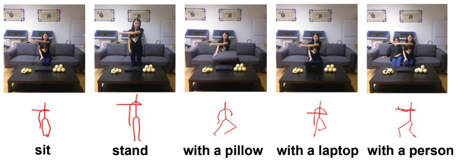

Fig. 5: Skeletons of action "wave" in SmartHome dataset [17].

SmartHome dataset [17] contains 6 types of actions performed 6 times by 9 subjects in 5 situations from single camera, resulting in 1620 sequences. Skeleton joints contain much noises, due to occlusions and the unconstrained poses of action performers. The noisy skeleton snaps of action "wave" are illustrated in Fig.5. Following the Cross-Subject (CS) protocol in [17], we use subjects #1, 3, 5, 7, 9 for training and subjects #2, 4, 6, 8 for testing.

3.2. Implementation Details

The implementation is derived from Keras[1] toolbox based on one NVIDIA GeForce GTX 1080 GPU and our codes are open source[2]. Each LSTM layer is composed of 100 LSTM neurons, and the number of neurons of the FC layer is equal to the number of action classes. The batch sizes for SmartHome dataset and NTU RGB+D dataset are 32 and 256 respectively. Adam [18] is adapted to train all the networks, and the initial learning rate is set to 0.001. Dropout [19] with a probability of 0.5 is used to alleviate overfitting. The fixed length of sequence N' for SmartHome dataset and NTU RGB+D dataset are set to 30 and 100 respectively.

3.3. Experimental Results

Evaluation of shape and motion evolution maps. Table 1 shows the recognition accuracies of learned SEM, MEM and the fusion of them(SME-MEM) on NTU RGB+D dataset and SmartHome dataset. It can be seen that MEM performs better than SEM, because MEM highlights the details of motion information of joints and emphasizes the temporal information between frames to perform LSTM model the long-range temporal information more effectively. Importantly, the recognition accuracy of **SEM-MEM** outperforms SEM and MEM by 8.24% and 4.79% respectively for CV protocol on NTU RGB+D dataset, showing complementary property of SEM and MEM. Fig.6 shows some action results. Especially for the

[1] https://github.com/fchollet/keras

[2] https://github.com/Damilytutu/SEM-MEM

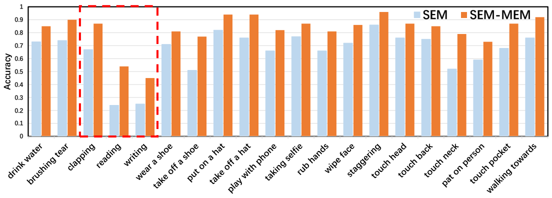

Fig. 6: Compare the accuracies of SEM and SEM-MEM for some actions on NTU RGB+D dataset (Cross-View protocol [12])

Table 1: Comparison of recognition accuracies among SEM, MEM and SEM-MEM on NTU RGB+D dataset and SmartHome dataset.

Methods	NTU RGB+D		SmartHome
	CS(%)	CV(%)	CS(%)
SEM	72.49	80.27	72.00
MEM	75.98	83.72	73.91
SEM-MEM	**81.53**	**88.51**	**80.32**

Table 2: Comparison of recognition accuracies using WAL on NTU RGB+D dataset and SmartHome dataset.

Methods	NTU RGB+D		SmartHome
	CS(%)	CV(%)	CS(%)
SEM+WAL	74.67	82.71	73.50
MEM+WAL	77.52	85.55	75.16
SEM-MEM+WAL	**82.86**	**89.94**	**81.83**

similar actions "clapping", "reading", "writing"(with red frame), the accuracy of **SEM-MEM** is morn than 20% higher than that of SEM. **Evaluation of weighted aggregate layer(WAL).** Table 2 shows the effect of WAL on NTU RGB+D dataset and SmartHome dataset. By referring to Table 1, SEM+WAL and MEM+WAL outperforms SEM and MEM by 2.44% and 1.83% respectively for CV protocol on NTU RGB+D dataset. Fig.7(a) illustrates the visualization of the adaptive weights for action "clapping". Since the deep LSTM networks usually accumulate more information over time, the learned weights usually increase correspondingly. The weights reflect the importance of shape and motion representation at each time step respectively. Furthermore, as shown in Fig.7(b), the training loss of **SEM-MEM+WAL** drops and converges faster than SME-MEM. These results verify that the proposed WAL contributes to speed up convergence and boost recognition performance.

Comparison with the state-of-the-arts. Table 3 shows the recognition accuracies of the state-of-the-art methods and our method on NTU RGB+D dataset. Since this dataset provides rich samples for training, deep models such as HBRNN-L [9] and Part-aware LSTM [12] achieve higher accuracy than most of hand-crafted methods such as Lie Group [20] and Dynamic Skeletons [21]. Our method surpasses some state-of-the-art LSTM-based methods by a notable margin, such as ST-LSTM+Trust Gate [10], Geometric Features [22] and GCA-LSTM [13]. Besides, compared with multi-CNN models [23, 24], our method outperforms Clips+CNN+MTLN [24] by 3.29% and 5.11% under CS and CV protocols respectively. Our proposed **SEM-MEM+WAL** achieves the highest accuracy of 82.86% and 89.94% under CS and CV protocols respectively. These results indicate the discriminative power of our method to jointly learn SEM and MEM. Table 4 shows the recognition accuracies of the state-of-the-art methods and our method on SmartHome dataset. Our proposed **SEM-MEM+WAL** achieves the highest accuracy of 81.83%, which outperforms Skeleton Visualization [23] by 3.22%. This improvement not only indicates the proposed method can work well against noisy data, but also proves the effectiveness and robustness of the proposed method.

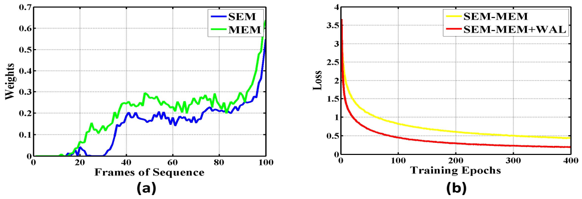

Fig. 7: (a) Visualization of the adaptive weights of SEM and MEM for action "clapping". (b) The loss curves of SEM-MEM and SEM-MEM+WAL on NTU RGB+D dataset. **(Best viewed in color)**

Table 3: Comparison with the state-of-the-art methods on NTU RGB+D dataset (Cross-Subject and Cross-View protocols [12]).

Methods	**Year**	**CS(%)**	**CV(%)**
Lie Group [20]	2014	50.08	52.76
Dynamic Skeletons [21]	2015	60.23	65.22
HBRNN-L [9]	2015	59.07	63.97
Part-aware LSTM [12]	2016	62.93	70.27
ST-LSTM+Trust Gate [10]	2016	69.20	75.70
Geomeric Features [22]	2017	70.26	82.39
GCA-LSTM [13]	2017	74.40	82.80
Skeleton Visualization [23]	2017	75.97	82.56
Clips+CNN+MTLN [24]	2017	79.57	84.83
SEM-MEM+WAL(Ours)	2017	**82.86**	**89.94**

Table 4: Comparison with the state-of-the-art methods on SmartHome dataset (Cross-Subject protocol [17]).

Methods	**Year**	**CS(%)**
ConvNets [25]	2015	67.22
JTM [26]	2016	71.11
SM+MM [17]	2017	77.92
Skeleton Visualization [23]	2017	78.61
SEM-MEM+WAL(Ours)	2017	**81.83**

4. CONCLUSIONS

This paper presents an improved LSTM-based network that jointly learns explicit SEM and MEM for skeleton-based human action recognition. Since MEM highlights the details of motion information and the fusion of SEM and MEM depicts the skeleton sequence more completely, the proposed network efficiently recognizes human actions especially similar actions from sequential data. Besides, the proposed WAL speeds up the convergence and boosts the performance by reserving more discriminative representations of input sequence. Experimental results on the currently largest NTU RGB+D dataset and public SmartHome dataset show that our method outperforms the state-of-the-art methods. Future works will focus on verifying the effectiveness of our method in wider applications.

5. REFERENCES

[1] W. Lin, M. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE TCSVT*, vol. 18, no. 8, pp. 1128–1139, 2008.

[2] H. Liu, Q. He, and M. Liu, "Human action recognition using adaptive hierarchical depth motion maps and gabor filter," *in Proc. ICASSP*, pp. 1847–1851, 2017.

[3] M. Liu and H. Liu, "Depth context: A new descriptor for human activity recognition by using sole depth sequences," *Neurocomputing*, vol. 175, pp. 747–758, 2016.

[4] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *in Proc. CVPR*, vol. 56, no. 1, pp. 116–124, 2011.

[5] J. Aggarwal and X. Lu, "Human activity recognition from 3D data: A review," *PRL*, vol. 48, pp. 70–80, 2014.

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

[7] L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *PR*, vol. 53, pp. 130–147, 2016.

[8] J. Zhang, W. Li, P. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," *PR*, vol. 60, pp. 86–105, 2016.

[9] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," *in Proc. CVPR*, pp. 1110–1118, 2015.

[10] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," *in Proc. ECCV*, pp. 816–833, 2016.

[11] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," *in Proc. AAAI*, vol. 2, pp. 3697–3703, 2016.

[12] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," *in Proc. CVPR*, pp. 1010–1019, 2016.

[13] J. Liu, G. Wang, P. Hu, L. Duan, and A. Kot, "Global context-aware attention LSTM networks for 3D action recognition," *in Proc. CVPR*, vol. 7, pp. 1647–1656, 2017.

[14] 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*, pp. 4263–4270, 2017.

[15] A. Graves, "Supervised sequence labelling with recurrent neural networks," *Springer*, vol. 385, 2012.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *MIT Press Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] M. Liu, Q. He, and H. Liu, "Fusing shape and motion matrices for view invariant action recognition using 3D skeletons," *in Proc. ICIP*, 2017.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, 2014.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," *in Proc. CVPR*, pp. 588–595, 2014.

[21] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *in Proc. CVPR*, pp. 5344–5352, 2015.

[22] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," *in Proc. WACV*, pp. 148–157, 2017.

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

[24] Q.H. Ke, M. Bennamoun, S.J. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," *in Proc. CVPR*, 2017.

[25] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," *in Proc. ACPR*, pp. 579–583, 2015.

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