Large-scale Multimodal Gesture Segmentation and Recognition based on Convolutional Neural 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 an effective method for continuous gesture recognition. The method consists of two modules: segmentation and recognition. In the segmentation module, a continuous gesture sequence is segmented into isolated gesture sequences by classifying the frames into gesture frames and transitional frames using two stream convolutional neural networks. In the recognition module, our method exploits the spatiotemporal information embedded in RGB and depth sequences. For the depth modality, our method converts a sequence into Dynamic Images and Motion Dynamic Images through rank pooling and input them to Convolutional Neural Networks respectively. For the RGB modality, our method adopts Convolutional LSTM Networks to learn long-term spatiotemporal features from short-term spatiotemporal features obtained by a 3D convolutional neural network. Our method has been evaluated on ChaLearn LAP Large-scale Continuous Gesture Dataset and achieved the state-of-the-art performance.

1. Introduction

Gesture recognition from visual information is an active research topic and has many potential applications in human computer interaction [33], human robot interaction, sign recognition and virtual reality. Due to tiny differences among similar gestures, complex scene background, different observation conditions, and noise in acquisition, effective gesture recognition remains challenging [30].

The majority of gesture recognition methods focuses on the isolated gesture recognition. However, cases containing unknown numbers, unknown orders and unknown boundaries of gestures occurs commonly in practice [20]. For such continuous gesture recognition, both the segmentation and the recognition problems need be solved either separately or simutaneously. The method proposed in this paper falls into the category of the former.

Many methods have been proposed for temporal segmentation using hand positions and motion [31, 5]. These methods are highly dependent on hand detection or sensitive to complex background. Instead of using motion information for temporal segmentation, the appearance-based approach is proposed in this paper. As shown in Figure 1, a continuous gesture sequence is composed of gesture frames that cover useful hand movement and transitional frames between two gestures. Segmentation of gestures can be achieved by classifing frames into two classes: gesture frames and trasitional frames. In this paper, this binary calssification is performned by fusing the independent classification using ConvNets on the RGB and depth frames. After classification of the frames, the middle point of any segment of transitional frames is defined as the boundary between two gestures. The segmented gestures are then recognized with the proposed multimodal gesture recognition networks.

Recognition of the segmented gestures are carried out on both RGB and depth video sequences. RGB sequences mainly contain appearance information such as color and texture and depth sequences mainly carries geometric and structural information, complementing the appearance information in the RGB sequences. Extensive works have been reported on video recognition from RGB modality [8, 44, 24, 37, 7] and for depth modality [45, 46, 49, 48, 47]. This paper aims to leverage the different types of information, apearance, geometric, motion and structural as well from the RGB and depth modality for robust gesture recognition.

Specifically, a multimodal gesture recognition network is proposed. For the RGB modality, the proposed method adopts 3D ConvLSTM to learn spatiotemporal features as describe in [58]. The 3D ConvLSTM adoptes Convolu-

[*] Both authors contributed equally to this work

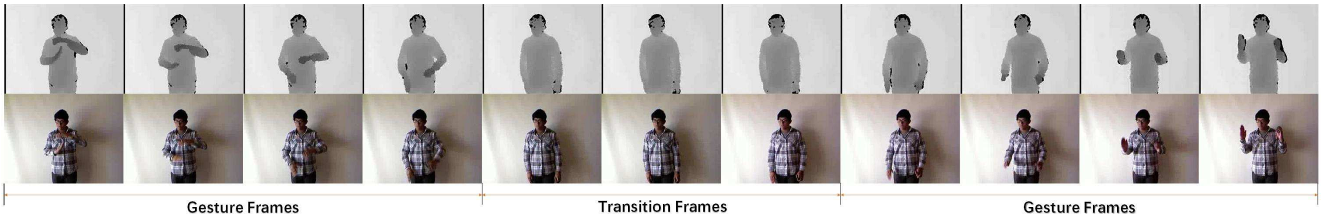

Figure 1. The sample gesture sequence. A continuous gesture sequence is composed of gesture frames and transitional frames between two guestures.

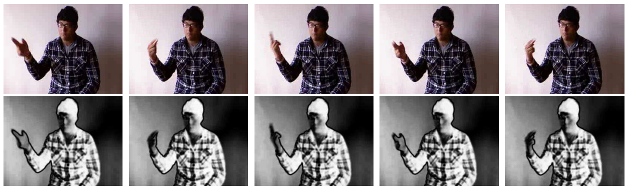

Figure 2. Illustration of a RGB sequence (top) and its saliency sequence (bottom).

tional LSTM Networks (ConvLSTM) [51] to learn long-term spatiotemporal features from short-term spatiotemporal features extracted using a 3D convolutional neural network (3DCNN) [19, 39]. The 3D ConvLSTMs performs gesture recognition from still video frames of a RGB sequence and its saliency sequence extracted using the method decribed in [1]. As shown in Figure 2, the saliency sequence helps eliminate interference of background. For depth modality, inspired by the performance of rank pooling method [15, 2, 14, 16, 13] on depth sequence [48, 47], this paper employs rank pooling to encode depth sequences into Depth Dynamic Images (DDIs). However, this process converts a video sequence into images and at the same time lose some temporal information. To address this problem, the Depth Motion Dynamic Image (DMDI) is introduced. The Depth Motion Dynamic Images apply rank pooling to the absolute differences (motion energy) between consecutive frames of a depth sequence. As shown in Figure 2, the Depth Motion Dynamic Image preserves both motion cues and structural information. The DDIs and DMDIs are fed to ConvNets to recognize gestures. The multiple 3D ConvLSTMs and ConvNets are combined through late fusion. The proposed method was evaluated on the ChaLearn LAP Large-scale continuous Gesture Recognition Challenge datasets (ChaLearn LAP ConGD). The results are state-of-the-art.

The rest of this paper is organised as follows. Section 2 reviews the related work on temporal segmentation and multimodal gesture recognition based on deep learning. Section 3 gives the details of the proposed method. Section 4 presents the experiments and the discussions. The paper is concluded in Section 5.

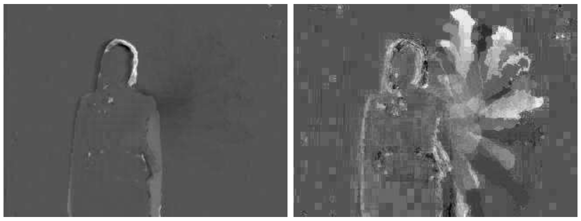

Figure 3. illustration of a Dynamic Image (left) and Motion Dynamic Image (right).

2. Related Work

In this section, the related works on temporal segmentation and multimodal gesture recognition based deep learning are briefly reviewed.

2.1. Temporal Segmentation

Various methods have been proposed for temporal segmentation. The exiting methods can be divided into four categories. The first category employs some models, such as Dynamic Time Warping (DTW) [41, 40, 10, 4], Hidden Markov Model (HMM) [54] and Conditional Random Fields (CRF) [53, 52], initially developed for speech recognition to decide boundaries of individual gestures. The second category localizes the starting and ending of the gestures through classification. Neverova et al. [30] built a binary classifier to distinguish the frames of the subject being on rest and performing actions. The third category is based on the position and motion of human hand. Peng et al. [31] design a temporal segmentation method based on the motion analysis of human hands. Chai et al. [5] use hand positions to realize the temporal segmentation based on the assumption that the subject puts the hand up when beginning a gesture and puts the hands down after performing one gesture. The last category is based on appearance. Upon the general assumption that the start and end frames of adjacent gestures are similar, correlation coefficients [27] and K-nearest neighbour algorithm with histogram of oriented gradient (HOG) [50] were used to identify the start and end frames of gestures. Jiang et al. [20] and Wang et al. [49] proposed a method based on quantity of move-

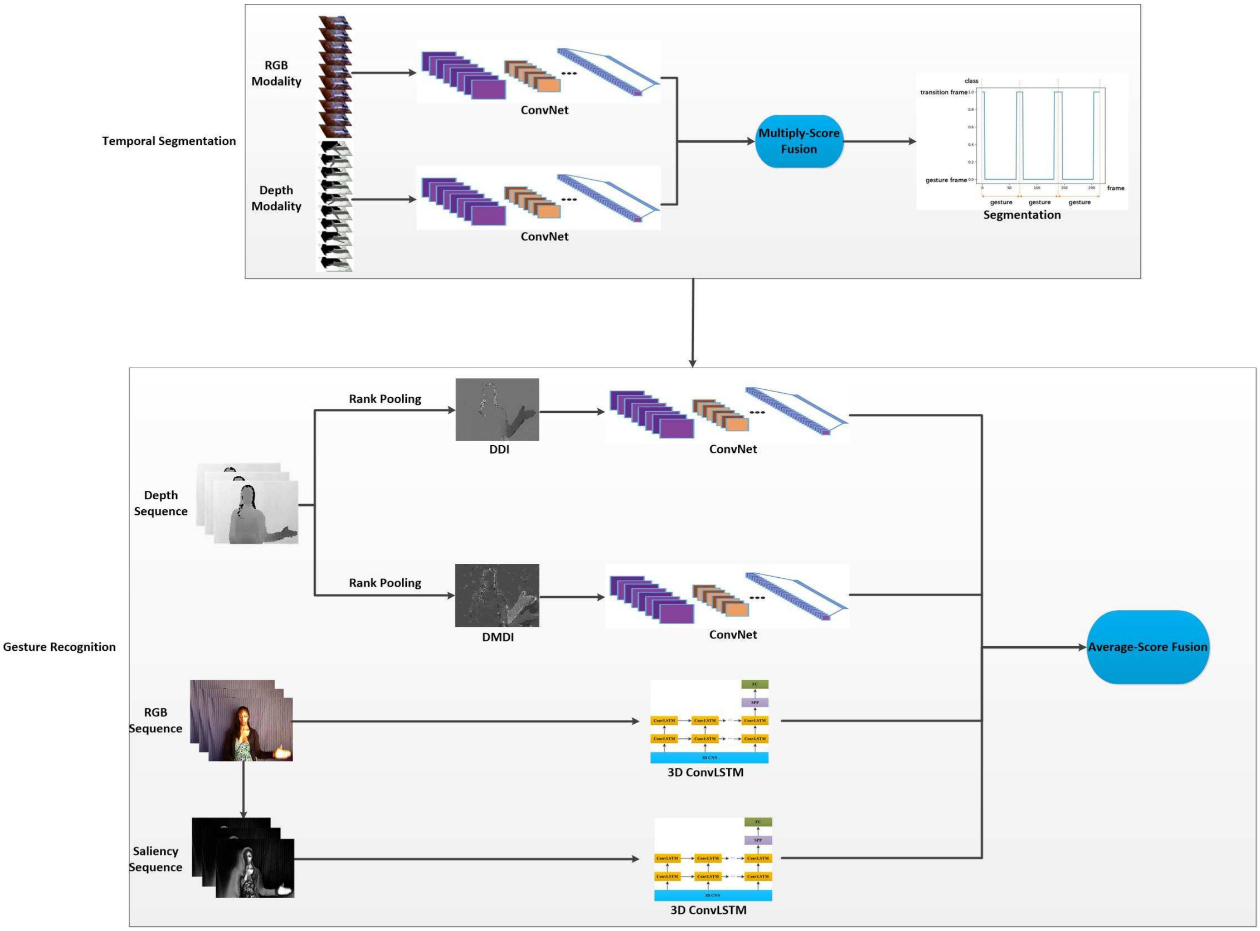

Figure 4. Overview of the proposed method.

ment (QOM). They first measure the QOM for each frame in a multi-gesture sequence and then threshold the quantity of movement to get candidate boundaries. Then, a sliding window is employed to refine the candidate boundaries to produce the final segmented gesture sequences in a multi-gesture sequence. This paper builds a binary classifier based on ConvNet to localize the transtional frames between gestures for temporal segmentation and its details will be presented in Section 3.1.

2.2. Gesture Recognition with Deep Learning

RGB and depth modalities have their own specific properties, and how to combine the strengths of both modalities with a deep learning approach is interesting. To address this problem, several methods have been proposed. These methods can be divided into three categories. The first one is CNN-based approach. Zhu et al. [57] fused RGB and depth in a pyramidal 3DCNN for gesture recognition. Duan et al. [9] proposed a convolutional two-stream consensus voting network (2SCVN) and a 3D depth-saliency ConvNet stream (3DDSN) for gesture recognition. Wang et al. [47] adopted scene flows to extract features which fuses the RGB and depth at feature level. The new representation based on CNN and named Scene Flow to Action Map (SFAM) was developed for gesture recognition.

The second approach is RNN-based. Pigou et al. [32] considered the depth as the fourth channel and ConvNet was adopted to frame-based appearance features. Temporal convolutions and RNN were combined to capture the temporal information. Li et al. [25] adopted 3DCNN to extract features separately from RGB and depth modalities, and used the concatenated for SVM classifier. Zhu et al. [58] presented a gesture recognition method combining 3DCNN and convolutional LSTM (convLSTM) based on depth and RGB modalities. Luo et al. [28] proposed to use a RNN-based encoder-decoder framework to learn a video representation for recognition by predicting a sequence of basic motions described as atomic 3D flows.

The last approach is other-structure-based approach. Shahroudy et al. [36] extracted hand-crafted features which are neither independent nor fully correlated from RGB and depth, and embedded the features into a space of factorized common and modality-specific components. Then they stacked layers of non-linear auto encoder-based component factorization to form a deep shared-specific analysis network.

All of these methods usually adopt the same algorithm on RGB and depth modalities. The proposed method instead uses different types of networks aiming to learn different types of features from different spatio-temporal modalities to improve the recognition accuracy.

3. Proposed Method

As shown in Figure 4, the proposed method consists of four phases: temporal segmentation, gesture recognition from depth modality, gesture recognition from RGB modality and score fusion of the outputs from the depth and RGB modalities for final gesture recognition. Given a continuous gesture sequence, the continuous gesture sequence is firstly segmented into isolated gesture sequences. On the one hand, Depth Dynamic Images and Depth Motion Dynamic Images are constructed from depth sequences and fed to the ConvNets. On the other hand, the RGB and saliency sequences are input to the 3D ConvLSTMs. The ConvNets for depth modality and 3D ConvLSTMs for RGB modality are designed to learn spatiotemporal features combining the strengths of RGB and depth modalities to improve the recognition.

3.1. Temporal Segmentation

As shown in Figure 1, a continuous gesture sequence is composed of gesture frames and transition frames. All frames in a gesture sequence can be classified into two classes. The transition frame is the boundary of two consecutive gestures. It can be treated as a two-class classification problem. To solve this classification problem, two stream ConvNets are adopted to classify each frame in RGB-D sequences. As shown in Figure 4, the RGB stream is trained from RGB modality and the depth stream performs from depth modality. Each stream is implemented using a ConvNet, softmax scores of which are combined by late fusion. Both stream ConvNets are image classification architecture, we can build upon the recent advances in large-scale image recognition method, and pre-train the network on a large image classification dataset, such as the ImageNet dataset [34]. The details are presented in setcion 4.2.1.

Given a continuous gesture sequence, this strategy allows us to assign each frame with a label "transitional frame" or "gesture frame". As shown in Figure 5, the beginning and the end of each gesture are typically transitional frames. In this paper, the middle frame of a continuous segment of transitional frames is defined as the final boundary between two gestures. The segmented gesture sequences will be input to the gesture recognition module.

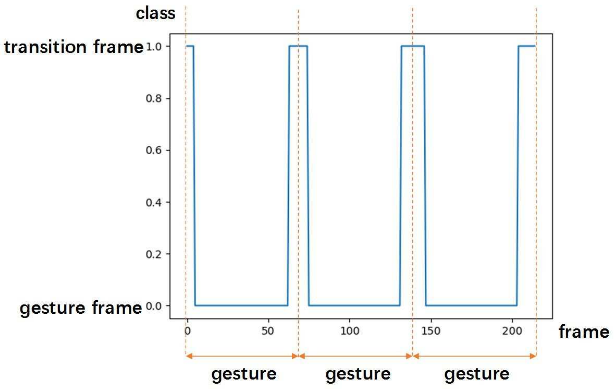

Figure 5. An example of the temporal segmentation. The sequence is segemented into three isolated gesture sequences, the middle point of continuous transitional frames is defined as the boundary of two gestures.

3.2. Gesture Recognition for Depth Modality

Firstly, four sets of dynamic images, Depth Dynamic Images (DDIs) and Depth Motion Dynamic Images (DMDIs) are constructed from an image sequence through bidirectional rank pooling. Each set of dynamic images 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

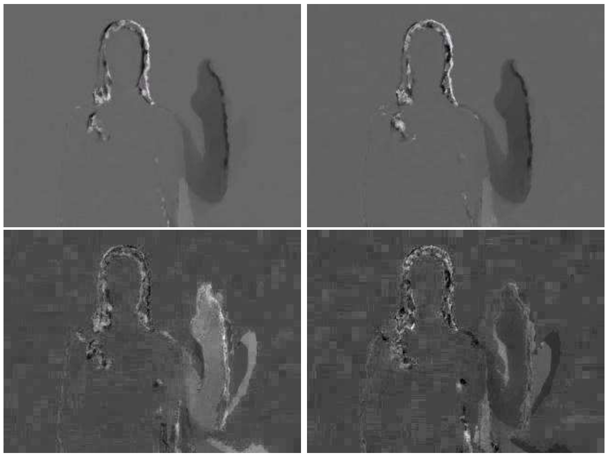

Figure 6. Samples of generated forward and backward DDIs and DMDIs for gesture Mudra1/Ardhapataka, the left images are dynamic images for forward, the right images are dynamic images. From up to bottom: DDIs and DMDIs.

from the work [2], the rank pooling is applied in a bidirectional way to convert one video sequence into two dynamic images. DDIs are constructed from depth sequence. Unlike DDIs, DMDIs are constructed from the absolute differences between consecutive frames through an entire depth sequence. Each dynamic image is fed into a ConvNet. The resulting dynamic images are illustrated in Figure 6. As shown, DDIs and DMDIs effectively capture the spatiotemporal information.

3.3. Gesture Recognition for RGB Modality

The 3D ConvLSTM network is described in detail by Zhu et al. [58]. As shown in Figure 7, a 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 [39] to learn short-term spatiotemporal features. Two-level ConvLSTM [51] 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) [17] is added on the top of ConvLSTM and connected to the full connected layer. Different from [58], both RGB sequence and its saliency sequence are input to the 3D ConvLSTM networks. The saliency sequence are extracted using the algorithm decribed in [1].

3.4. Score Fusion for Classification

Given a pair of RGB and depth video sequences, DDIs and DMDIs are generated from the depth sequence and fed into seperately trained ConvNets, and the RGB sequence and its saliency sequence are fed into the 3D ConvLSTM networks. Average-score fusion is used to fuse the classification output of all nextworks. 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 LAP ConGD Dataset [42] and evaluation protocols are described. The experimental results of the proposed methods on the dataset are reported. The final results was tested by the challenge organisers.

4.1. Datasets

The ChaLearn Gesture Dataset (CGD) was recorded by Microsoft Kinect sensor [56]. It includes color and depth video sequences provided by the sensor as it does not provide the human pose information. The ChaLearn LAP ConGD Dataset are derived from ChaLearn Gesture Dataset (CGD). 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 labels performed by 21 different individuals. The detailed information of the ChaLearn LAP ConGD dataset is shown in Tabel 1.

4.2. Network Training

4.2.1 Network Training for Temporal Segmentation

To training the ConvNets for temporal segmentation, eight frames around the bounary of two gestures were taken as training samples of the class “trasitional frames” and the rest frames were considered as “gesture frames”. The VGG-16 [38] was adopted. The ConvNets were fine-tuned from the pre-trained models on ILSVRC-2015 [34]. The network

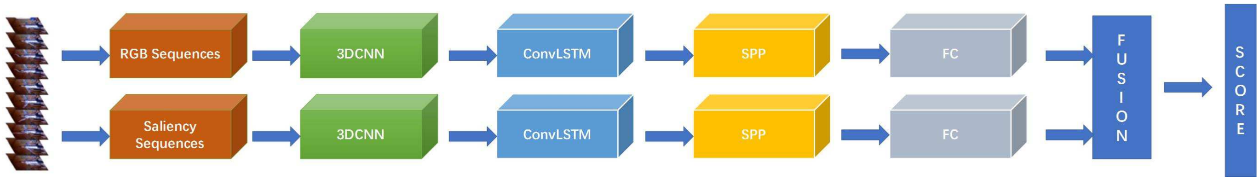

Figure 7. The framework of a 3D ConvLSTM. A RGB sequence and saliency sequence are fed into the 3D ConvLSTM.

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 1. Statistics of the ChaLearn LAP ConGD Dataset

weights were 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 64 samples was shuffled randomly. All the images were resized to 224 × 224. The learning rate for fine-tuning was set to 0.001, and then it was decreased according to a fixed schedule. This was kept the same for all training sets. The training underwent 90K iterations and the learning rate is dropped to its 0.1 every 40K iterations.

Methods	Mean Jaccard Index $\overline{J_S}$
MFSK [42]	0.0918
MFSK+DeepID [42]	0.0902
Wang et al. [49]	0.2403
Chai et al. [5]	0.2655
Camgoz et al. [3]	0.2809
Proposed Method	**0.5214**

Table 2. Comparison of the proposed method and other methods on the validation set of ConGD

4.2.2 Network Training for Depth Modality

After the construction of DDIs and DMDIs, four ConvNets were trained on the four channels individually. In this paper, the ResNet-50 [18] was adopted as the ConvNet model. We fine-tuned the ConvNets separately with pre-training models on ILSVRC-2015 [34]. The network weights were 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 used the rectification (RELU) activation function. At each iteration, a mini-batch of 16 samples was shuffled randomly. All the images were resized to 224 × 224. The learning rate for fine-tuning was set to 10^{-4}, and then it was decreased according to a fixed schedule, which was same for all training sets. The training underwent 90K iterations and the learning rate is dropped to its 0.96 every 40K iterations.

4.2.3 Network Training for RGB Modality

The 3D ConvLSTM was implemented based on the tensorflow and Tensorlayer platforms. RGB sequences and saliency sequences based networks were trained separately. We fine-tuned the networks on RGB modality based on the pre-training model on SKIG [26] and then fine-tuned the networks on saliency sequences based on the pre-training model of the RGB modality. Batch normalization makes training processes easier and faster. The initial learning rate was set to 0.1 and dropped to its $\frac{1}{10}$ every 15K iterations. The weight decay was initialized to 0.004 and decreased to 0.00004 after 40K iterations. At most 60K iterations are needed for training. At each iteration, the batch-size was 13, the temporal length of each clip was 32 frames, and the crop size for each image was 112.

4.3. Evaluation on ChaLearn LAP ConGD Dataset

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.

$$J_{s,i} = \frac{G_{s,i} \bigcap P_{s,i}}{G_{s,i} \bigcup P_{s,i}} \tag{4}$$

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.

Rank by test set	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	**0.5957**	0.5950
3	PaFiFA	0.3646	0.3744
4	Deepgesture	0.3190	0.3164
–	Proposed Method	0.5214	0.5307

Table 3. Performance comparison with other teams in ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge

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.

$$J_s = \frac{1}{l_s} \sum_{i=1}^{L} J_{s,i} \qquad (5)$$

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.

$$\overline{J_S} = \frac{1}{n} \sum_{j=1}^{n} J_{s_j} \qquad (6)$$

Tabel 2 compares the performance of the proposed method and that of exiting methods on validation set. It can be seen that our proposed method achieve the state-of-the-art results compared with existing methods.

Table 3 compares the performance of the proposed method and that of ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge [21]. Our mean Jaccard Index is 0.5307 in test set. It can be seen that our method is among the top performance.

5. Conclusion

The paper presents an effective method for large-scale multimodal gesture segmentation and recognition. The video sequences are first segmented into isolated gesture sequences by classifying the frames into gesture frames and transition frames. For each segemented gesture sequence, our proposed method explores the effective spatiotemporal information based ConvNets for depth modality and 3D ConvLSTMs for RGB modality. Experimental results on ChaLearn LAP ConGD Dataset verified the effectiveness of our proposed method.

Acknowledgment

Huogen Wang and Pichao Wang gratefully acknowledge the 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] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE, 2009. 2, 5

[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. 2, 5

[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] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici. Gesture recognition using skeleton data with weighted dynamic time warping. In *VISAPP (1)*, pages 620–625, 2013. 2

[5] 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. 1, 2, 6

[6] J. Y. Chang. Nonparametric gesture labeling from multi-modal data. In *ECCV Workshops (1)*, pages 503–517, 2014.

[7] 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

[8] 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

[9] 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. 3

[10] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan. Principal motion components for one-shot gesture recognition. *Pattern Analysis and Applications*, 20(1):167–182, 2017. 2

[11] 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.

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

[13] 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. 2

[14] 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. 2

[15] 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. 2

[16] B. Fernando and S. Gould. Learning end-to-end video classification with rank-pooling. In *International Conference on Machine Learning*, pages 1187–1196, 2016. 2

[17] 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. 5

[18] 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. 6

[19] 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

[20] 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. 1, 2

[21] 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. 7

[22] 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.

[23] 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.

[24] 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

[25] 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. 3

[26] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, volume 4, page 8, 2013. 6

[27] Y. M. Lui. Human gesture recognition on product manifolds. *Journal of Machine Learning Research*, 13(Nov):3297–3321, 2012. 2

[28] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. *arXiv preprint arXiv:1701.01821*, 2017. 3

[29] 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.

[30] 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, 2

[31] X. Peng, L. Wang, Z. Cai, and Y. Qiao. Action and gesture temporal spotting with super vector representation. In *ECCV Workshops (1)*, pages 518–527, 2014. 1, 2

[32] 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. 3

[33] 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

[34] 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. 4, 5, 6

[35] M. Ryoo and J. Aggarwal. Stochastic representation and recognition of high-level group activities. *International journal of computer Vision*, 93(2):183–200, 2011.

[36] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[37] 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

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[39] 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, 5

[40] J. Wan, V. Athitsos, P. Jangyodsuk, H. J. Escalante, Q. Ruan, and I. Guyon. Csmmi: Class-specific maximization of mutual information for action and gesture recognition. *IEEE Transactions on Image Processing*, 23(7):3152–3165, 2014. 2

[41] J. Wan, Q. Ruan, W. Li, and S. Deng. One-shot learning gesture recognition from rgb-d data using bag of features. *The Journal of Machine Learning Research*, 14(1):2549–2582, 2013. 2

[42] 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

[43] 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.

[44] 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

[45] 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

[46] 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

[47] 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, 2, 3

[48] 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, 2

[49] 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

[50] D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 7–12. IEEE, 2012. 2

[51] 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. 2, 5

[52] H.-D. Yang and S.-W. Lee. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters*, 34(16):2051–2056, 2013. 2

[53] H.-D. Yang, S. Sclaroff, and S.-W. Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1264–1277, 2009. 2

[54] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286. ACM, 2011. 2

[55] 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.

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

[57] 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. 3

[58] G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3d convolution and convolutional lstm. *IEEE Access*, 2017. 1, 3, 5