Spatio-Temporal and View Attention Deep Network for Skeleton based View-invariant Human Action Recognition

Yan Feng[1], Ge Li[2], Chunfeng Yuan[3]
{fywmh@163.com[1], lige420@126.com[2], cfyuan@nlpr.ia.ac.cn[3]}

School of Information Science & Technology, Qingdao University of Science & Technology, Qingdao 266061[1,2], National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190[3]

Abstract. In this paper, we propose a spatio-temporal and view attention based deep network model to avoid the disturbance of the view and noise in skeleton data for human action recognition. Our model consists of two sub-networks which are built on the Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM). The view-specific sub-network incorporating spatio-temporal attention learns discriminative features from single input view by paying more attention to key joints and frames. The following view attention sub-network obtains common view-invariant representations shared among views and it contains a view attention module to select the discriminative views. Finally, we propose a regularized cross-entropy loss to ensure the effective end-to-end training of the network. Experimental results demonstrate the effectiveness of the proposed model on the current largest NTU action recognition dataset.

Keywords: Action recognition, skeleton, view-invariant, attention model.

1. Introduction

Recently, with the development of the depth sensors (such as Microsoft Kinect [1], Asus Xtion PRO LIVE [2] and Intel RealSence [3]), more and more attention has been paid to the research of the skeleton based action recognition.

View-invariant action recognition is a major challenge in action recognition because view variations result in the complex changes of skeletal data. On one hand, in practical scenario, the different views of the same action captured by different cameras at the same time are different, such as the front view and the side view. On the other hand, the different setups of the cameras to different subjects lead to different skeletal views, such as the different heights and distances between the cameras and the subjects. Moreover, actors may dynamically change their motion directions. As illustrated in Fig. 1, the skeletons of the same posture are quite different when the posture is captured from different views.

There are many approaches [4-14, 26] contributing to the view-invariant research in skeleton action recognition. However, most of these methods mine the discriminant information only from signal view or utilizes the traditional methods which are too complex and redundant to effectively train the network. In this paper, we propose a new deep network consisting of view-specific sub-network and view attention sub-network. Our deep network first learns the view-specific discriminant information and then learns the view-invariant

MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276579

information. Moreover, Our deep network incorporates the spatiao-temporal and view attention mechanism to focus on key joints, frames and views for the accuracy improvement.

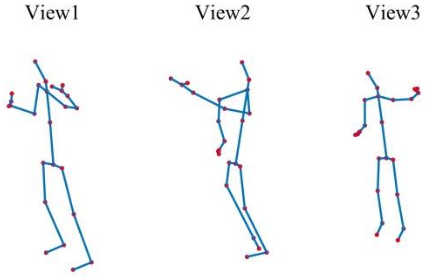

Fig. 1. Visualization of the skeleton data from different views captured by different cameras about one action ‘taking a selfie’.

Specifically, the view-specific sub-network in our deep network aims to extract the view discriminantive information. Its input is the combination of the 3D joints of each frame. But, it is not the fact that every joint should have the same effect on action recognition either in same frame or between different frames. Therefore, we propose a spatial attention module and a temporal attention module incorporated into our deep network to pay more attentions to discriminative joints and frames.

The view attention sub-network in our deep network extracts the potential representations shared by all views. From multiple views, the skeletal sequences also have different effects on the action recognition. At the same time, the side views always generate inaccurate skeletal information because of the serious self-occlusion and noise. Therefore, we propose a view attention module based view attention sub-network, which automatically assigns different view attention weights to different view features and then focuses on the key view sequence.

2. Related work

2.1. Skeleton action recognition based on LSTM

View-invariant skeleton action recognition plays an important part in skeleton-baded human action recognition. The approaches [4-14] about the view-invariant skeleton action recognition can be categoried into three classes. (1) Some researchers transform the joint information to obtain view-invariant descriptors. Evangelidis et al. [4] propose a local skeleton descriptor that encodes the relative positions of joint quadruples to a compact (6D) view-invariant skeletal feature, referred to as skeletal quad. (2) Some researchers transfer the actions to a view-invariant high-level space. For example, Rahmani et al. [14] propose a large corpus of multi-view training data by fitting synthetic 3D human models to real motion capturing data and rendering the human poses from numerous viewpoints. (3) Methods based on Long Short-Term Memory(LSTM) network [21] transform the viewpoints to another viewpoint. Zhang et al. [13] introduce a view adaptive recurrent neural network that enables the network itself to adapt to the most suitable observation viewpoints from end to end.

In contrast to the above works, we propose a LSTM based view-invariant deep network with two attention sub-networks to extract the effective skeleton features from different views.

2.2. Attention model

Our method is also related with the attention model [9,12,22,23,25] which attempts to mine more discriminative information. Song et al. [9] propose an end-to-end spatial and temporal attention model for human action recognition from skeleton data. This method learns to focus on discriminative joints of skeleton within each frame of input and pays attention of different levels to outputs of different frames.

The input of our multi-view model are multiple skeleton sequences in different views for the same action. View variation cannot be ignored. To address this, we extend our view-specific sub-network and view attention sub-network by adding spatio-temporal and view attention modules to effectively utilize the multi-view information for action recognition. To the best of our knowledge, we are the first to incorporate the attention mechanism into view-invariant action recognition from skeleton data.

3. Spatio-temporal and view Attention Deep Network

For skeleton-based action recognition, the main difficulty lying in that skeleton data is that action recognition is sensitive to views and contains noise. We propose a spatio-temporal and view attention deep network to avoid the disturbance of the view and the noise in skeleton data for human action recognition.

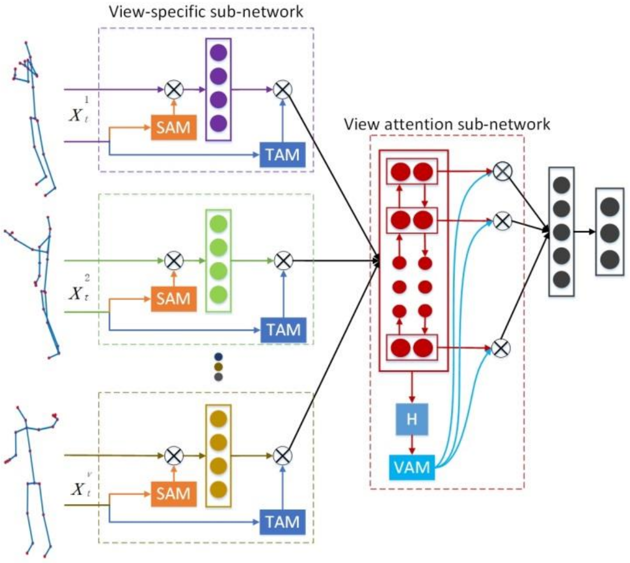

Fig. 2. Overall architecture of our spatio-temporal and view attention deep network, which consists of several view-specific sub-networks in the first layer (the different views denoted by purple, green and gold colors in the picture) and the view attention sub-network shared in the second layer (red color). The spatial attention module (SAM) is denoted in orange color; Temporal attention module (TAM) is denoted in blue color; View attention module (VAM) is denoted in wathet blue color.

The architecture of proposed network is shown in Figure 2. Firstly, the skeleton sequences from multiple views are the input of the view-specific sub-network. Secondly, the output of the view-specific sub-network is the input of the view attention sub-network. Finally, the information from view attention sub-network is sent to fully-connected layer and SoftMax layer for action classification.

In this section, we first introduce the view-specific sub-network which contains a spatial attention module, stacked LSTMs and a temporal attention module. We then describe the view attention sub-network which is constructed by the Bi-LSTM and a view attention module.

3.1. View-specific sub-network

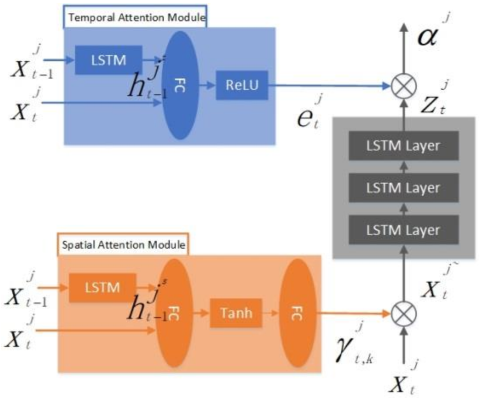

Fig. 3. Overall architecture of the view-specific sub-network, which consists of a spatial attention module (orange), a temporal attention module (blue) and a stacked LSTMs (grey).

We propose a view-specific sub-network based on LSTM to capture the discriminative features. Fig. 3 shows the architecture of the proposed view-specific sub-network, which consists of v stacked LSTMs and corresponding spatiao-temporal attention modules. v is the number of view sequences from the same action. The stacked LSTMs has multiple LSTM layers. The spatial as well as temporal attention module is composed of a LSTM layer, a fully connected layer, and an activation layer.

3.1.1. The Spatial Attention module

The Spatial Attention Module (SAM) pays different spatial attention weights to different joints in order to select key joints within each frame. LSTM is the basic structure in spatial attention module, and it outputs the joint selective gate which can adaptively focus on the discriminative joints. The specific structure of SAM is shown in Fig. 3.

Specially, given v views, a sample of j^{th} view is denoted as x^j. At each time step t, the full set of K joints is defined as $x_t^j = [x_{t,1}^j, x_{t,2}^j, \dots, x_{t,K}^j]$, with $x_{t,k}^j \in \mathbb{R}^3$. The K is the number of joints in each frame. The output scores s_t^j of the SAM indicates the importance of K joints:

$$s_t^j = W_{es}\tanh\left(W_{xs}x_t^j + W_{hs}h_{t-1}^{js} + b_s\right) + b_{es} \quad , \qquad (1)$$

where W_{es}, W_{xs}, W_{hs} are the learnable parameter matrixes, b_s, b_{es} are the bias vectors, and h_{t-1}^{js} is the hidden state from an LSTM layer in the spatial channel. For the k^{th} joint, we use the activation as their joint selective gate:

$$\gamma_{t,k}^{j} = \frac{\exp\left(s_{t,k}^{j}\right)}{\sum_{\Upsilon=1}^{K}\exp\left(s_{t,\Upsilon}^{j}\right)} \quad , \tag{2}$$

We define the activation $\gamma_{t,k}^{j}$ as the attention weight corresponding to the k^{th} joint in the t^{th} frame under the j^{th} view. By the SAM, the input of our stacked LSTMs is changed to:

$$x_t^{j\sim} = \left[x_{t,1}^{j\sim}, \dots, x_{t,K}^{j\sim}\right] \quad , \tag{3}$$

where $x_{t,k}^{j\sim} = \gamma_{t,k}^{j} \cdot x_{t,k}^{j}$.

3.1.2. The Stacked LSTMs

Our stacked LSTMs consists of multiple LSTM layers for feature extraction and temporal modeling. There are v stacked LSTM layers for v views. They are used to learn discriminative features for each view. The input of the stacked LSTMs is $x_t^{j\sim}$ in (3). And their output is denoted as z_t^{j} as illustrated in Fig. 3.

3.1.3. The Temporal Attention module

The motivation behind the Temporal Attention Module (TAM) is that different frames in the video have different importance contributing to action recognition. The TAM automatically assigns different attention weights to different frames. The temporal attention module utilizes LSTM as the basic unit. The whole architecture of TAM is illustrated in Fig. 3. Specially, the input x_t^{j} of TAM is same with SAM. The temporal attention weight at each time step is learnt as follows:

$$e_t^{j} = \mathrm{ReLU}\left(W_{e1}x_t^{j} + W_{e2}h_{t-1}^{jt} + b_e\right) \quad , \tag{4}$$

where W_{e1} and W_{e2} are learnable parameters, b_e is bias, and h_{t-1}^{jt} is the hidden state in an LSTM layer on temporal channel as illustrated in Fig. 3.

Finally, the output of the view-specific sub-network is modulated as $\alpha^{\mathrm{j}} = e_t^{j} \cdot \mathrm{z}_t^{j}$.

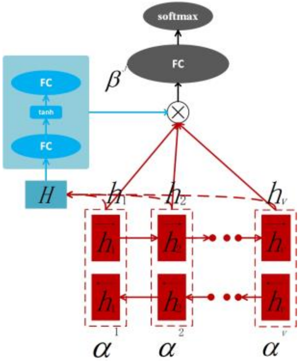

Fig. 4. The architecture of our proposed view attention sub-network, which consists of a Bi-LSTM and a view attention module.

3.2. View attention sub-network

The view attention sub-network consists of a Bi-LSTM and a view attention module, as shown in Fig.4. The Bi-LSTM learns the contextual information about multi-view features provided by multiple view-specific sub-networks. At the same time, the view attention module adaptively assigns different weights to aggregate these multi-view features.

3.2.1. The Bi-LSTM module

The output of v view-specific sub-networks is used as the input of the view attention sub-network. The Bi-LSTM employs a bidirectional LSTM layer, which captures a aggregated representation from v view-specific sub-networks to address the view independency problem. Given a view sequence, we concatenate v outputs of view-specific sub-networks as follows:

$$z = (\alpha^1, \alpha^1, \dots, \alpha^v)^T \quad . \qquad (5)$$

Each entry in the sequence z is independent with each other. We use the sequence z as the input of the Bi-LSTM to gain the dependency between v views:

$$\overrightarrow{h_j} = \overrightarrow{\text{LSTM}}\left(w_j, \overrightarrow{h_{j-1}}\right) \ , \qquad (6)$$

$$\overleftarrow{h_j} = \overleftarrow{\text{LSTM}}\left(w_j, \overleftarrow{h_{j-1}}\right) \ , \qquad (7)$$

Then we concatenate each $\overrightarrow{h_j}$ with $\overleftarrow{h_j}$ to obtain a hidden state h_j. For simplicity, we denote the v hidden states as H:

$$\text{H}=(h_1,h_2,\dots h_v) \quad . \qquad (8)$$

3.2.2. The View Attention module

Our View Attention Module (VAM) is based on the self-attention mechanism that provides a set of weight vectors for the hidden states from the Bi-LSTM. The view attention module (VAM) is a non-linear combination which consists of two fully-connected layers and a activation layer as shown in Fig.4. The hidden states H=$(h_1,h_2,\dots h_v)$ is used as the input sequence:

$$\beta^j = \text{softmax}(w_{v2}\tanh(w_{v1}h_j)) \ . \qquad (9)$$

For the view sequence classification, we multiply the view attention weight $\beta^j$with each entry h_j in view representation set H. We compute the weighted sum by multiply the annotations β^j and LSTM hidden states h_j:

$$\text{o} = \sum_{j=1}^{v} \beta^j h_j \quad . \qquad (10)$$

Finally, the output is obtained by a softmax classifier.

4. The Regularized Objective Function

In the proposed network, the attention modules learn different attention weights assigned to different joints, frames and views, the stacked LSTMs extract the discriminative features at each view and the Bi-LSTM captures view-invariant representation. Therefore, the whole network consists of interaction between attention weights, input data and hidden output data. Our network is an end-to-end framework, and we jointly train the stacked LSTMs, Bi-LSTM and the attention modules.

We utilize a regularized cross-entropy loss function to jointly train our model:

$$\mathrm{L} = -\sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda_1 \sum_{k=1}^{K} \left(1 - \frac{\sum_{j=1}^{\mathrm{v}} \sum_{t=1}^{T} \gamma_{t,k}^{j}}{vT}\right)^2 + \frac{\lambda_2}{Tv} \sum_{j=1}^{v} \sum_{t=1}^{T} \left\| e_t^j \right\|_2 + \lambda_3 \| W_{sv} \|_1 \quad , \qquad (11)$$

where $y = (y_1, y_2, \dots, y_C)^T$ denotes the label, $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C)^T$ denotes the predicted result, and $\lambda_1, \lambda_2, \lambda_3$ are the parameters which balance the three regularization terms. We use the Stochastic Gradient Descent (SGD) algorithm to minimize the loss function.

The first term in the loss function represents the cross-entropy loss term. The first regularization term is designed to make the spatial attention module to dynamically focuses on the key joints in each frame within each view. The second regularization term is designed to express the temporal domain attention module that dynamically focus on key frames. Finally, the last regularization term with the l_1 norm is to reduce overfitting of the whole network. W_{sv} denotes the connection matrix of the network.

5. Experiment

We test our network on the NTU RGB+D dataset [11], which is currently the largest action recognition dataset. It consists of 56,880 action samples of 60 classes, containing 4 different modalities of data for each sample: RGB videos, depth map sequences, 3D skeletal data and infrared videos. The 3D skeletal data contains three dimensional locations of 25 major body joints, at each frame. We use the 3D skeletal data to evaluate our algorithm with X-subject evaluation protocol in which 20 subjects are used for training, and the remaining 20 subjects are for test.

For parameter settings, our experiments are preformed based on the Keras framework [25], and the TensorFlow[26] backend. The LSTM layers in attention modules and the bi-LSTM layer are all composed of 128 LSTM neurons, and the stacked LSTM layers compose 256 LSTM neurons. We set $\lambda_1, \lambda_2 and\ \lambda_3$ to 0.01, 0.001 and 0.0005 for the NTU dataset. We set the learning rate, decay rate and momentum to 0.001, 1e-6 and 0.9, respectively. The dropout probability in our network is 0.2. The batch size is 256.

5.1. Evaluation of our network

To evaluate the effectiveness of our network, we compare three networks with the following three different configurations:

- 'Deep-LSTM Network': This network is constructed by ablating three attention modules from our network. It contains three stacked LSTM layers that capture the discriminative features under each view and the Bi- LSTM layer that captures the common features shared multiple views.
- 'Deep-LSTM Network ⊕ ST-AM': This network is the network which only ablates the view attention module from our network. This network captures the discriminative features with the spatiao-temporal attention modules, then learns the common feature shared multiple views.
- 'Deep-LSTM Network ⊕ ST-AM ⊕ VAM': This network is our proposed network.

The results of the three different networks on the NTU dataset for the X-subject setting are listed in Table 1. At the same time, we also present the results of other methods in Table 1. We can see that our proposed 'Deep-LSTM Network ⊕ ST-AM ⊕ VAM' outperforms the other skeleton-based methods by about 5% accuracy. The result of 'Deep-LSTM Network ⊕

ST-AM ⊕ VAM' outperforms 'Deep-LSTM Network ⊕ ST-AM' by 3.4%, which indicates the effectiveness of view attention module. It outperforms our deep network with attention modules by 6.1%, which demonstrates the effectiveness of our spatiao-temporal and view attention modules. To the best of our knowledge, our method achieves the best result for skeleton-based action recognition.

Table 1. Results(accuracies) on the NTU RGB+D dataset.

Method	X-subject
Skeletal Quads [4]	38.6%
Lie Group[5]	50.1%
Dynamic Skeletons [6]	60.2%
HBRNN [7]	59.1%
Part-aware LSTM [8]	62.9%
ST-LSTM [11]	69.2%
STA-LSTM [9]	73.4%
GCA-SLTM [12]	74.4%
'Deep Network'	71.6%
'Deep Network ⊕ ST-AM'	76.3%
'Deep Network ⊕ ST-AM ⊕ VAM'	79.7%

Table 2. Accuracy comparison for different LSTM layer numbers on the NTU RGB+D dataset.

# Layer	Accuracy
1	65.8%
2	69.7%
3	79.7%
4	71.4%

Moreover, we test the effect of the number of LSTM layers for recognition accuracies by testing our network with different numbers of stacked LSTM layers in the view-specific subnetworks. The results of our 'Deep-LSTM Network ⊕ ST-AM ⊕ VAM' with different numbers of stacked LSTM layers on the NTU RGB+D dataset are listed in Table 2. We can observe that when the number of layers increases in some extent, the accuracy increases. However, if the number of layers exceeds the range, the experimental accuracy drops again. In our other experiment, we choose the 3 LSTM layers.

5.2 Visualization and Discussion

To better understand our network, we visualize and analyze the learnt view and spatio-temporal attention weights. Fig. 5 shows the skeletons from different views of same posture.

At each frame, the size of the red circles denotes the spatial attention weights. We can see that different joint has different attention weights in each frame and same joint in different frames has different attention weights.

At each sequence, the blue values of edges in Fig. 5 are used to visualize of the temporal attention weights. Different frames have different temporal weights.

At each view, we define the gray value of background as the view attention weights. In the Fig. 5, we can see that the view 3 has the largest view attention weight.

From the abundant observations, we find that the attention modules can learn key joints, frames and views.

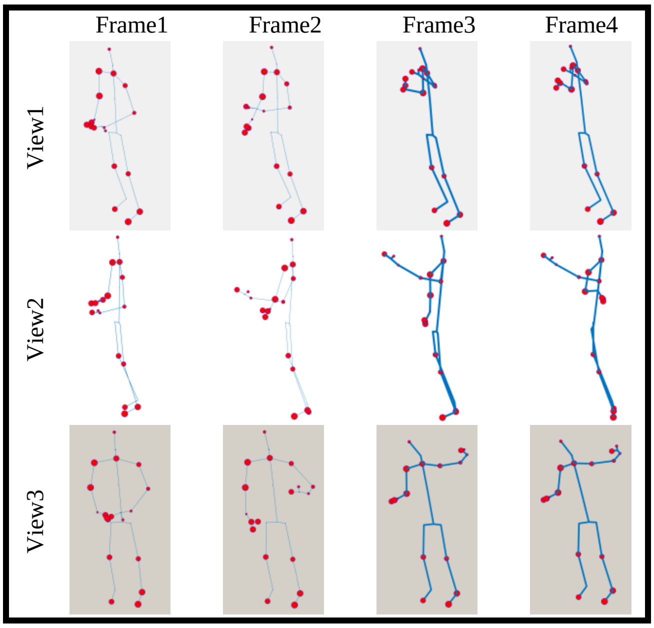

Fig. 5. Visualization of Attention modules.

6. Conclusion

In this paper, we have proposed a deep network with attention modules for human action recognition from skeleton data. The view-specific sub-network in our network has captured the discriminative features by the stacked LSTMs and adaptively focusd on key joints and key frames by spatiao-temporal attention modules. The view attention sub-network has learnt the view dependency common features shared among all views by Bi-LSTM and selectively focusd on discriminative view by view attention module. The experiments have proven the effectiveness of our attention network compared with other state-of-the-art methods.

Acknowledgments. This work is supported by the National Natural Science Foundation of China(61472196, 61672305, 11771188) and the Natural Science Foundation of Shandong Province in China (ZR2015FM012).

References

[1] Z. Zhang. Microsoft kinect sensor and its effect. IEEE MultiMedia, 19(2):4–10, 2012.
[2] ASUS Xtion PRO LIVE, https://www.asus.com/3D-Sensor/ Xtion_PRO/ (2011).
[3] Intel RealSense. https://software.intel.com/en-us/realsense.
[4] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads: Human action recognition using joint quadruples. In ICPR, 2014.
[5] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In CVPR, 2014.

[6] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In CVPR, 2015.

[7] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In CVPR, 2015.

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In CVPR, 2016.

[9] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to- end spatio-temporal attention model for human action recog- nition from skeleton data. In AAAI Conference on Artificial Intelligence, 2017. 2, 3, 5, 6, 7

[10] M. Kan, S. Shan, X. Chen. Multi-view Deep Network for Cross-View Classification[C]// Computer Vision and Pattern Recognition. IEEE, 2016:4847-4855.

[11] J. Liu, A. Shahroudy, D. Xu, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 816-833.

[12] J. Liu, G. Wang, P. Hu, et al. Global context-aware attention lstm networks for 3d action recognition[C]//Proc. Comput. Vis. Pattern Recognit. 2017: 1647-1656.

[13] P. Zhang, C. Lan, J. Xing, et al. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data[J]. 2017.

[14] H. Rahmani, A. Mian. 3D Action Recognition from Novel Viewpoints[C]// Computer Vision and Pattern Recognition. IEEE, 2016:1506-1515.

[15] D. Gavrila and L. Davis. 3D model-based tracking of humans in action: a multi-view approach. In CVPR, 1996.

[16] T. Syeda-Mahmood, A. Vasilescu, and S. Sethi. Action recognition from arbitrary views using 3D exemplars. In ICCV, 2007.

[17] A. Yilmaz and M. Shah. Action sketch: a novel action representation. In CVPR, 2005.

[18] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham. 3D pose from motion for cross-view action recognition via non- linear circulant temporal encoding. In CVPR, 2014. 1, 2, 4,

[19] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In CVPR, 2015. 1, 2, 3, 7, 8

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In International Conference on Machine Learning (ICML), 2013.

[21] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 1997.

[22] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[23] S. Sharma, R. Kiros, R. Salakhutdinov. Action recognition using visual attention[J]. arXiv preprint arXiv:1511.04119, 2015.

[24] Franc ̧ois Chollet. 2015. Keras. https://github. com/fchollet/keras.

[25] W. Zhu, C. Lan, J. Xing, W. Zeng, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In AAAI Conference on Artificial Intelligence, 2016.

[26] H. Yan, X. Jiang, T. Sun, et al. 3D human action recognition based on the spatial-temporal moving skeleton descriptor[C]// Mutimedia and Expo (ICME), 2017.