Two Streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition

Xiujuan Chai, Zhipeng Liu, Fang Yin, Zhuang Liu, Xilin Chen
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
University of Chinese Academy of Sciences, Beijing, 100049, China
Cooperative Medianet Innovation Center, China

{chaixiujuan, xlchen}@ict.ac.cn, {zhipeng.liu, fang.yin, zhuang.liu}@vipl.ict.ac.cn

***Abstract*—In this paper, we tackle the continuous gesture recognition problem with a two streams Recurrent Neural Networks (2S-RNN) for the RGB-D data input. In our framework, the spotting-recognition strategy is used, that means the continuous gestures are first segmented into separated gestures, and then each isolated gesture is recognized by using the 2S-RNN. Concretely, the gesture segmentation is based on the accurate hand positions provided by the hand detector trained from Faster R-CNN. While in the recognition module, 2S-RNN is designed to efficiently fuse multi-modal features, i.e. the RGB and depth channels. The experimental results on both the validation and test sets of the Continuous Gesture Dataset (ConGD) have shown promising performance of the proposed framework. We ranked 1st in the ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge with the mean Jaccard Index of 0.286915.**

***Index Terms*—continuous gesture recognition; recurrent neural networks; spotting-recognition; multi-modal features**

I. Introduction

Vision based gesture recognition has become a popular research topic because of its wide applications, such as human-computer interaction and sign language translation and so on.

Recently, some new data capture sensors have promoted the development of gesture recognition. Microsoft Kinect is a motion input device which can provide RGB and depth image simultaneously. The multimodal data complement each other effectively and form more powerful gesture representation than single modality. Therefore, much more work on gesture or sign language recognition is explored on such RGB-D data [1] [2]. To evaluate the gesture recognition algorithms objectively and fairly, some research groups released several gesture databases [3]–[8]. Among them, ChaLearn LAP RGB-D Continuous Gesture Dataset (ConGD) is a large scale dataset with clear training and testing protocols and a challenge is organized based on it.

In this paper, we will describe our spotting-recognition framework used in the challenge. In our method, the continuous gestures are firstly segmented into isolated gestures based on the assumption that the subject puts the hands down after performing each gesture. Then the segmented isolated gestures are recognized with our proposed 2S-RNN. In our recognition algorithm, the two paralleled simple Recurrent Neural Network (SRNN) layers, which take the feature from different modalities as input, are effectively fused by a fusion layer. A followed Long Short-Term Memory (LSTM) layer aims to model the contextual information of the temporal gesture sequences.

The remainder of this paper is organized as follows: Section II briefly reviews the related work on continuous gesture recognition. Section III gives the details of the proposed continuous gesture recognition framework. Experimental results and discussion are presented in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

In this section, we will briefly review the previous work on continuous gesture recognition.

Different from isolated gesture recognition, in the continuous gesture recognition task, the boundaries of the gestures need to be detected, namely, gesture segmentation. Most methods dealt with gesture segmentation and recognition in separate procedures [9], [10]. Others realized segmentation and recognition simultaneously, such as the methods by dynamic programming [11] [12] and Viterbi decoding [2] [13]. Hoai et al. [12] used a discriminative temporal extension of the spatial bag-of-words model and the inference over the segments was done efficiently with dynamic programming. Pitsikalis et al. [13] extended HMM with multiple hypotheses rescoring fusion scheme. The sequences were segmented based on the best hypothesis.

Dynamic gesture recognition is one kind of temporal sequence recognition problem, which is similar to speech recognition , but more challenging. In the early time, gesture recognition borrowed some models from speech recognition, such as Dynamic Time Warping (DTW) [11], Hidden Markov Model (HMM) [2] and Conditional Random Fields (CRF) [14] [15]. Besides these traditional models, there are some other methods for gesture recognition. Pfister et al. [16] proposed a method to boost the performance of one-shot learning by using weakly supervised learning. Ong et al. [17] used Hierarchical Sequential Interval Pattern Trees (HSP-Tree) to realize segmentation and recognition simultaneously. On a continuous sign language dataset with vocabulary size of 40, the accuracies of signer-dependent and signer-independent

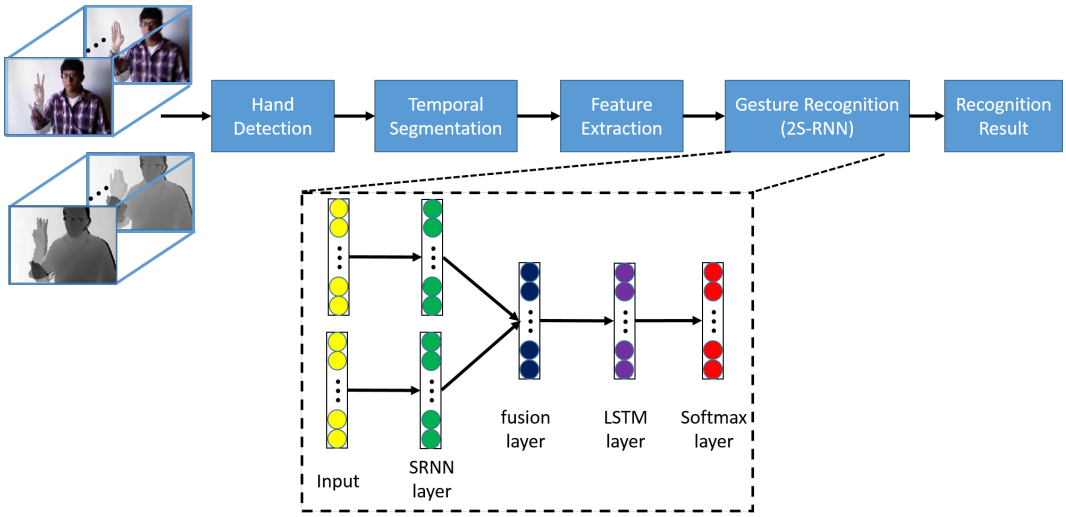


Fig. 1: The main idea of our continuous gesture recognition framework.

evaluations are 71% and 54% respectively. Recently, with the development of deep learning, some neural network related methods are also applied on gesture recognition. Tran et al. [18] extracted the spatiotemporal features with deep 3D Convolutional Neural Networks (3D-CNN) and used SVM classifier for classification. Molchanov et al. [19] proposed Recurrent 3D Convolutional Neural Networks (R3DCNN) which performed detection and recognition of gestures from multi-modal data simultaneous and supported online recognition with zero or negative lag. On their multi-modal dataset with 25 gesture types, the best accuracy is 83.8%. Also considering the multimodal information, this paper presents a novel two streams RNN architecture to tackle the continuous gesture recognition problem.

III. Methodology

In this section, we will describe the spotting-recognition framework for large continuous gesture recognition. Figure 1 shows the main idea of our whole framework.

For all frames of a given RGB-D continuous gesture video, first, the hands are located with a pre-trained hand detector and the face is also detected as a reference [20]. Considering on the constraint of the hand positions, the video sequence is segmented into several isolated gestures. With the hand trajectory and hand posture features extracted from each data modality, by using the proposed 2S-RNN, each isolated gesture is classified into its most likely category.

A. Temporal Segmentation

We use hand positions to realize the temporal segmentation based on the assumption that the subject puts the hands up when beginning to sign a gesture and puts the hands down after performing one gesture. Thus the hand detection is very crucial for our temporal segmentation, and also for the subsequent recognition module. Previously, hand regions are usually detected by their color [21] or multiple proposals [22]. These skin-color-based methods have certain capabilities to detect hands, but they are sensitive to illumination and background. Recently, a series of region-based convolutional neural networks (R-CNNs) [23] have been proposed for object detection, which give us a new aspect for hand detection. In this paper, we treat hand detection as a two-class (hand and non-hand) classification problem.

We use the latest incarnation of region-based CNNs, Faster R-CNN [24], to build our hand detection model. Faster R-CNN is an end-to-end network that takes an image (of any size) as input and outputs a set of rectangular detected objects, each with an objectness score. The whole network has two main parts as shown in Fig.2. The first part is Region Proposal Network (RPN) which can generate high-quality region proposals. RPN is a kind of fully-convolutional network (FCN) [25] and it can be trained end-to-end specifically for the task to generate detection proposals. The second part is an object detection network, called Fast R-CNN [26]. Fast R-CNN is a single-stage network that takes a set of object proposals as input. Each proposal is pooled into a fixed size feature map through RoI (Region of Interest) pooling and then mapped to a feature vector by fully connected layers (FCs). Next, the feature vector is sent to two sibling layers, then the class probabilities and per-class bounding-box (bbox) regression offsets are output. The detail of Faster R-CNN can be referred to [24]. Figure 3 shows an example of the temporal segmentation result for a continuous gesture sequence.

B. Two Streams RNN

In the continuous gesture recognition problem, it is important to explore how to integrate the information from RGB and depth modalities, which are intrinsically complementary to each other. Therefore, as shown in Fig. 1, a two streams RNN architecture is proposed in order to make full use of the two channels' features. In consideration of the number of parameters, the deeper the network, the more over-fit it will be. While the shallow network with fewer parameters is insufficient to the temporal feature representation. In order to

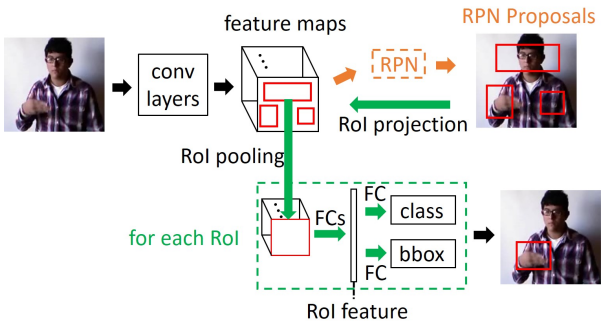


Fig. 2: The flow chart of Faster R-CNN algorithm. The figure shows two part of Faster R-CNN. The orange part is a Region Proposal Network which can generates high quality region proposals used by Fast RCNN. The green part is Fast RCNN architecture for each region of interest (RoI).

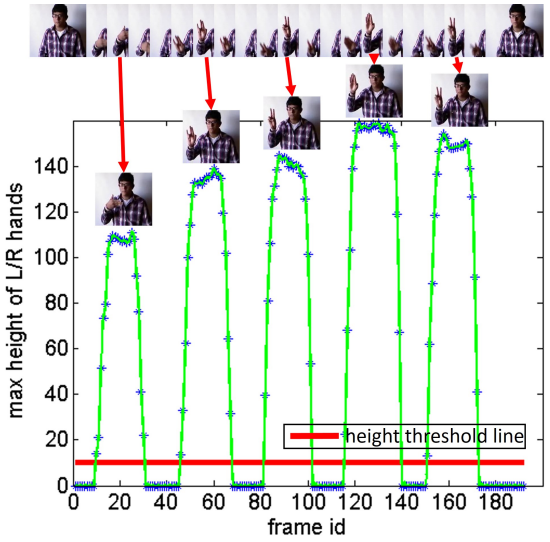


Fig. 3: An example of the temporal segmentation result for a continuous gesture sequence.

get a trade-off between the parameters and the representation ability, we propose a 2S-RNN architecture in this paper.

The key point of the RNNs is the recurrent connection which has the ability to memorize previous inputs to persist in the network's internal state, and therefore influences the network output [27]. With the memory function, RNNs can model the contextual information of a temporal sequence. Different memory neurons have different functions. In our method, There are two neurons, simple recurrent neural network (SRNN) and LSTM, which are illustrated in Fig. 4. In SRNN, given an input sequence whose length is T, $X = \left(x^0, \ldots, x^{T-1}\right)$, the hidden states of a recurrent layer $H = \left(h^0, \ldots, h^{T-1}\right)$ and the output $Y = \left(y^0, \ldots, y^{T-1}\right)$ can be derived as follows [27].

$$h^t = \theta\left(W_{xh}x^t + W_{hh}h^{t-1} + b_h\right) \tag{1}$$

$$y^t = O\left(W_{ho}h^t + b_o\right) \tag{2}$$

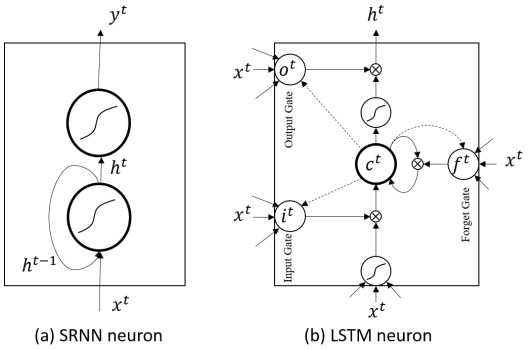


Fig. 4: The principle of SRNN and LSTM neuron

where t belongs to $(0, 1, \cdots, T-1)$, b_h and b_o are bias vectors, W_{xh}, W_{hh}, W_{ho} denote the connection weights, $\theta(\cdot)$ and $O(\cdot)$ are activation functions in the hidden layer and output layer. The SRNN neuron is very simple and has few parameters. However, it has the disadvantages of the vanishing gradient and error blowing up [28]. In order to solve these problems, one LSTM layer is added in our model. In LSTM, the activations of the memory cell and three gates are given as follows:

$$i^t = \sigma\left(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i\right) \tag{3}$$

$$f^t = \sigma\left(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f\right) \tag{4}$$

$$c^t = f^t c^{t-1} + i^t \tanh\left(W_{xc}x^t + W_{hc}h^{t-1} + b_c\right) \tag{5}$$

$$o^t = \sigma\left(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^t + b_o\right) \tag{6}$$

$$h^t = o^t \tanh\left(c^t\right) \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, all the vectors b denote bias vectors and all the matrices W are the connection weights.

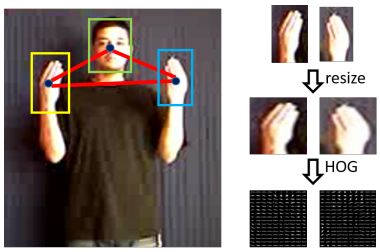
The proposed 2S-RNN consists of four layers, i.e., SRNN layer, fusion layer, LSTM layer and softmax layer. In the first SRNN layer, the RGB and depth features are fed into two paralleled SRNNs. Then the two SRNNs are combined in the fusion layer. In the following step, LSTM neuron is used to model the context information of each gesture and the final is the softmax layer to get the class label with the corresponding probability.

C. Training

To well train the networks, a training set Ω is prepared first, which contains plenty of gesture samples and the corresponding class labels. For the convenience of writing, RGB is denoted as "R" and depth is denoted as "D". For the j-th sample with T frames, the features for RGB and depth channels are denoted as $X^1_{Rj} = \left(x^{1(0)}_{Rj}, \ldots, x^{1(T-1)}_{Rj}\right)$ and $X^1_{Dj} = \left(x^{1(0)}_{Dj}, \ldots, x^{1(T-1)}_{Dj}\right)$ respectively, which are taken as the input of the first SRNN layer. Via Eq.1 and Eq.2, the output can be obtained, $Y^1_{Rj} = \left(y^{1(0)}_{Rj}, \ldots, y^{1(T-1)}_{Rj}\right)$ and $Y^1_{Dj} = \left(y^{1(0)}_{Dj}, \ldots, y^{1(T-1)}_{Dj}\right)$. For the fusion layer, the newly

TABLE I: Information of the ChaLearn LAP ConGD Dataset

Sets	# of Labels	# of Gestures	# of RGB Videos	# of Depth Videos	# of Performers	Label Provided	Temporal Segmention Provided
Training	249	30442	14314	14314	17	Yes	Yes
Validation	249	8889	4179	4179	2	Yes	Yes
Testing	249	8602	4042	4042	2	No	No



(a) skeleton pair feature (b) HOG feature

Fig. 5: Feature representation of each image frame

concatenation representation is taken as the input of the second recurrent LSTM layer, which is generated by

$$X_j^2 = Y_{Rj}^1 \oplus Y_{Dj}^1 \tag{8}$$

where $\oplus$ is the concatenation operator. For the LSTM layer, the output Y_j^2 can be derived from Eqn.(3-7). Then Y_j^2 are normalized by the softmax function to get each class probability $p(j_k)$. The objective function of our 2S-RNN is multiclass logloss:

$$L(\Omega) = -\sum_{j=0}^{m}\sum_{k=0}^{n-1} \delta(k-r)\, p(j_k|\,\Omega_j) \tag{9}$$

where $\delta(\cdot)$ denotes the Kronecker function, r is the groundtruth label of the gesture sample Ω_j, n is the number of gesture's classes and m denotes the number of gesture samples.

In optimizing the objective function, BPTT algorithm [27] is utilized to obtain the partial derivatives of the objective function and RMSprop method [29] is used to minimize the objective function. Finally, we will obtain the trained parameters, i.e. the connection weights and bias vectors.

D. Recognition & Implementation

Given a segmented RGB-D gesture video, the trained 2S-RNN model will be used to predict the corresponding label with its probability.

In our implementation, the gesture is characterized by the hand skeleton pair feature [30] and HOG feature frame-by-frame, which represent the hand motion and hand posture respectively. Figure 5 gives one example to show the feature we used. The centroids of face and two hands are selected as the key points and the skeleton pair feature is constructed by the relative distances between each pair of three points by dividing the longest distance among the three lines. As for hand representation, we extract the HOG features from the separated two hand regions, which are resized to 32×32. The final feature vector is integrated by concatenating the skeleton pair feature and the HOG feature. Here, one point should be noted that the feature vector is generated for each modality of RGB and depth.

Since the dimension of the original HOG is high, PCA is used for HOG feature dimensionality reduction. The feature dimension for the final hand shape representation is reduced to 81 from 324 for RGB and depth hand images, with nearly 90% energy reserved.

As for programming platform, hand detection is implemented in Caffe [31]. The 2S-RNN model is trained in keras [32] with cuDNN4 on a Titan X GPU.

In our proposed 2S-RNN, each layer has its own settings. The first layer has two paralleled SRNNs and each has 165 neurons and the third LSTM layer has 330 neurons.

IV. Experiments

In this section, our proposed method is evaluated on the Large-scale Continuous Gesture Recognition Dataset of the ChaLearn LAP challenge 2016(ChaLearn LAP ConGD Dataset). First, we give a brief overview on ConGD and its evaluation protocol. Then, we show the performances on different features and also different networks in order to verify the effectiveness of the proposed 2S-RNN method. Finally, the final test results of the top three winners are given and we win the first place.

A. Dataset and Evaluation Protocol

Totally, the ChaLearn LAP ConGD Dataset includes 47933 RGB-D gestures from 22535 RGB-D continuous gesture videos. The data is performed by 21 different signers and split into three mutually exclusive subsets, i.e. the training, validation and testing sets. The detailed information of the database is shown in Table I.

In order to measure the performance of different methods, the Mean Jaccard Index $\overline{J_S}$ [5] is adopted as the evaluation criteria for the continuous gesture recognition. This score measures the average overlap between the predicted gesture labels and the ground truths .

B. Evaluation on Different Features

In gesture representation, different features reveal different aspects of signs [33]. For example, HOG feature of RGB hand region describes the static appearance of hand posture. While the HOG feature of depth hand region characterizes both the hand posture and the change of fingers' distance. The skeleton pair feature describes the dynamic hand motion. Therefore, the discriminative abilities of different feature are accordingly different. This section shows the experimental

TABLE II: Performance comparision of different features on validation set

	RS	RH	RS+RH	DH	DS+DH	**R+D**
Score	0.0211	0.1878	0.2031	0.2229	0.2377	**0.2655**

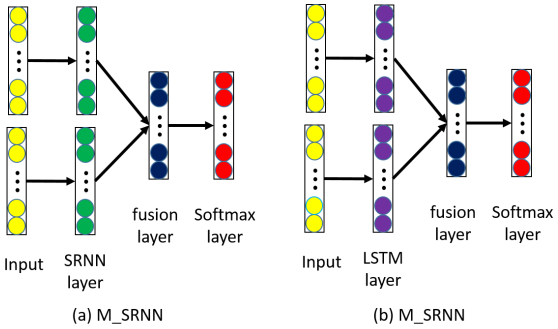


Fig. 6: The flow chart of other two networks

results on the continuous gesture recognition with different features. We conducted the experiments by using six different features, which are HOG feature of hand region in RGB image (RH), skeleton pair feature in RGB image (RS), the fused feature in RGB image (RH + RS), HOG feature of hand region in depth image (DH), the fused feature in depth image (DH + DS) and two streams of feature from RGB and depth modals (R+D). The experiment on skeleton pair feature of depth channel (DS) is omitted because DS is the same with RS.

The performance comparison of different features on validation set is given in Table II. From the results, it can be obviously seen that HOG feature is highly superior to the skeleton pair feature. The combined feature, (both RH+RS and DS+DH) improves the performance than single HOG feature in its own channel, but not very distinctive. In a whole, the function of skeleton pair feature is not satisfied. Intuitively, the gestures are essentially upper-body activities and to well characterize the activity, at least five important skeleton points are needed, including face, left elbow, right elbow, left hand and right hand. However, we can only get three key points, face and two hands, which cannot represent dynamic gesture motion properly and thus play a small role in gesture recognition. The combined feature in depth channel (DS+DH) performs better than that of the RGB channel (RS+RH), which indicates that depth feature has more powerful discriminative ability. This comparison is precisely consistent with our above-mentioned statement. The best result is achieved by the fused feature of these two modalities since the RGB and depth information are complementary to each other in some extent.

C. Evaluation on Different Networks

In order to illustrate the effectiveness of our key recurrent layers, e.g. SRNN and LSTM, the comparison experiments are carried out on two simple networks. The first one (shortened as M_SRNN) is shown in Fig. 6 (a), which has single SRNN layer, followed by a fusion layer and a softmax layer.

TABLE III: Performance comparison of different methods on validation set

Method	Test Set	Score
M_SRNN	Validation	0.2369
M_LSTM	Validation	0.2597
MFSK [5]	Validation	0.0918
MFSK+DeepID [5]	Validation	0.0902
2S-RNN(ours)	Validation	**0.2655**

The second network (M_LSTM) ,as shown in Fig. 6 (b), is constructed by replacing the SRNN layer in M_SRNN with the LSTM layer. Besides these two networks, two baseline methods, e.g. MFSK and MFSK+DeepIDare [5], are also taken as the comparisons for the previous evaluations on the ConGD. MFSK and MFSK+DeepID utilize mixed features around sparse keypoints (MFSK) [34] and Deep hidden IDentity (Deep ID) features [35]. Support vector machine (SVM) is adopted as their gesture classifiers.

The experimental results are listed in Table III. The performance of each recurrent layer can be directly seen and the fusion of our 2S-RNN can improve the performance a step further. As for the other two baseline methods, our 2S-RNN gets much higher recognition result.

TABLE IV: Comparison of the results from the first three winners in this Challenge

Rank	Team	Score
1	**ICT_NHCI(ours)**	**0.286915***
2	TARDIS [36]	0.269235
3	AMRL [37]	0.265506

D. Evaluation on Testing Set

In this section, the final results on the testing set of this ChaLearn LAP large-scale Continuous Gesture Recognition Challenge are given. Table IV lists the results of the first three winners and our group won the first place. The other two teams, TARDIS [36] and AMRL [37], adopted the method of 3D Convolutional Neural Networks and obtained similar results.

Actually, the recognition performance on this ConGD only reached 30% roughly, which shows that the dataset is very challenging. So in the near future, both the segmentation and the recognition modules should be explored in a further step to enhance the performance on continuous gesture recognition.

V. Conclusion

This paper presents an effective spotting-recognition framework for large-scale continuous gesture recognition. First, the continuous gesture sequence is segmented into several isolated gestures according to the accurate hand detection. To recognize

*Our most recent test result is 0.3186, which is evaluated by the organizing team for the ground truth labels of the test data are not released yet. The improvement of the performance comes from the hand detection module.

each segmented gesture, a two streams RNN architecture is designed, which can fuse the RGB and depth features effectively. In each modality, the HOG and skeleton pair features are concatenated to generate the powerful gesture representation. The 2S-RNN can model the contextual information of the temporal gesture sequences and make full use of the RGB and depth information. Experimental results on ChaLearn LAP ConGD Dataset demonstrate the effectiveness of the proposed method.

VI. Acknowledgements

This work was partially supported by 973 Program under contract No 2015CB351802, Natural Science Foundation of China under contracts Nos.61390511, 61472398, 61532018, and the Youth Innovation Promotion Association CAS.

References

[1] H. Wang, Q. Wang, and X. Chen, "Hand posture recognition from disparity cost map," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 722–733.

[2] 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*. ACM, 2011, pp. 279–286.

[3] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 445–452.

[4] X. Chai, H. Wang, and X. Chen, "The devisign large vocabulary of chinese sign language database and baseline evaluations." *Technical Report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS*, 2014.

[5] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *Computer Vision and Pattern Recognition workshop (CVPRW), 2016*.

[6] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner, *Results and Analysis of the ChaLearn Gesture Challenge 2012*, 2013.

[7] S. Ruffieux, D. Lalanne, and E. Mugellini, "Chairgest: a challenge for multimodal mid-air gesture recognition for close hci," in *ACM on International Conference on Multimodal Interaction*, 2013, pp. 483–488.

[8] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1493–1500.

[9] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–8.

[10] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, *Multi-scale Deep Learning for Gesture Detection and Localization*, 2015.

[11] S. Celebi, A. Aydin, T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping," in *VISAPP*, 2013, Conference Paper, pp. 620–625.

[12] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *CVPR*. IEEE, 2011, pp. 3265–3272.

[13] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multimodal gesture recognition via multiple hypotheses rescoring," *JMLR*, vol. 16, pp. 255–284, 2015.

[14] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *PAMI*, vol. 31, no. 7, pp. 1264–1277, 2009.

[15] 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*, vol. 34, no. 16, pp. 2051 – 2056, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513002559

[16] T. Pfister, J. Charles, and A. Zisserman, "Domain-adaptive discriminative one-shot learning of gestures," in *European Conference on Computer Vision*, 2014, pp. 814–829.

[17] E. J. Ong, N. Pugeault, and R. Bowden, "Sign spotting using hierarchical sequential patterns with temporal intervals," in *Proceeedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1931 – 1938.

[18] T. Du, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[19] 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 networks," in *CVPR*, 2016.

[20] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascae for multi-view face detection with alignment awareness." *Neurocomputing(Under review)*.

[21] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," *IN PROC. GRAPHICON-2003*, pp. 85–92, 2004.

[22] O. D. Cortazar, A. Megia-Macias, and A. Vizcaino-De-Julian, "Hand detection using multiple proposals," in *British Machine Vision Conference*, 2011, pp. 75.1–75.11.

[23] 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*, 2014, pp. 580–587.

[24] 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*, pp. 1–1, 2016.

[25] 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*, 2015, pp. 3431–3440.

[26] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[27] A. Graves, "Supervised sequence labelling with recurrent neural networks," *Studies in Computational Intelligence*, vol. 385, 2012.

[28] J. Kolen and S. Kremer, *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*. Wiley-IEEE Press, 2009.

[29] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[32] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[33] H. Wang, X. Chai, Y. Zhou, and X. Chen, "Fast sign language recognition benefited from low rank approximation," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[34] J. Wan, G. Guo, and S. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," 2015.

[35] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[36] N. C. Camgöz, H. Simon, K. Oscar, and B. Richard, "Using convolutional 3d neural networks foruser-independent continuous gesture recognition," in *International Conference on Pattern Recognition workshop (ICPRW), 2016*.

[37] W. Pichao, L. Wanqing, L. Song, Z. Yuyao, G. Zhimin, and O. Philip, "Large-scale continuous gesture recognition using convolutional neural networks," in *International Conference on Pattern Recognition workshop (ICPRW), 2016*.