GL-PAM RGB-D GESTURE RECOGNITION

Benchao Li [‡] *Wanhua Li* [†] *Yongyi Tang* [‡] *Jian-Fang Hu*[†] *Wei-Shi Zheng*[†*]

[‡]School of Electronics and Information Engineering, Sun Yat-sen University, Guangzhou, China
[†]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
{libch3, liwh9, tangyy8}@mail2.sysu.edu.cn, {hujf5, zhwshi}@mail.sysu.edu.cn

ABSTRACT

The existing approaches for RGB-D gesture recognition mainly developed their systems based on the global features extracted from full sequences, which makes them unreliable for capturing some important movements. In this paper, we propose to combine the global and local context information extracted from posture, appearance, and motion sequences. Our experimental results on a large scale RGB-D gesture dataset show that the proposed global and local contexts can complement well with each other for efficiently characterizing gestures, and thus achieve the 2nd place in the ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge (Round 2).

Index Terms— RGB-D gestures recognition, global-local features

1. INTRODUCTION

Recognizing human gestures from RGB-D sequences has received an increasing amount of attention in the research community, for its potential applications in HCI, robotics and sign language comprehension etc. Gestures, as a special type of human action, are meaningful body motions involving physical movements of the fingers, hands and arms [1].

Over the last two decades, the vision community has done many works for recognizing RGB-D gestures. Many prior works focus on developing their systems based on the hand-crafted features like HOG, HOF [2] and SURF [3]. For instance, Wan *et al.* [4] collect a spatiotemporal feature named MFSK while Konecny *et al.* [5] extract static HOG as appearance descriptor and dynamic HOF as motion descriptor for the RGB-D gesture recognition. Wan *et al.* [6] represent each video as a bag of 3D MOSIFT features, and use a nearest neighbor classifier to predict gestures.

Recently, with the success of deep neural networks in various challenges [7, 8], many video-based RGB-D gesture recognition methods [9–11] have been proposed to learn a robust video representation. Molchanov *et al.* [12] combine information from multiple spatial scales of hands and deploy their framework based on 3D convolutional networks. Nishida *et al.* [13] propose a multi-stream recurrent neural network for gesture recognition which can be trained end-to-end without any domain-specific engineering. Neverova et al [14] proposed a multi-modal framework that operates at three temporal scales corresponding to dynamics postures for gesture localization.

However, almost all the existing methods intend to build their recognition system based on the context depicted in the full sequences, which may ignore some useful local gesture context. Hands and arms have been widely used compared with other body parts for gesturing partly due to the fact that it is a natural form of medium for human communication [15]. Indeed, the local regions around the hands and arms of actors could contain some informative cues for gesture recognition [1]. Hence, we aim to extract features from the full sequences and the local regions across various modalities simultaneously, obtaining global sequences and local sequences of those various modalities, in order to characterize gestures from a more comprehensive view.

In addition to RGB-D data, skeleton data is also found to be useful for characterizing actions/gestures [16–18], owing to its high level representation and robustness to variations of appearances, and surrounding distractions. It is reasonable to use the body posture to improve the gesture representation. Hence, based on the skeleton modality we extracted global and local features form RGB, depth and skeleton sequences.

Overall, in order to capture the global and local information for gesture recognition at the same time, we propose a new Global-Local Posture, Appearance and Motion (GL-PAM) framework for RGB-D gesture recognition. The main contributions are summarized as follows: 1) Global-local sequences for appearance, motion and posture modalities are proposed. Experiments show that global and local sequences can complement with each other sufficiently; 2) To our best knowledge, we first introduced skeleton modal for Chalearn IsoGD dataset; 3) Our method ranked the second in the ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge (Round 2) [19], which achieved the accuracy of 67.02% on the test subset.

*Corresponding author.

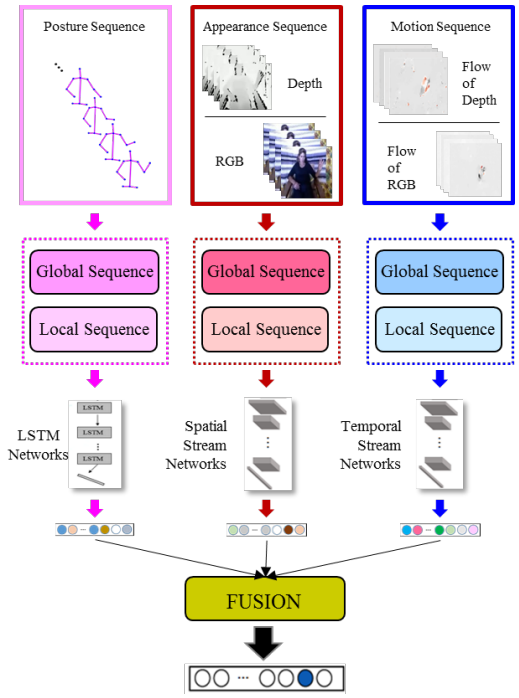

Fig. 1. The overview of the proposed framework (best viewed in color).

2. GL-PAM FRAMEWORK

The overall framework of our GL-PAM model is presented in Figure 1. In the framework, we aim to extract a set of discriminative features representation encoding the global and local posture, appearance, and motion information depicted in RGB-D and skeleton gesture sequences. Specifically, the global and local appearance cues are mined from RGB and depth sequences by two spatial convolutional networks to characterize the shape and texture messages for a gesture. Also, two different temporal networks are used to extract the local and global motion information from optical flow data. Two LSTM architectures are utilized to characterize the dynamic global and local posture cues. Finally, the (score) outputs of the above networks (spatial and temporal networks and LSTM networks) are fused together for recognizing gestures. In the following, we will introduce our framework in detail.

2.1. Global-Local Sequences Construction

Our method needs to extract global and local features from RGB-D and skeleton sequences. We empirically find that the local regions near the hands (left and right) of actors are more informative for distinguishing human gestures than other regions. Thus, we use the regions around the hands of actors to form our local sequences. The overall procedure of constructing local posture, appearance and motion data is illustrated in Figure 2. In the following, we describe how to construct the global and local sequences in detail.

2.1.1. Global and Local Posture Sequences

We extract pose dynamics from skeleton sequences for the gesture modeling. We apply regional multi-person pose estimation(RMPE) [20] framework to extract the trajectories of the key joints for each gesture video. In detail, we only consider the joints of upper part of the body, which contains most of the gesture information. As shown in Figure 2, RMPE presents 12 joints indicating the locations of the upper body part of each actor. These joints are considered as the global posture data for each frame. Following the implementations in [21], we compute the relative positions between each pair of joints and concatenate them together, obtaining a vector for each frame of the video sequence, which depicts the global information.

Complementarily, we construct a local posture sequence for the gesture representation. Considering that the hands could contain more gesture information than other body parts (e.g. head and foot etc.), we extract the joints of hands, elbows and shoulders to form a local posture sequence, and calculate the relative positions of the local posture sequences. Finally, the relative positions extracted form both the global and local posture sequences are feed into the LSTM networks to predict the results, respectively.

2.1.2. Global-Local Appearance Sequences

The appearance, depicted in the RGB and depth sequences, usually contains some characteristic appearance information (e.g. shape and texture etc.) of gesture. Some gestures, such as gesture of number *"1"*, can be identified from the appearance. Specifically, RGB video can deliver some useful color information, while the depth channel is insensitive to illumination variations and reliable for capturing body silhouette. Thus, the depth and RGB channels are very important for characterizing appearance information and they complement each other. Here, the original RGB and depth sequences are treated as the global appearance sequences.

However, only using the global context information is not enough for some complex gestures, such as gesture of Music Notes *do*. Therefore, we construct the local appearance sequences to capture the fine-grained cues. Specifically, we extract six local image patches around the joints of *Left Shoulder*, *Left Elbow*, *Left Wrist*, *Right Shoulder*, *Right Elbow* and *Right Wrist*. In detail, as for shoulders and elbows, we extract four 33×33-sized image patches, and two image patches with size of 65×65 for hands. Here, the patches corresponding to the hands are larger than that of elbows and shoulders, this is because that the regions around the hands typically have larger variations thus could depict more informative gesture information. The six patches are concatenated together to form a new image, which is illustrated in Figure 2. Both the global and local appearance (RGB and depth) sequences are fed into spatial convolutional networks to obtain appearance feature representation.

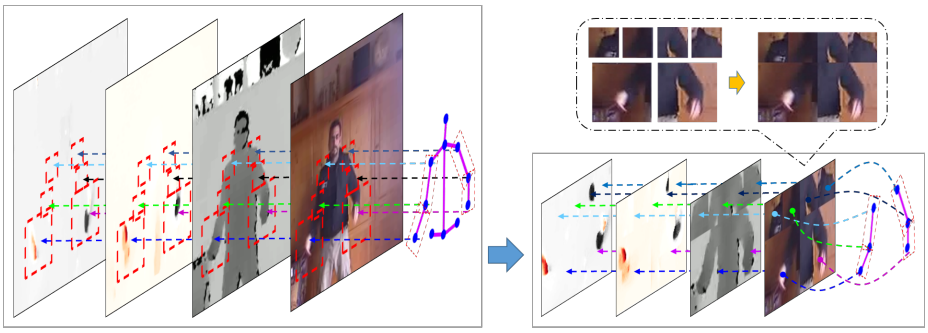

Fig. 2. The process of generating local sequences for each global sequences. For each global sequence of posture, appearance or motion, the regions (or joints) corresponding to hands and arms are extracted to produce our local sequences. (best viewed in color).

2.1.3. Global-Local Motion Sequences

Using postures and appearance is not sufficient for distinguishing gestures with long-term motions. For example, *Door Closed* and *Door Opened*. Here, we consider exploring the long-term motions of RGB-D sequences for gesture recognition. For capturing motion cues, we extract optical flows from both RGB and depth sequences using TV-L1 [22].

Similar to the construction of local appearance sequences, we extract the local patches around the hands and arms of the actor of interest, which could deliver most of the movements occurred in gestures. The concatenation of these patches along the temporal dimension forms our local motion sequences. Some example samples can be found in Figure 2. All the global and local motion sequences are fed into temporal networks. Thus, a total of four different temporal networks are trained for capturing the long-term motions from global and local perspectives.

2.2. Score Fusion

To this end, we can obtain ten score features, which encode the gesture context from different perspectives. Here, we fuse these features together to achieve a robust gesture recognition. Existing multi-modal feature fusion approaches can be divided into two categories: early fusion and late fusion. Here, we adopt the late fusion strategy, i.e., the score outputs obtained for each individual global and local modal features of posture, appearance and motion are integrated together. The fusion procedure is formulated as following,

$$y = \frac{1}{n}\sum^{n} y_i, \tag{1}$$

where n is the count of sequences and y_i is the predicted result of one certain sequence.

3. EXPERIMENTS

3.1. Dataset

We evaluate our methods on a large-scale gesture recognition dataset, Chalearn IsoGD. To the best of our acknowledge, this is the largest set that can be available for the evaluation of recognizing gestures. Overall, it has 47933 RGB-D videos from 249 gesture classes performed by 21 different individuals. For evaluation, these videos are divided into three subsets, where the videos in the training and validation/testing subsets are performed by different subjects. Since the labels of the test subset was not available for evaluation, all the results are calculated on the validation subset.

3.2. Implementation Details

Our networks were implemented with TensorFlow and three TITAN X GPUs are used to train the model. Uniform sampling with temporal jitter is utilized for temporal augmentation as described in [10]. We uniformly sample 16 frames from each video. For training the convolutional networks, all the sampled frames are resized to 256 $\times$ 256, where 224 $\times$ 224-sized regions are randomly cropped and the 16 frames are stacked in channel. We feed the relative position vectors of the sampled frames into it LSTM to compute discriminative features among posture information.

LSTM Networks. For LSTM networks, we use the architecture with two LSTM layers followed by a ReLU function. The number of neurons in each LSTM layer is 128. The networks are trained using stochastic gradient descent (SGD) algorithm, where the batch size is set as 128. The learning rate is reduced by 0.5 when the loss does not decrease in a reasonable condition, with a initialization of 1. The training process stops after 100 epochs.

Convolutional Networks. We use the VGG-16 as the backend of the spatial and temporal networks. The parameters were learned using mini-batch stochastic gradient descent with momentum, where the momentum rate is set as 0.9. The

Modal	Global(%)	Local(%)
Skeleton	31.24	30.24
Depth	42.31	25.38
Depth-OF	32.7	30.7
RGB	44.83	26.09
RGB-OF	45.71	40.78

Table 1. Results for Single Modal Feature.

Modal	No-Softmax (%)	Softmax (%)
Skeleton	34.16	33.18
Depth	43.93	43.29
RGB	47.29	46.12
Depth-OF	41.65	40.68
RGB-OF	50.02	49.36
All modalities	59.70	58.72

Table 2. Results of Fusing Global and Local Features.

Team	Recognition Rate (%)
ASU [24]	67.71
SYSU_ISEE(ours)	**67.02**
Lostoy	65.97
AMRL	65.59
XDETVP	60.47
baseline	67.26

Table 3. Comparison the performance of our method with other team.

weight for decay is set to 0.0005. The batch size is set to 32. The learning rate is initialized as 0.001 and it will be reduced by 2 whenever the validation error stops to decrease. We also set the rate of dropout as 0.5 for the first two fully-connected layers.

Testing. To obtain a robust recognition result, similar to other work [7], we randomly sampled 16 clips from the entire RGB-D sequences and then obtain a recognition result for each clip. The results of all the 16 clips are finally averaged.

3.3. Experimental Results

3.3.1. Results for Single Modal Feature

Table 1 presents the results using a single modal feature. We use "RGB-OF" to indicate the methods using optical flow computed from RGB sequences, and "Depth-OF" stands for optical flow from depth sequences. "Global" ("Local") indicates that the corresponding features are extracted from the global (local) gesture sequences. From the table, we can find that motion features extracted from global sequences achieves the highest accuracy among all these features, which demonstrates the efficacy of the proposed motion feature for characterizing gestures. After scanning the most gesture samples's raw depth sequences, we find that the depth sequences of the dataset have some noise, which causes the inaccuracy in the computation of depth's optical flow. Therefore, the recognition result for depth motion is poor.

We can also observe that posture features gets the lowest result, owing to the fact that skeleton carries much less information than appearance and motion. Actually, the skeleton loses the very grained hand pose discriminative cues for gesture recognition. For example, the gesture for number "1" is only different in fingers compared to the gesture for number "2", but joints of the actor has no difference at all.

3.3.2. Results of Fusing Global and Local Features

The recognition accuracy results of fusing the global and local features are presented in Table 2, where each column indicates the result for the corresponding modal. "All modalities" stands for the results of fusing all the global and local features of posture, appearance and motion. "No-softmax" means that the fused features are the scores un-normalized by the softmax layer. In contrast, "softmax" means the features are probability scores normalized by a softmax operator. According to Table 1 and Table 2, although the recognition accuracy of local sequences are not very satisfying compared with the corresponding global sequences, they complement well with each other respectively. We can also note that fusing un-normalized scores can produce a better recognition accuracy, which has a similar observation in [23]. In the following experiments, we use the un-normalized scores as our feature for fusion.

3.3.3. Comparison with the State-of-the-art

Here, we fuse all the features together and obtain a second place in the challenge of ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge 2017 (Round 2). More detailed results are presented in Table 3, as well as other competitors. As shown, our method obtained an accuracy of 67.02% on test set, which is very close to the best result achieved in this challenge. The first place team ASU [24] only obtain their sophisticated features only based on the global sequences, which may lose the complementary cues in local sequences.

4. CONCLUSION

In this paper, we presented a GL-PAM framework for RGB-D gesture recognition. Both global and local appearance, motion and postures sequences are proposed to represent RGB-D gestures from a comprehensive perspective. Experimental results show that the proposed global and local features complement well each other for characterizing gestures.

Acknowledgements

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(61522115, 61661130157, 61472456, U1611461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), and the Royal Society Newton Advanced Fellowship (NA150459).

5. REFERENCES

[1] J. Sonkusare, N.B. Chopade, R. Sor, and S.L. Tade, "A review on hand gesture recognition system," in *Proceedings of the IEEE International Conference on CCCA*, 2015, pp. 790–794.

[2] I. Laptev and T Lindeberg, "On space-time interest points," in *IJCV*, 2005.

[3] G. Willems, T. Tuytelaars, and L.V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008.

[4] J.Wan, G.Guo, and S.Z.Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE Transactions on PAMI*, pp. 1626–1639, 2016.

[5] J. Konecny and M. Hagara, "One-shot-learning gesture recognition using hog-hof," in *JMLR*, 2014, vol. 15, pp. 2513–2532.

[6] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from rgb-d data using bag of features," in *Journal of Machine Learning Research*, 2013, pp. 2549–2582.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[9] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *Proceeding of ICPRW*, 2016.

[10] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen, "Large-scale isolated gesture recognition using pyramidal 3d convolutional networks," in *Proceeding of ICPRW*, 2016.

[11] G. Zhu, L. Zhang, P. Shen, and J Song, "Multimodal gesture recognition using 3-d convolution and convolutional lstm," in *IEEE Access*, 2017, pp. 4517–4524.

[12] P. Molchanov, S. Gupta, K. Kim, and J Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *CVPR Workshop*, 2015.

[13] Noriki Nishida and Hideki Nakayama, "Multimodal gesture recognition using multi-stream recurrent neural network," in *PSIVT*, 2015.

[14] Natalia Neverova, Christian Wolf, Taylor, and et al., "Multi-scale deep learning for gesture detection and localization," in *ECCV 2014 Workshops*, 2015.

[15] S.S. Rautaray and A Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," in *Artificial Intelligence Review*, 2012, pp. 1–54.

[16] J.K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," in *Pattern Recognition Letters*, 2014, vol. 48, pp. 70–80.

[17] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeleton data: A review," in *arXiv preprint arXiv:1601.01006*, 2016.

[18] L.L. Presti and M.La Cascia, "3d skeleton-based human action classification: a survey," 2016, vol. 53, pp. 130–147.

[19] Wan Jun, Sergio Escalera, and et al., "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *ICCV Workshops*, 2017.

[20] Haoshu Fang, Shuqin Xie, Yuwing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *arXiv:1612.00137*.

[21] L. Wang, Y. Liu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition.," in *IEEE Transactions on PAMI*, 2016.

[22] J.S. Prez, E. Meinhardt-Llopis, and G. Facciolo, "Tv-l1 optical flow estimation," in *IPOL Journal*, 2013, pp. 137–150.

[23] Jiali. Duan, Jun. Wan, Shuai. Zhou, Xiaoyuan. Guo, and Stan Li, "A unified framework for multi-modal isolated gesture recognition," in *In ACM TOMM*, 2017.

[24] Miao Qiguang, Li Yunan, Ouyang Wanli, and et al., "Multimodal gesture recognition based on the resc3d network," in *ICCV*, 2017.