Large-scale Multimodal Gesture Recognition Using Heterogeneous Networks

Huogen Wang*[1], Pichao Wang*[2], Zhanjie Song[3], Wanqing Li[4]
[1]School of Electrical and Information Engineering, Tianjin University, China
[1,2,4]Advanced Multimedia Research Lab, University of Wollongong, Australia
[3]School of Mathematics, Tianjin University, China
{[1]hw823,[2]pw212}@uowmail.edu.au,[3]zhanjiesong@tju.edu.cn,[4]wanqing@uow.edu.au

Abstract

This paper presents the method designed for the 2017 ChaLearn LAP Large-scale Gesture Recognition Challenge. The proposed method converts a video sequence into multiple body level dynamic images and hand level dynamic images as the inputs to Convolutional Neural Networks (ConvNets) respectively through bidirectional rank pooling and adopts Convolutional LSTM Networks (ConvLSTM) to learn long-term spatiotemporal features from short-term spatiotemporal features extracted using a 3D convolutional neural network (3DCNN) at body and hand level. Such a heterogeneous network system learns effectively different levels of spatiotemporal features that are complementary to each other to improve the recognition accuracy largely. The method has been evaluated on the 2017 isolated and continuous ChaLearn LAP Large-scale Gesture Recognition Challenge datasets and the results are ranked among the top performances.

1. Introduction

Gesture recognition from visual information is an active topic with many potential applications in human computer interaction [30], human robot interaction, sign recognition and virtual reality. Due to subtle differences among similar gestures, complex scene background, different observation conditions, and noises in acquisition, robust gesture recognition is very challenging [28].

Gesture recognition aims to recognize and understand meaningful movement of human bodies [1] in which arms and hands play crucial roles. Only few gestures can be identified from their spatial or structure information. In fact, motion cues and structure information simultaneously characterize a unique gesture. How to learn spatiotemporal features effectively is always the key in gesture recognition.

With the recent development of deep learning, a few methods [38, 39, 6, 37, 23, 33, 5] have been developed for gesture recognition based on ConvNets or RNNs. The ConvNet based methods use a single dynamic image [12, 2, 11, 13, 10, 41, 40] or a depth motion map [38, 39, 42] to represent the spatiotemporal information of a video sequence, and then the dynamic images and depth motion maps are fed into ConvNets for final classification. Dynamic images are compact, however, temporal information is inevitably lost to some extent during the conversion of video into its dynamic images. RNN based methods usually cascade ConvNets or 3DCNNs with RNN [5, 27, 47]. It tends to overemphasize the temporal information and overlook structure information.

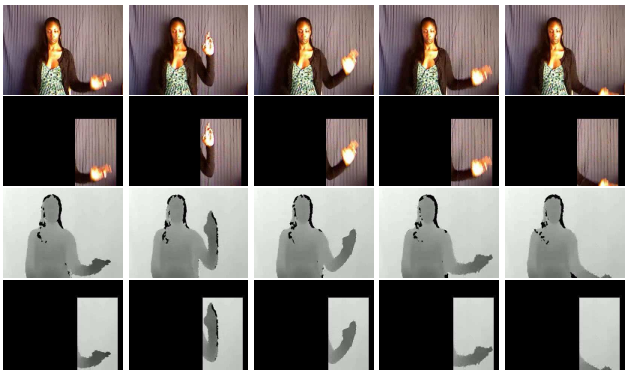

Figure 1. Examples of image frames at body level and hand level. From up to bottom: body level RGB images,hand level RGB images, body level depth images and hand level depth images.

Unlike previous methods utilizing a single type of network or cascading multiple types of networks, we investigate a different architecture based on heterogeneous networks. The heterogeneous networks consist of two separate recognition components built upon multiple ConvNets and 3D ConvLSTMs, which are then combined by late fusion. The ConvNets are trained to recognize gestures from dynamic images, whilst the 3D ConvLSTMs perform gesture recognition from still video frames. The ConvNets are

*Both authors contributed equally to this work

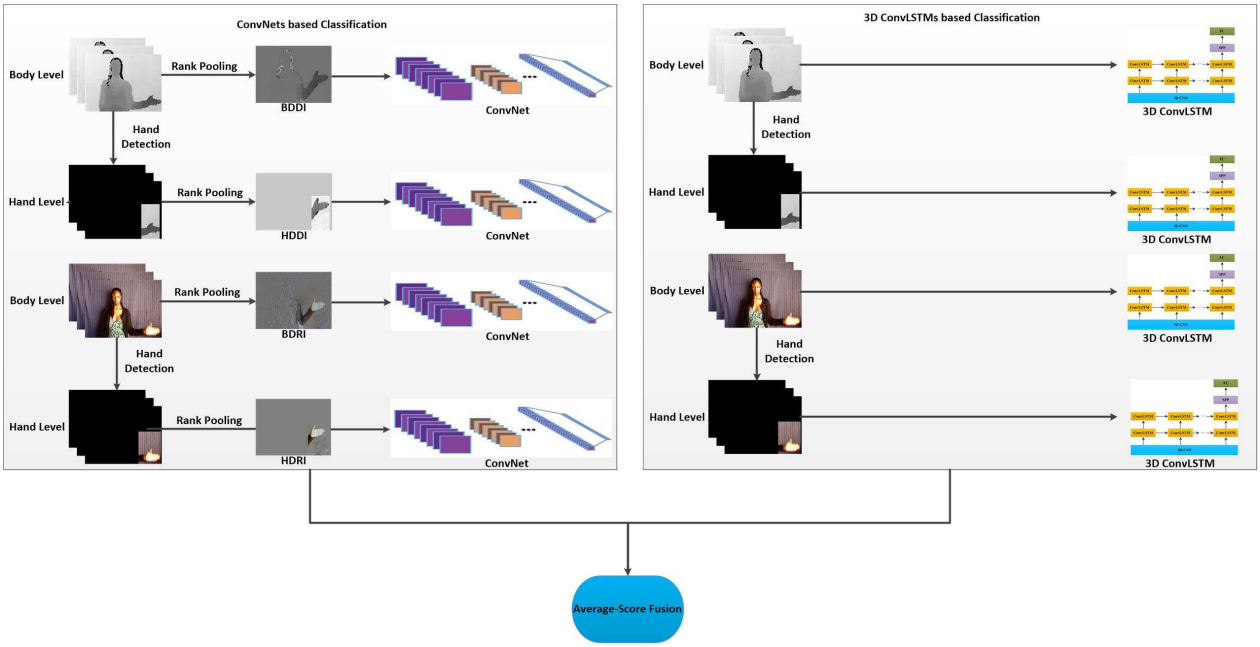

Figure 2. Overview of the proposed method.

pre-trained on the ImageNet dataset [32] and the 3D ConvLSTMs are same as those described in [47]. This architecture is able to learn spatiotemporal features for gesture recognition avoiding to overemphasize either spatial or temporal features. One of the sources of interference to gesture recognition is background, especially cluttered background. To address this problem, this paper proposes to apply heterogeneous networks on video sequences at two spatial levels, namely, body and hand levels. As shown in Figure 1, this simple scheme can reduce the influence of background and clothes and learn spatiotemporal features from global to fine-grained levels.

The continuous gesture sequences can be segmented into several isolated gestures based on the quantity of movement (QOM) [19, 42], so continuous gesture recognition can be converted to isolated gesture recognition. The proposed method is evaluated on the 2017 ChaLearn LAP Large-scale Gesture Recognition Challenge datasets. The results have shown that its performances on the Isolated and Continuous Gesture Recognition tasks are ranked among the top.

The rest of this paper is organised as follows. Section 2 reviews the related work on deep learning based gesture recognition. Section 3 describes the proposed method. Section 4 presents the experimental results and the discussions. The paper is concluded in Section 5.

2. Related Work

Recently, deep learning methods based on the Convolutional Neural Network (ConvNet) and Recurrent Neural Network (RNN) have achieved remarkable success in gesture recognition [9, 8, 7]. Exiting deep learning based gesture recognition can be divided into four categories. The first approach applies ConvNet to extract spatial features from individual frames and fuse the temporal information later. Karpathy et al. [21] explored four temporal fusion methods, and proposed that slow fusion can get more global information in both spatial and temporal dimensions. Ng et al. [29] explored several temporal pooling methods and showed that max pooling in the temporal domain leads to significant improvements.

The second approach is to encode spatial and temporal features from RGB and stacked optical flow separately. Simonyan et al. [33] proposed a two-stream ConvNet architecture which incorporates spatial and temporal networks, and the two streams are fused together at later stage. Methods extending the two-stream networks were also proposed by integrating improved trajectories [36], motion vector [44] and Motion History Image [22].

The third approach is to extend convolutional operation into temporal domain [18, 34]. C3D based methods demonstrated the state-of-the-art result on the 2016 ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge. Zhu et al. [46] embedded pyramidal input and pyramidal fusion strategies into the C3D model for gesture recognition. Li et al. [24] applied the C3D model on the RGB and depth data respectively. Molchanov et al. [27] proposed recurrent 3D convolutional neural networks which integrate C3D and LSTM for gesture recognition.

The fourth approach is to encode the video into single image that contain the spatiotemporal information and then apply ConvNet for image-based recognition. Wang et al. [38, 39] encode depth map sequences into texture color images using the concepts of Depth Motion Maps (DMM) and pseudo-coloring. Bilen et al. [2] and Fernando et al. [12, 13, 10, 11] proposed to adopt rank pooling to encode the video into one dynamic image and used pre-training models over ImageNet dataset for fine-tuning, their methods achieved the promising results on RGB data. Wang et al. [41] encoded the depth map sequences into three kinds of dynamic images with rank pooling which can learn the posture and motion information from three different levels.

The first two approaches learn spatiotemporal features separately or at different stages. The other two approaches learn spatiotemoral features simultaneously, but the features are different. The proposed gesture recognition method in this paper combine the last two approaches to improve the performance by using the score fusion. The fusion method can make full use of the different kinds of features and improve the recognition accuracy largely.

3. Proposed Method

As shown in Figure 2, the proposed method consists of four components: a faster R-CNN for hand detection, ConvNets-based classification from dynamic images, 3D ConvLSTMs based classification from video sequences and score fusion of the outputs from the ConvNets and 3D ConvLSTMs for final gesture recognition. Given an isolated gesture sequence, the faster R-CNN proposed in [31] detects hand region frame by frame and the hand images are cropped using the biggest bounding box of hands in the frames. On the one hand, dynamic images are constructed for the body and hand parts from the RGB-D sequences and fed to the ConvNets. On the other hand, the RGB-D sequences of the body and hands are input to the 3D ConvLSTMs. The ConvNets and 3D ConvLSTMs are designed to learn complementary spatiotemporal features at the two spatial levels (i.e. body and hand) to imporve the recognition.

3.1. Faster R-CNN based Hand Detection

In order to learn the spatiotemporal features at body level and hand level respectively, hand detection on both RGB and depth images is crucial. Hand regions are usually detected by color or multiple cues, but these methods are sensitive to illumination and background. Inspired by the promising performance of the region-based convolutional neural networks (R-CNNs) [15] in object detection, the proposed method in this paper adopt the faster R-CNN [31] to detect the hand regions.

The Faster R-CNN consists of two modules. The first, called the Regional Proposal Network (RPN) [26], is a fully convolutional network for generating object proposals that will be fed into the second module. The second module is the Fast R-CNN detector [14] whose purpose is to refine the proposals. The key idea is to share the same convolutional layers for the RPN and Fast R-CNN detector up to their own fully connected layers. Now the image only passes through the ConvNet once to produce and then refine object proposals. The Faster R-CNN is an end to end network that takes an image as input and outputs a set of rectangular detected objects with the class probabilities.

After hand region detected frame by frame in a video sequence, the biggest bounding box of the hand can be detected through the whole sequence. Then the hand level images can be cropped.

3.2. ConvNets-based Classification

Firstly, the four sets of dynamic images, Body Level Dynamic Depth Images (BDDIs), Hand Level Dynamic Depth Images (HDDIs), Body Level Dynamic RGB Images (BDRIs) and Hand Level Dynamic RGB Images (HDRIs) are constructed from an image sequence through bidirectional rank pooling. Each level dynamic image is represented by two motion images, forward and backward.

3.2.1 Rank Pooling

Given a sequence with k frames, which can represented as $X =< x_1, x_2, \cdots, x_t, \cdots, x_k >$. And $\varphi(x_t) \in \mathbb{R}^d$ be a representation or feature vector extracted from each frame x_t. Let $V_t = \frac{1}{t}\sum_{\tau=1}^{t}\varphi(x_t)$ be time average of these features up to time t. At each time t, a score $r_t = \omega^T \cdot V_t$ is assigned. In general, later times are associated with larger scores, so the score satisfies $r_i > r_j \Leftrightarrow i > j$. The process of rank pooling is to find ω^* that satisfies the following objective function:

$$\begin{aligned} &\underset{\omega}{argmin}\frac{1}{2}\left\|\omega\right\|^2 + \lambda\sum_{i>j}\varepsilon_{ij}, \\ s.t. \quad &\omega^T \cdot (V_i - V_j) \geq 1 - \varepsilon_{ij}, \varepsilon_{ij} \geq 0 \end{aligned} \tag{1}$$

The parameters ω^* represent the information that frame representation V_t comes before the frame representation V_{t+1}, and can be used as a descriptor of the sequence. ε_{ij} is the smallest non-negative number.

3.2.2 Construction of Dynamic Images

In this paper, we apply the rank pooling directly on the pixels of video sequence to form dynamic images. Different from the work [2], the rank pooling is applied in a bidirectional way to convert one video sequence into two dynamic images. Each dynamic image is fed into a ConvNet. The resulting dynamic images are illustrated in Figure 3. As shown, dynamic images effectively capture the structure information.

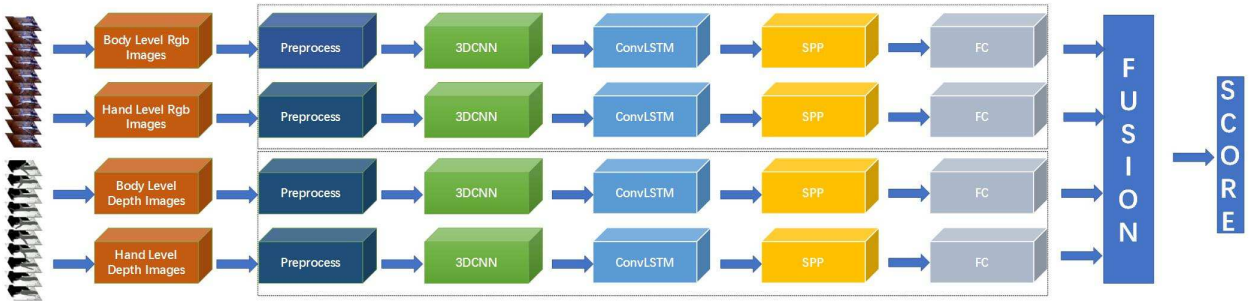

Figure 4. The framework of 3D ConvLSTM. Body level depth images, body level RGB images, hand level depth images and hand level RGB images are fed into 3D ConvLSTM.

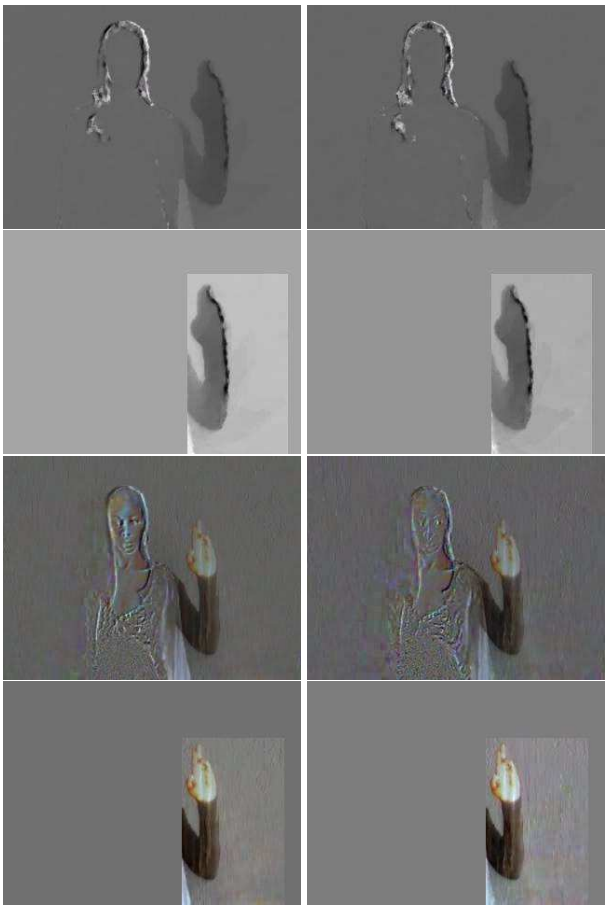

Figure 3. Samples of generated forward and backward BDDIs, HDDIs, BDRIs and HDRIs for gesture Mudra1/Ardhapataka, the left images are dynamic images for forward, the right images are dynamic images. From up to bottom: BDDIs, HDDIs, BDRIs and HDRIs.

3.3. 3D ConvLSTMs based Classification

The 3D ConvLSTM network is described in detail by Zhu et al. [47]. As shown in Figure 4, the 3D ConvLSTM network is composed of four components: Input preprocessing, 3D Convolutional Networks, Convolutional LSTM, Spatial Pyramid Pooling. This method uses uniform sampling with temporal jitter based on pyramid input to down sample each gesture sequence into a fixed length. The sampling process can be described as follow.

$$Idx_i = \frac{S}{L} \cdot (i + jit/2) \tag{2}$$

Where Idx_i is the index of ith sampled frame, and jit is a random value sampled form the uniform distribution between -1 and 1. And the sampling result can be represented as follow.

$$US = \{Idx_1, Idx_2, \cdots, Idx_L\} \tag{3}$$

After this sampling process, the video sequence is fed into 3DCNN [34] to learn short-term spatiotemporal features. Two-level ConvLSTM [43] is adopted to learn long-term spatiotemporal features from short-term spatiotemporal features. The final output of the high level ConvLSTM layer is considered as the final long-term spatiotemporal features for each gesture. The output of ConvLSTM has same spatial size as the output of 3D convolutional networks. The full-connected layers need to have fixed-size/length input by their definition. So the spatial pyramid pooling (SPP) [16] is added on the top of ConvLSTM and connected to the full connected layer. Different from [47], we feed four sets of still video frames, including body level depth images, body level RGB images, hand level depth images and hand level RGB images, into the 3D ConvLSTM networks. Hand level depth images and hand level RGB images focus the motions of hands, which help reduce the influence of background.

3.4. Score Fusion for Classification

Given a pair of RGB and depth video sequences, eight levels dynamic images are generated and fed into eight independently trained ConvNets, and the four sets of still video frames are also fed into the 3D ConvLSTM networks. For recognition, average-score fusion is used. The score

vectors outputted by ConvNets and 3D ConLSTMs are averaged in an element-wise way, and the max score in the resultant vector is assigned as the probability of the test sequence. The index of this max score corresponds to the predicted class label.

4. Experiments

In this section, the ChaLearn Gesture Datasets (CGD) [35] and evaluation protocols are described. The experimental results of the proposed method on the datasets are reported. The final results were obtained by the challenge organisers running the code on the test datasets.

4.1. Datasets

The full ChaLearn Gesture Dataset (CGD) was recorded by Microsoft Kinect sensor [45]. It includes color and depth video sequences provided by the sensor but no human pose information was acquired. The ChaLearn LAP IsoGD Dataset and the ChaLearn LAP ConGD Dataset are derived from ChaLearn Gesture Dataset (CGD). The ChaLearn LAP IsoGD Dataset includes 47,933 RGB-D gesture video, and each RGB-D video representing one gesture instance. There are 249 types of gestures performed by 21 different individuals. The detailed information of the ChaLearn LAP IsoGD dataset is shown in Tabel 1. The ChaLearn LAP ConGD Dataset includes 47,933 RGB-D gesture instances in 22,535 RGB-D gesture videos. Each RGB-D video may represent one or more gestures, and there are also 249 gestures performed by 21 different individuals. The detailed information of the ChaLearn LAP ConGD dataset is shown in Tabel 2.

4.2. Network Training

4.2.1 Network Training for ConvNet

After the construction of BDDIs, HDDIs, BDRIs and HDRIs, eight ConvNets are trained on the eight channels individually. In this paper, the ResNet-50 [17] is adopted as the ConvNet model. We fine-tune the ConvNets on BDDIs and HDDIs with pre-training models on ILSVRC-2015 [32], and then fine-tune the ConvNets on BDRIs and HDRIs with pre-training models on BDDIs and HDDIs separately. The network weights are learned using mini-batch stochastic gradient descent with the momentum being set to 0.9 and weight decay being set to 0.0001. All hidden weight layers use the rectification (RELU) activation function. At each iteration, a mini-batch of 16 samples is sampling 16 shuffled training samples. All the images are resized to 224×224. The learning rate for fine-tuning is set to 10^{-4}, and then it is decreased according to a fixed schedule, which is kept the same for all training sets. For the ConvNet, the training undergoes 90K iterations and the learning rate is dropped to its 0.96 every 40K iterations.

Methods	Accuracy
Body Level (ConvNet)	49.14%
Hand Level (ConvNet)	50.36%
Hand Level + Body Level (ConvNet)	53.65%
Body Level (3D ConvLSTM)	51.31%
Hand Level (3D ConvLSTM)	48.32%
Hand Level + Body Level (3D ConvLSTM)	53.09%
Body Level (ConvNet+3D ConvLSTM)	57.85%
Hand Level (ConvNet+3D ConvLSTM)	54.67%
All-Score Fusion	**60.81%**

Table 3. The result of different schemes on validation set of ChaLearn LAP IsoGD

4.2.2 Network Training for 3D ConvLSTM

The 3D ConvLSTM is implemented based on the tensorflow and Tensorlayer platforms. Four level video sequences based networks were trained separately. We fine-tuned the networks on depth modality based on the pre-training model on SKIG [25] and then fine-tuned the networks on RGB modality based on the pre-training model of the depth modality. Batch normalization makes training processes easier and faster. The initial learning rate is set to 0.1 and dropped to its $\frac{1}{10}$ every 15K iterations. The weight decay is initialized as 0.004 and decreases to 0.00004 after 40K iterations. At most 60K iterations are needed for training. At each iteration, the batch-size is 13, the temporal length of each clip is 32 frames, and the crop size for each image is 112.

4.3. Evaluation on ChaLearn LAP IsoGD

For the isolated gesture recognition challenge, the recognition rate r is used as the evaluation criteria. The recognition rate is calculated as follow.

$$r = \frac{1}{n}\sum_{i=1}^{n}\delta(p_l(i), t_l(i)) \tag{4}$$

where n is the number of samples; p_l is the predicted label; t_l is the ground truth; $\delta(j_1, j_2) = 1$, if $j_1 = j_2$, otherwise $\delta(j_1, j_2) = 0$.

The result of individual levels and different networks on validation set are listed in Table 3. The following conclusions can be derived: (i) body level and hand level are complementary, as their fusion improves on both; (ii) Score fusion of ConvNet and 3D ConvLSTM greatly improves the final result.

Table 4 compares the performance of the proposed method and that of exiting methods on validation set. It can be seen that the proposed method achieved the state-of-the-art results compared with both hand-crafted features based methods and deep learning methods.

Sets	# of Gestures	# of RGB Videos	# of Depth Videos	# of Subjects
Training	35878	35878	35878	17
Validation	5784	5784	5784	2
Testing	6271	6271	6271	2
All	47933	47933	47933	21

Table 1. Information of the ChaLearn LAP IsoGD Dataset

Sets	# of Gestures	# of RGB Videos	# of Depth Videos	# of Subjects
Training	30442	14134	14134	17
Validation	8889	4179	4179	2
Testing	8602	4042	4042	2
All	47933	22535	22535	21

Table 2. Information of the ChaLearn LAP ConGD Dataset

Methods	Accuracy
MFSK [35]	18.65%
MFSK+DeepID [35]	18.23%
Scene Flow [40]	36.27%
Wang et al. [41]	39.23%
Pyramidal C3D [46]	45.02%
Duan et al. [7]	49.17%
Li et al. [24]	49.2%
C3D+ConvLSTM [47]	51.02%
Proposed Method	**60.81%**

Table 4. Comparison of proposed method with other method on the validation set of ChaLearn LAP IsoGD

The results on the ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge dataset are summarized in Table 5 [20]. The final ranking is based on the test evaluation phase. We can see that our method is among the top performances, Our team obtains 65.59% accuracy on the test dataset and place fourth in the ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge (Round 2) in ICCV 2017.

4.4. Evaluation on ChaLearn LAP ConGD

The proposed method has been also applied to continuous gesture recognition. The continuous gesture sequence was first segmented into several isolated gesture sequences based on quantity of movement (QOM) [19, 42]. For continuous gesture recognition, the Jaccard index (the higher the better) is adopted to measure the performance. The Jaccard index measures the average relative overlap between true and predicted sequences of frames for a given gesture. For a sequence s, let $G_{s,i}$ and $P_{s,i}$ be binary indicator vectors for which 1-value correspond to frames in which the i^{th} gesture label is being performed. The Jaccard Index for i^{th} class is defined for the sequence s as follow.

Methods	Mean Jaccard Index $\overline{J_S}$
MFSK [35]	0.0918
MFSK+DeepID [35]	0.0902
Wang et al. [42]	0.2403
Chai et al. [4]	0.2655
Camgoz et al. [3]	0.2809
Proposed Method	**0.5957**

Table 6. Comparison of the proposed method with other methods on the validation set of ChaLearn LAP ConGD

$$J_{s,i} = \frac{G_{s,i} \bigcap P_{s,i}}{G_{s,i} \bigcup P_{s,i}} \tag{5}$$

where $G_{s,i}$ is the ground truth of the i^{th} gesture label in sequence s, and $P_{s,i}$ is the prediction for the i^{th} label in sequence s.

When $G_{s,i}$ and $P_{s,i}$ are empty, $J_{s,i}$ is defined to be 0. Then for the sequence s with l_s true labels, the Jaccard Index J_s is calculated as follow.

$$J_s = \frac{1}{l_s} \sum_{i=1}^{L} J_{s,i} \tag{6}$$

where L is the number of gesture labels. For all testing sequences $S = s_1, \cdots, s_n$ with n gestures, the mean Jaccard Index $\overline{J_S}$ is used as the evaluation criteria and calculated as follow.

$$\overline{J_S} = \frac{1}{n} \sum_{j=1}^{n} J_{s_j} \tag{7}$$

Tabel 6 compares the performance of the proposed method and that of exiting methods on the validation dataset. It can be seen that the proposed method achieve the state-of-the-art results.

Rank by test set	Team	Recognition Rate r (valid set)	Recognition Rate r (test set)
1	ASU	**64.40%**	**67.71%**
2	SYSU_ISEE	59.70%	67.02%
3	Lostoy	62.02%	65.97%
4	**AMRL(ours)**	60.81%	65.59%
5	XDETVP	58.00%	60.47%
-	baseline [7]	49.17%	67.26%

Table 5. Comparison of the performance of our method with others in ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge

Rank by testing dataset	Team	Mean Jaccard Index $\overline{J_S}$ (valid set)	Mean Jaccard Index $\overline{J_S}$ (test set)
1	ICT_NHCI	0.5163	**0.6103**
2	**AMRL(ours)**	**0.5957**	0.5950
3	PaFiFA	0.3646	0.3744
4	Deepgesture	0.3190	0.3164

Table 7. Comparison of the performance of the proposed method with other methods in ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge

The results of ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge are listed in Table 7 [20]. The final ranking is based on the test evaluation phase. It can be seen that our method is among the top performance. Our mean Jaccard Index is **0.5950** on the test dataset, which is placed second and very close to the best performance of the ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge (Round 2) in ICCV 2017. It should be noticed that the proposed method achieves the best performance on the validation dataset.

5. Conclusion

This paper presents an effective method for large-scale multimodal gesture recognition using heterogeneous networks. The proposed method take both advantages of ConvNet based on dynamic image method and 3D ConvLSTM method. The evaluation results demonstrate that our heterogeneous networks can learn effectively different levels of spatiotemporal features and these features are complementary to each other.

Acknowledgment

Huogen Wang and Pichao Wang gratefully acknowledge financial support from China Scholarship Council. Zhanjie Song was partly supported by National Natural Science Foundation of China (Grant No.61379014) and Natural Science Foundation of Tianjin (Grant No.16JCYBJC15900).

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011. 1

[2] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016. 1, 3

[3] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 49–54. IEEE, 2016. 6

[4] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen. Two streams recurrent neural networks for large-scale continuous gesture recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 31–36. IEEE, 2016. 6

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 1

[6] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 1

[7] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Li. A unified framework for multi-modal isolated gesture recognition. In *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),(under review, round 2)*, 2017. 2, 6, 7

[8] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, et al. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 67–73. IEEE, 2016. 2

[9] S. Escalera, V. Athitsos, and I. Guyon. Challenges in multimodal gesture recognition. *Journal of Machine Learning Research*, 17(72):1–54, 2016. 2

[10] B. Fernando, P. Anderson, M. Hutter, and S. Gould. Discriminative hierarchical rank pooling for activity recognition. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1924–1932, 2016. 1, 3

[11] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2017. 1, 3

[12] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015. 1, 3

[13] B. Fernando and S. Gould. Learning end-to-end video classification with rank-pooling. In *International Conference on Machine Learning*, pages 1187–1196, 2016. 1, 3

[14] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[16] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 4

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 2

[19] F. Jiang, S. Zhang, S. Wu, Y. Gao, and D. Zhao. Multi-layered gesture recognition with kinect. *The Journal of Machine Learning Research*, 16(1):227–254, 2015. 2, 6

[20] W. Jun, S. Escalera, A. Gholamreza, H. J. Escalante, X. Baró, I. Guyon, M. Madadi, A. Juri, G. Jelena, L. Chi, and X. Yiliang. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. 6, 7

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[22] C. Lea, A. Reiter, R. Vidal, and G. D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2016. 2

[23] G. Lefebvre, S. Berlemont, F. Mamalet, and C. Garcia. Inertial gesture recognition with blstm-rnn. In *Artificial Neural Networks*, pages 393–410. Springer, 2015. 1

[24] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 25–30. IEEE, 2016. 2, 6

[25] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, volume 4, page 8, 2013. 5

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3

[27] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016. 1, 2

[28] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multi-scale deep learning for gesture detection and localization. In *Workshop at the European conference on computer vision*, pages 474–490. Springer, 2014. 1

[29] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, pages 1–10, 2015. 2

[30] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015. 1

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2017. 3

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 5

[33] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 1, 2

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2, 4

[35] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016. 5, 6

[36] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015. 2

[37] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 1

[38] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona. Convnets-based action recognition from depth maps

through virtual cameras and pseudocoloring. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1119–1122. ACM, 2015. 1, 3

[39] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona. Action recognition from depth maps using deep convolutional neural networks. *IEEE Transactions on Human-Machine Systems*, 46(4):498–509, 2016. 1, 3

[40] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. *arXiv preprint arXiv:1702.08652*, 2017. 1, 6

[41] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 7–12. IEEE, 2016. 1, 3, 6

[42] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona. Large-scale continuous gesture recognition using convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 13–18. IEEE, 2016. 1, 2, 6

[43] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 4

[44] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2016. 2

[45] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 5

[46] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 19–24. IEEE, 2016. 2, 6

[47] G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3d convolution and convolutional lstm. *IEEE Access*, 2017. 1, 2, 4, 6