Global and Local Spatial-Attention Network for Isolated Gesture Recognition

Qi Yuan[1], Jun Wan[2(✉)], Chi Lin[3], Yunan Li[4], Qiguang Miao[4], Stan Z. Li[2], Lihua Wang[1], and Yunxiang Lu[1]

[1] School of Software, Beihang University, Beijing, China
{qiyuan,wanglihua}@buaa.edu.cn, yxlu_2000@163.com
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jun.wan,szli}@nlpr.ia.ac.cn
[3] University of Southern California, Los Angeles, CA, USA
linchi@usc.edu
[4] School of Computer Science and Technology, Xidian University & Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, China
yn_li@stu.xidian.edu.cn, qgmiao@mail.xidian.edu.cn

Abstract. In this paper, we focus on isolated gesture recognition from RGB-D videos. Our main idea is to design an algorithm that can extract global and local information from multi-modality inputs. To this end, we propose a novel attention-based method with 3D convolutional neural network (CNN) to recognize isolated gesture recognition. It includes two parts. The first one is a global and local spatial-attention network (GLSANet), which takes into account the global information that focuses on the context of the frame and the local information that focuses on the hand/arm actions of the person, to extract efficient features from multi-modality inputs simultaneously. The second part is an adaptive model fusion strategy to fuse the predicted probabilities from multi-modality inputs. Experiments demonstrate that the proposed method has achieved state-of-the-art performance on the IsoGD dataset.

Keywords: Gesture recognition · Fusion strategy · RGB-D video

1 Introduction

Video based dynamic gesture recognition plays an important role in human-computer interaction (HCI) [1]. Isolated gesture recognition and continuous gesture recognition are two major tasks [2]. The former focuses on gesture classification merely while the latter also pays attention to temporal segmentation that needs to separate each gesture from a video containing continuous gestures.

In this paper, we focus on isolated gesture recognition. In the task of isolated gesture recognition, most of the deep learning based methods [3–7] are adapted from general action recognition. However, the general action recognition task

Z. Sun et al. (Eds.): CCBR 2019, LNCS 11818, pp. 84–93, 2019.
https://doi.org/10.1007/978-3-030-31456-9_10

that based on widely used action recognition datasets, such as HMDB51 [8], UCF101 [9], and Kinetics [10] focuses more on the general human activity in the videos. But gesture recognition is a fine-grained action recognition task that focuses more on detailed hand gestures and arm movements, it is hard to extract distinguishable features to classify different gestures from the entire frame in spatial with general action recognition derived methods.

Therefore, we propose a global and local spatial-attention network, dubbed as GLSANet, which considered the global information that focuses on the global context and local information that focuses on the hand actions from multi-modality inputs simultaneously, to classify the isolated gestures on publicly used large-scale gesture recognition dataset IsoGD [2]. Besides, we propose an adaptive fusion strategy to fuse the probabilities of multi-modality inputs. The results show that we achieve state-of-the-art performance. The main contributions of this work can be summarized as follows:

- We propose an attention based network that not only embedded the global information into 3D CNNs along with the original RGB-D videos, but also focused on the local hand/arm regions based on the skeleton points.
- We develop a class-constrained fusion strategy to fuse the predicted probabilities of all the global/local attention models from multi-modality inputs.
- We achieve state-of-the-art performance on the IsoGD dataset.

2 Related Work

Most of deep learning based gesture recognition methods are adapted from action recognition tasks. Generally, action recognition models can be roughly divided into two categories. One is 2D CNN based methods that extract spatial features from several video frames followed by a temporal reasoning scheme to extract temporal features from these spatial features, and the other one is 3D CNN based methods that treat the spatial dimension the same as the spatial dimension and extract spatio-temporal features uniformly with 3D convolutional kernels.

2D CNN Based Methods. Wang *et al.* [11] proposed three representations of depth sequences, referred to respectively as Dynamic Depth Images (DDI), Dynamic Depth Normal Images (DDNI) and Dynamic Depth Motion Normal Images (DDMNI) that are constructed from a sequence of depth maps using bidirectional rank pooling [12] to capture the spatio-temporal information. 2S-RNN [13] used an LSTM layer to fuse the color and depth features extracted by a 2D CNN separately. Kopuklu *et al.* [14] proposed a data level fusion strategy to fuse optical flow information into static images as better representatives of spatio-temporal states of action.

3D CNN Based Methods. C3D [15] was first proposed to extract spatio-temporal features with a single model in action recognition, C3D treat the spatial dimension the same as the spatial dimension by using 3D convolution kernel

in the network. Li *et al.* [7] used C3D to extract features from multi-modality inputs. Miao *et al.* [3] proposed a multi-modality gesture recognition method based on the ResC3D network which leverages the advantages of both residual network [16] and C3D model. Zhu *et al.* [17] presented a pyramidal 3D convolutional network framework for gesture recognition, in which the author used a pyramid input scheme to extract multi-scale contextual information and a pyramid fusion scheme to fuse the features from pyramid input.

3 The Proposed Method

In this section, we describe the proposed GLSANet algorithm to handle the isolated gesture recognition problem. The overall structure of our framework is illustrated in Fig. 1.

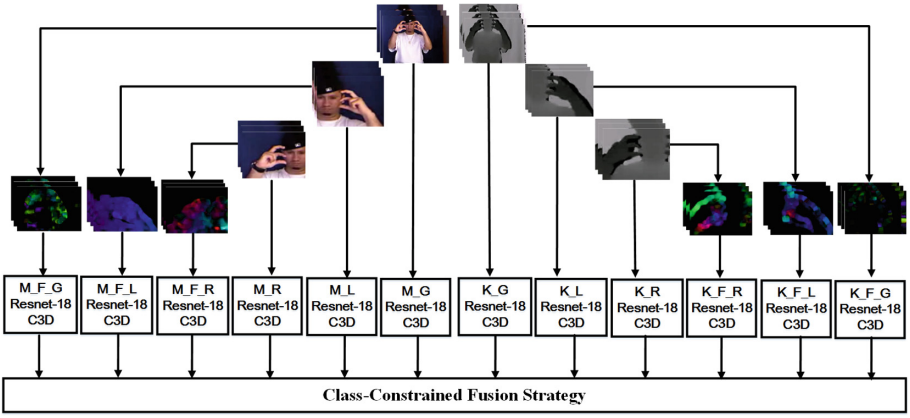
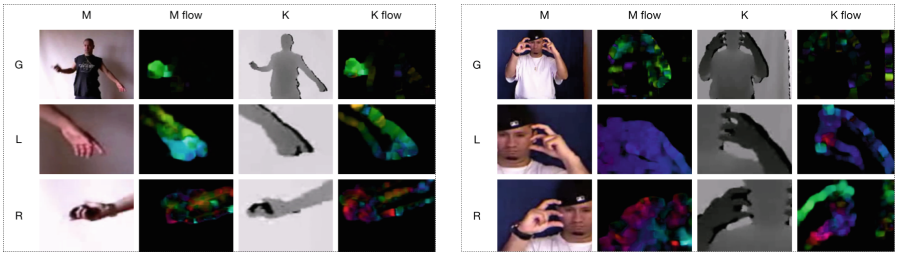


Fig. 1. The overview of our GLSANet network, which includes 12 sub-branches. A weight fusion layer is designed to merge the predictions of all branches to get the final results (M: RGB, K: depth, L: the left hand region, R: the right hand region, F: optical flow, G: global information (the whole image)).

3.1 Global and Local Spatial Attention Network

Intuitively, the motions of body parts (*e.g.*, arms and hands) are important to gesture recognition. Thus, as shown in Fig. 1, we design the GLSANet to focus on the local context with attention mechanism, especially for the left and right hands, together with the global context provided by the entire video for better extracting the essential features of the gesture.

Considering the outstanding performance of C3D and ResNet models in the recognition task [18], we use them to extract features for both original global or local videos. The C3D model with a 3D convolutional and pooling structure can describe a video concurrently from both spatial and temporal domains, while the



(a) Gesture 106 (b) Gesture 162

Fig. 2. Two samples of the training data in the IsoGD dataset. M, M flow, K, K flow are correspond to RGB, RGB flow, depth and depth flow data respectively; G, L, R are correspond to a frame of global, left and right hand attention respectively

ResNet has a better convergence performance with the deep network. Integrating them can help to achieve a good result for video based gesture recognition.

We can obtain the gesture and body keypoints via CPM [19] algorithm. Therefore, for each sub-model of GLSANet, we can get the regions of hands via the coordinates of hands/body. The cropping region size is determined related to the shoulder's width ϕ_w, which is set to $2 \times \phi_w \times c$, c is a parameter used to control the size. Thus, the cropping area $\mathcal{A}$ is:

$$\mathcal{A} = \Big\{(x - \phi_w \cdot c, y - \phi_w \cdot c), (x + \phi_w \cdot c, y + \phi_w \cdot c)\Big\} \tag{1}$$

where the first point is the top left coordinate and the second one is the bottom right coordinate. We adjust the parameter c to ensure that it not only includes the hand but is also smaller than 112×112 (we set $c = 0.6$ in our experiments).

The cropped left and right hand regions are shown in the second and third rows of Fig. 2. Besides, we also extract the optical flow for both RGB and depth videos after cropping operation on all of the videos.

Global Spatial Information. Although dynamic gesture recognition is most relevant to the arm or hands, the global information from other parts of the performer's body and environments (*i.e.*, face, background) also provides useful information to increase the robustness of our method. Thus, the raw RGB and depth videos are fed into the C3D network, respectively to capture the global context. Similarly, the optical flow from raw RGB and depth video encode the dynamic gesture in sequences, which refers to temporal information of the video. So optical flow information is also encoded by the C3D network.

Local Spatial Information. To extract more precise information of the gesture, we attempt to focus on the hands, and crop the videos with centroids on the hands, which is helpful to draw local attentions to the hands and extract details of complex gestures. Similar to the global one, the optical flow of the cropped right/left hand is also calculated and fed to the C3D network.

In this way, we can obtain the predictions of 12 C3D models. Then we use an adaptive loss weight layer to fuse the results to get the final prediction.

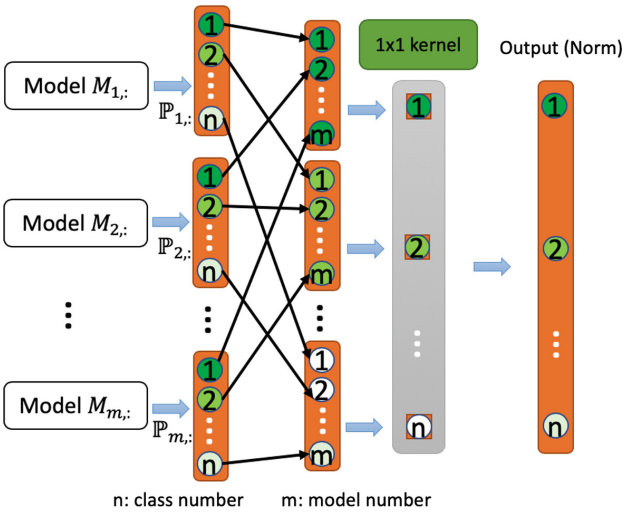


Fig. 3. The class-constrained weight fusion strategy. The 1^{st} column: the classification model; the 2^{nd} column: the output probability vectors of the input models; the 3^{rd} column: class-constrained vector for each class; the 4^{th} column: the 1×1 kernel for convolutional operation; the 5^{th} column: the final predicted result.

3.2 Class-Constrained Fusion Strategy

Inspired by our previous fusion method [20], we also applied a similar fusion strategy for the final prediction. However, compared with [20], the class-constrained fusion strategy has two traits. First, because of the lack of right or left hand in some gestures (*i.e.*, Fig. 3), it would lead to the unavailable local model for the prediction voting. Therefore, we develop a more general scheme for adaptive fusion of all models even if some models are missing. Second, we employ a series of mathematical expressions to derive the fusion processing.

We suppose there are m classification models used for final result fusion, and the $P_{i,j}$ is the predicted probability value of the i-th model $\mathcal{M}_i$ for the j-th class, and n is the class number. The final fusion value for the j-th class can be calculated as:

$$y_j = \sum_{i=1}^{m} (\omega_{i,j} \times \mathbb{P}_{i,j}) = \mathbb{P}_{:,j} \bigotimes \mathbb{K} \ for \ j = 1, \cdots, n$$
$$\mathbb{P}_{i,j} = \begin{cases} P_{i,j}, \ j = 1, \cdots, n & P_{i,j} \text{ exist} \\ 0 & P_{i,j} \text{ not exist} \end{cases} \tag{2}$$

where y_j is the final result for the j-th class and $\omega_{:,j}$ is the weight vector for the j-th class. It can be achieved by using a convolution with a 1×1 kernel $\mathbb{K}$ ($\bigotimes$ is the convolution operation). $\mathbb{P}_{i,j} \in R^{m \times n}$ is a piecewise function which means if the j-th model is missing, then we can directly set $\mathbb{P}_{i,j}$ to zero. We can

see that no matter what value for the weight w, it doesn't influence the final weight fusion voting as shown in Eq. (2). Because y_j is only calculated by its corresponding j-th probabilities of m models in Eq. (2), we call it as the class-constrained fusion strategy. Finally, we normalize $y_j, j = 1, \cdots, n$ via normalized exponential function in the range [0, 1] and then get the final fusion result. The structure are show in Fig. 3

4 Experiments

In this section, we illustrate our experiments on IsoGD dataset. First, the experimental setup is presented, including the running environments and settings. Then, the performances and comparisons on the IsoGD dataset is given. At last, an ablation study analyzes and discusses the effect of each strategy including our spatial attention mechanism and weight scheme.

4.1 Experiment Setup

Our experiments are conducted on three NVIDIA Titan Xp GPUs with PyTorch [21]. For the GLSANet algorithm, we utilize the stochastic gradient descent (SGD) optimization strategy and train the model for up to 20 epochs with the batch size 32, and the initial learning rate and the momentum are set as 0.001 and 0.9, respectively. For the fusion network, we utilize the SGD algorithm to optimize the model with the initial learning rate 0.01, momentum 0.9, and 100 epochs (batch size is 128).

4.2 Experiments on the IsoGD Dataset

Table 1 shows all the results of different combinations of either data modalities or global/local attention models. In the header, M, K and F represent the modalities of RGB, $depth$ and $flow$, while L and R represent attention model on either the left or right hand, respectively. We divide our experiments into four groups according to different attention strategies, namely only global models (marked as g), global and right-hand-attention models (marked as gr), global and left-hand-attention models (marked as gl), and finally the global and both right/left-hand-attention models (marked as glr). Each group (*i.e.*, g) has nine models that have used varied data modalities.

Effectiveness of Local Attention Models. As shown in Table 1, the global and left hand attention model gl_9 achieves an improvement at 2.52% from 70.56% (g_9) to 73.08% (gl_9), while the global and right hand attention model gr_9 has improved 1.14% from 70.56% (g_9) to 71.70%(gr_9). It also shows that the global and left hand attention model gl_9 works better than the right hand attention model gr_9 because most gestures from the IsoGD dataset are with the left hand Moreover, we also can see an improvement of about 0.92% from 73.08% (gl_9) to 74.00% (glr_9). The same similar can be found in the other pairs of the result

Table 1. Accuracies of different fusion combinations on the IsoGD dataset. Column header (varied data modality): M: RGB image, K: depth image, L: local attention image from the left hand, R: local attention image from the right hand, F: optical flow image; Row header (model ID using different training data): *g*: the model using global data, *gr*: the model using both the global and right hand data; *gl*: the model using both the global and left hand data; *glr*: the model using the globe, left hand and right hand data.

ID	Data													
	M_G	M_L	M_R	K_G	K_L	K_R	M_F_G	M_F_L	M_F_R	K_F_G	K_F_L	K_F_R	Valid	Test
g_1	√												56.52%	59.56%
g_2				√									56.21%	64.55%
g_3							√						56.95%	60.90%
g_4										√			56.31%	64.76%
g_5	√						√						57.59%	61.62%
g_6				√						√			57.35%	65.59%
g_7	√			√									64.75%	68.94%
g_8							√			√			65.30%	69.81%
g_9	√			√			√			√			66.08%	70.56%
gr_1	√		√										56.79%	60.15%
gr_2				√		√							56.90%	65.22%
gr_3							√		√				57.22%	61.22%
gr_4										√		√	57.11%	65.08%
gr_5	√		√				√		√				58.18%	62.43%
gr_6				√		√				√		√	58.21%	66.34%
gr_7	√		√	√		√							65.53%	69.99%
gr_8							√		√	√		√	66.17%	70.47%
gr_9	√		√	√		√	√		√	√		√	67.39%	71.70%
gl_1	√	√											57.26%	60.93%
gl_2				√	√								57.92%	65.87%
gl_3							√	√					57.88%	61.41%
gl_4										√	√		57.56%	65.68%
gl_5	√	√					√	√					59.04%	62.86%
gl_6				√	√					√	√		58.80%	67.80%
gl_7	√	√		√	√								66.72%	70.44%
gl_8							√	√		√	√		66.39%	70.99%
gl_9	√	√		√	√		√	√		√	√		68.45%	73.08%
glr_1	√	√	√										57.50%	61.17%
glr_2				√	√	√							58.58%	66.46%
glr_3							√	√	√				58.02%	61.98%
glr_4										√	√	√	58.00%	66.11%
glr_5	√	√	√				√	√	√				59.42%	63.85%
glr_6				√	√	√				√	√	√	60.15%	68.20%
glr_7	√	√	√	√	√	√							67.58%	71.50%
glr_8							√	√	√	√	√	√	66.98%	71.39%
glr_9	√	√	√	√	√	√	√	√	√	√	√	√	**69.76%**	**74.00%**

such as g_7, gr_7, gl_7, glr_7. This proves that both of our left- and right-hand local attention can improve the accuracy of the whole framework.

Effectiveness of Model Fusion. We use the last group models (from glr_1 to glr_9) to illustrate the effectiveness of different data modalities in Table 1. The model glr_1 is trained via global RGB images and glr_2 using global depth images. The performance of the depth model (glr_2, 66.46%) is better than the RGB model (glr_1, 61.17%) about 5%. Similarly, the depth optical flow model (glr_4) is also better than the RGB flow model glr_3 by about 4%. When the RGB and optical flow of the RGB data have used (*i.e.*, glr_5), the performance can also be improved compared with only RGB data used (i.e., glr_1). When all the multiple data modalities are used, namely the RGB-depth-flow model can significantly improved a high accuracy of 74.00%.

Comparison with State-of-the-Arts. Table 2 gives the comparisons with the previous methods, in which our proposed method performs better than all published methods on both the validation set and the testing set. The best accuracy published before on the validation set is 64.40% from Miao *et al.* [3], and the best accuracy on the testing set is 68.42% from Lin *et al.* [20]. Our result of the GLSANet achieves 69.76% on the validation set and 74.00% on the testing set, which improves the accuracy at 5.36% and 5.58%, respectively.

Table 2. Comparison of different methods on the IsoGD dataset.

Method	Backbone	Fusion strategy	Modality of data	Evaluation	
				Valid	Test
Li *et al.* [5] '16	C3D	SVM	RGB-D	49.20%	56.90%
Wang *et al.* [11] '16	VGG-16	Score fusion	depth (DDI+ DDNI+DDMNI)	39.23%	55.57%
Zhu *et al.* [17] '16	pyramidal C3D	Score fusion	RGB-D	45.02%	50.93%
Zhu *et al.* [22] '17	C3D, convLSTM	Score fusion	RGB-D	51.02%	/
Li *et al.* [7] '17	C3D	SVM	RGB-D flow	54.50%	60.93%
Miao *et al.* [3] '17	ResC3D	SVM	RGB-D flow	64.40%	67.71%
Wang *et al.* [23] '17	convLSTM, Resnet-50, C3D	Score fusion	RGB-D saliency	60.81%	65.59%
Zhang *et al.* [24] '17	convLSTM, C3D	Score fusion	RGB-D flow	58.00%	60.47%
Duan *et al.* [4] '17	2S CNN, C3D	Score fusion	RGB-D saliency	49.17%	67.26%
Lin *et al.* [20] '18	Skeleton LSTM, C3D	Adaptive weight fusion	RGB-D Skeleton	64.34%	68.42%
GLSANet (Ours)	C3D	Adaptive weight fusion	RGB-D flow, skeleton	**69.76%**	**74.00%**

4.3 Weight Fusion Analysis

Besides the weight fusion strategy, we also show the maximum and average fusion strategies. The maximum fusion is to select the maximum probability as the predicted result, while the average fusion is to calculate the average value of all the models as the final fusion result. The comparisons among three fusion strategies on the IsoGD dataset are shown in Table 3. We can see that the weight fusion strategy get the best performance on both the validation and testing sets, which is much better than the max and average fusion strategies.

Table 3. Comparisons among different fusion methodologies on the IsoGD dataset.

Fusion method	Validation set	Testing set
Max fusion	61.67%	65.51%
Average fusion	62.02%	65.97%
Weight fusion	**69.76%**	**74.00%**

5 Conclusion

In the paper, we propose a novel gesture recognition architecture GLSANet and an improved adaptive fusion strategy. On the one hand, resnet based C3D network plays an important role in extracting global and local spatial attention features. On the other hand, the proposed adaptive fusion strategy fuses results of each category from different morality input efficiently. The state-of-the-art performance demonstrates the effectiveness of our method. Although the proposed method shows remarkable results, several venues still need further exploration.

Acknowledgments. This work has been partially supported by the Chinese National Natural Science Foundation Projects #61876179, #61872367, and by Science and Technology Development Fund of Macau (Grant No. 0025/2018/A1). We acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

References

1. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015)
2. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In: CVPRW, pp. 56–64 (2016)
3. Miao, Q., et al.: Multimodal gesture recognition based on the ResC3D network. In: ICCVW, pp. 3047–3055 (2017)
4. Duan, J., Wan, J., Zhou, S., Guo, X., Li, S.: A unified framework for multi-modal isolated gesture recognition. TOMM **9**(4) (2017)

5. Li, Y., et al.: Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model. In: ICPR, pp. 25–30. IEEE (2016)
6. Li, Y., et al.: Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model PRL (2017)
7. Li, Y., et al.: Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model. TCSVT **28**, 2956–2964 (2017)
8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV, pp. 2556–2563. IEEE (2011)
9. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild arXiv preprint arXiv:1212.0402 (2012)
10. Kay, W., et al.: The kinetics human action video dataset arXiv preprint arXiv:1705.06950 (2017)
11. Wang, P., Li, W., Liu, S., Gao, Z., Tang, C., Ogunbona, P.: Large-scale isolated gesture recognition using convolutional neural networks. In: ICPR, pp. 7–12. IEEE (2016)
12. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T.: Rank pooling for action recognition. TPAMI **39**(4), 773–787 (2017)
13. Chai, X., Liu, Z., Yin, F., Liu, Z., Chen, X.: Two streams recurrent neural networks for large-scale continuous gesture recognition. In: ICPR, pp. 31–36. IEEE (2016)
14. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: data level fusion strategy for hand gesture recognition. In: CVPR, pp. 2103–2111 (2018)
15. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV, pp. 4489–4497 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
17. Zhu, G., Zhang, L., Mei, L., Shao, J., Song, J., Shen, P.: Large-scale isolated gesture recognition using pyramidal 3D convolutional networks. In: ICPR, pp. 19–24. IEEE (2016)
18. Tran, D., Ray, J., Shou, Z., Chang, S.F., Paluri, M.: ConvNet architecture search for spatiotemporal feature learning arXiv preprint arXiv:1708.05038 (2017)
19. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
20. Lin, C., Wan, J., Liang, Y., Li, S.Z.: Large-scale isolated gesture recognition using masked Res-C3D network and skeleton LSTM. In: FG (2018)
21. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
22. Zhu, G., Zhang, L., Shen, P., Song, J.: Multimodal gesture recognition using 3D convolution and convolutional LSTM. IEEE Access **5**, 4517–4524 (2017)
23. Wang, H., Wang, P., Song, Z., Li, W.: Large-scale multimodal gesture recognition using heterogeneous networks. In: ICCVW, pp. 3129–3137 (2017)
24. Zhang, L., Zhu, G., Shen, P., Song, J., Shah, S.A., Bennamoun, M.: Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In: ICCV, pp. 3120–3128 (2017)