A Framework For Multimodal Sign Language Recognition Under Small Sample Based On Key-frame Sampling

Jianyu Wang, Jianxin Chen, Yihao Cai
Nanjing University of Posts and Telecommunications, Nanjing, China.

ABSTRACT

Sign language recognition is challenging, due to the scarcity of available annotated corpora and the difficulty of large vocabulary. In this paper, we study the task based on a Chinese SL database-DEVISIGN, but it only has a few samples to train the deep network on the scratch. First, we segment the hand to eliminate the disturbance of irrelevant factors. By analyzing the special movement tendency of sign words, we propose two novel Key-frame selection schemes. Since no other datasets can have similar data distribution with our preprocessed data, we invent a novel cross-sampling approach, which successfully prevent the overfitting under small sample. To enhance the diversity of data, we take several sampling-based videos as input, and learn spatiotemporal features based on R(2+1)D-18 layers, which is successful in action recognition tasks. Finally, it is shown that our solution can obtain the state-of-the-art performance.

Keywords: isolated sign language recognition, key-frame sampling mechanism, multimodal fusion

1. INTRODUCTION

Nowadays, hearing loss has severely negative impacts on the life quality of deaf people [1]. It is quite difficult for a deaf person to interact with common people as the common people seldom gain the knowledge of sign language. Hopefully, automatic sign language recognition is possible to bridge the communication gap.

However, there are some limits on sign language recognition [2]. Because of the large vocabularies, large-scale formal sign language datasets are not available as the normal gesture datasets such as Chalearn LAP IsoGD Database [3] and the Sheffield Kinect Gesture dataset (SKIG) [4].

Herein, we choose the Chinese large-scale sign language database DEVISIGN [5], which covers 2000 standard Chinese sign language words, for researching, although there are not many predecessors taking it for examination, Figure. 1 shows some examples of the collected sign data by Kinect. The first row represents videos without pre-processing, the left is the illustration figure, the middle is RGB image and the right is the depth image. The results on the second row vividly shows the results of hand segmentation and optical flow computation, the left picture is processed RGB image, the middle picture is processed depth image and the right part is optical-flow image.

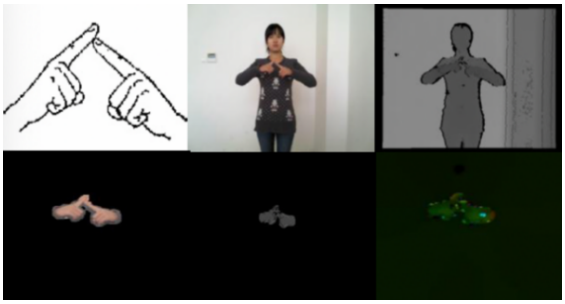

Figure 1. An example of the sign word 'human', RGB data, depth data and optical flow data.

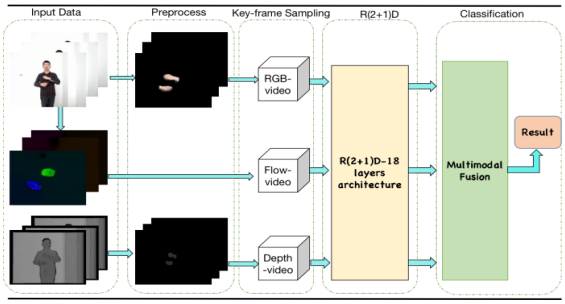

Figure 2. Pipeline of the proposed framework.

However, among previous Chinese sign language recognition works in DEVISIGN, few people considered about deep-learning method because of the small sample, and thus many researchers failed to capture the temporal information. What's more, even in other common isolated gesture datasets, many methods using deep networks had to abandon the model-size for better classification results like C3D [6], C3D+ConvLSTM [7] and Res-C3D [8]. Therefore, in order to improve the classification results while not enhancing the parameters in the model, we propose a multimodal isolated sign language

Fifth International Workshop on Pattern Recognition, edited by Xudong Jiang, Chuan Zhang, Yinglei Song,
Proc. of SPIE Vol. 11526, 115260A · © 2020 SPIE · CCC code: 0277-786X/20/$21 · doi: 10.1117/12.2574424

recognition method using a R(2+1)D network [9], which has been proved valid for they outperform many innovative strategies like C3D [10], P3D [11] and I3D [12] while keeping relative suitable model-size in action recognition tasks. Compared to the previous factorized P3D (an effective method by interleaving three blocks in sequence on networks), it is simple and homogeneous while the same (2+1)-decomposition is used in all blocks. Therefore, we want to take a try to implement this outstanding architecture in gesture recognition tasks since its performance in action recognition.

Furthermore, using our novel cross-sampling approach can successfully train deep network under small sample when no pre-trained models can be found. To recap, as illustrated in Figure. 2, our contributions in this paper include:

- We first utilize deep (2+1)D network based on convolution factorization in gesture-related tasks, which are quite useful since they can improve the model capacity by doubling the nonlinearities, and our method first captures temporal information and outperforms previous state-of-the-art on DEVISIGN-D.
- Rather than taking uniform sampling [13, 14, 15], which are commonly utilized to normalize the length of video clips, we design two new schemes relied on the skeleton data and optical flow data, which capture the key information about sign language well.
- Although it seems that finding pre-trained models might be impossible due to unique data distribution of our preprocessed data, we successfully implement the cross-sampling strategies as an optional approach to prevent overfitting under small sample.

2. MULTIMODAL FUSION FRAMEWORK

As illustrated in Figure. 2, our deep architecture is composed of four modules. At first, we adopt the temporal segmentation based on mixed segmentation method with RGB data, depth data and skeleton data, and the result is illustrated in Figure. 1. At the same time, we collect the optical flow videos generated from the RGB stream, which is illustrated in Figure. 1.

Since CNN-based networks require fixed input dimension, we invent two new key-frame sampling approaches with skeleton data and optical flow data by concluding the immanent tendency of sign language performing, which proved to outperform the uniform sampling a lot. Then, with the novel cross-sampling method finetuned using R(2+1)D architecture, we train the network successfully. Moreover, we enhance the diversity of data by using different sampling strategies as input. Finally, we fuse the scores from three channels to get the final multimodal classification result.

2.1. Data preparation

Since sign language recognition is quite different from action recognition that in the whole video, eliminating the disturbance of variant illumination, different clothes and human face pixels might be a useful improvement. Therefore, reducing the data complexity by undermining interference of irrelevant factors might also be useful to prevent overfitting under small sample.

In order to select good approaches to segment the hand, the multimodal information is utilized, e.g. depth information, skeleton data and several color spaces. First, enhancing the multiple thresholding algorithm in several spaces in RGB, HSV and YCrCb [16] by implementing the Otsu thresholding to get better skin detection results. Second, hand positions are roughly determined by the positions of hand joints. Third, according to the hand position in each video and relative depth information, a depth-based mask is obtained. Finally, the bitwise AND operation is used to synthesize the color-based mask and the depth-based mask to get effective hand segmentation, which is shown in Figure.1.

In order to get robust optical flow computation, we generate the optical flow from the RGB data stream and we save optical flow values as RGB images to use as another modality of data to enhance the classification performance. An optical flow image example is shown in Figure. 1.

2.2. Key frame sampling

2.2.1 Video analysis

CNN-based networks require fixed input dimension, which means that we have to take some sampling schemes to choose frames from raw videos. Uniform sampling is widely utilized in previous studies [13, 14, 15]. However, uniform sampling fails to capture the key information in isolated sign language video as uniform sampling will ignore sign video distinct attributes.

As we know, in most isolated sign language video, the action can be divided into three phases [8]: beginning, climax and ending. To get an understanding of their features, we will analyze the movement in the following discussion based on Figure. 3. In Figure. 3, in the beginning stage, the signer's movement is slight until he actually raises his hand in twentieth

frame, which implies that the beginning stage has little information about the signs difference. Moreover, from the 20th to the 80th frame, it is the climax stage, the movement is quite sharp, especially around the 50th frame. Therefore, the climax phase must be the most important stage which includes key information about the isolated sign language.

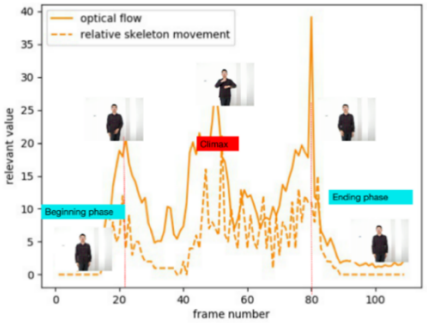

Figure 3. The relative movement tendency of the sign word 'the weak'. The value of the Y axis represents the relative value. In order to observe the tendency, all of them are multiplied by a constant number.

2.2.2 Key frame Definition

By analyzing the movement in each sign word, we find that when the movement is quite sharp in a video segment, this segment must contain much information with high representativeness. Therefore, we propose two novel key-frame sampling methods, one approach based on optical flow (OF-S) and the other method based on skeleton-data (S-S). As OF-S and S-S work processes are the same, here we list the procedure of S-S.

We divide the whole video into several segments to analyze their movement tendency, respectively. Firstly, we uniformly divide them into n clips. For each clip, we compute the sum of the absolute value of difference about left and right hands' 3D coordinates in adjacent frames d_{ki} ($1 \leqslant i \leqslant n$). For example, the total sum of these clips' value is computed as follows:

$$S_K = d_{k1} + d_{k2} + \cdots + d_{ki} + \cdots d_{kn} \tag{1}$$

Then, the weight of the i-th clip is defined as

$$W_{ki} = \frac{d_{ki}}{S_K} \tag{2}$$

Then for each video segment, the needed number of frames are obtained according to the weight. Let S denote the required number of frames, we define

$$Se_i = W_{ki} \times S \tag{3}$$

Finally, uniform sampling in each video segment is used to select the frames according to the value of Se_i. For example, in our experiments, for each video containing m frames, we set $n = \frac{2 \times m}{S}$, and S is set to 32. Then the segment interval is 16.

2.3. Learning spatiotemporal features based on R(2+1)D-18 model

As the dimension of input data has been determined, we focus on the model to extract spatiotemporal feature. Nowadays, R(2+1)D is a high performing network in action recognition tasks. Furthermore, we select R(2+1)D-18 layers [9] for several reasons. On the one hand, R(2+1)D maintains the advantage of R3D, and has better model capacity while maintaining approximately equal number of parameters. On the other hand, we only have two GPUs in our lab to train neural networks, which means it is hard to train very deep networks. Therefore, we apply this R(2+1)D-18 layers architecture in our task.

Figure 4. R3D block vs R(2+1)D block a)The residual building-block used in the R3D architecture. b) The residual building-b lock used in the R(2+1)D architecture.

The structure of 3D block and (2+1)D block are illustrated in Figure.4. The main difference between 3D block and (2+1)D block is that (2+1)D block decomposes the N_i 3D convolutional filters (the size is $N_{i-1} \times t \times d \times d$, where t denotes the temporal extent of the filter and d denotes the spatial dimension of the filter) into M_i 2D Convolutional filters(the size $is\ \ N_{i-1} \times 1 \times d \times d$) and Ni temporal convolutional filters(the size is $M_{i-1} \times t \times 1 \times 1$). Hyperparameter M_i signifies the intermediate subspace between the spatial and the temporal convolutions, N_i denotes the number of filters in the *i-th* block. The relation between M_i and N_i is

$$M_i = \left\lfloor \frac{td^2 N_{i-1} N_i}{d^2 N_{i-1} + t N_i} \right\rfloor \tag{4}$$

Therefore, the number of parameters in R(2+1)D block approximates to the number of parameters in full 3D-Convolution while doubling the nonlinearities, which leads to better capacity.

2.4. Multimodal fusion

In this paper, late fusion is implemented by fusing the scores on the prediction results of the multimodal data. Besides, we also try to implement early fusion by concatenating the features to improve the performance. After several attempts, we take the feature-level fusion by sending concatenated features to a linear SVM to get the final classification results, as illustrated in Table. 5.

3. EXPERIMENTS

3.1. Datasets

DEVISIGN [5] is a Chinese vocabulary database, which covers 2000 Sign Language words captured by Kinect. We select the subset-DEVISIGN-D for examination. It contains 500 daily used vocabularies, including 36 static gestures in DEVISIGN-G. The data covers 8 different signers. Among them, the vocabularies are recorded twice for 4 signers and once for other 4 signers. Since having to do the signer independent tasks, the probe set should contain data from 4 different signers.

Table 1. Data partition of DEVISIGN-D datasets in the evaluation protocols.

Training Data	Probe Data
P01_1, P01_2, P02_1, P02_2, P03_1, P03_2, P04_1, P04_2	P05_1, P06_1, P07_1, P08_1

3.2. Networking training

We conduct our experiments on a PC with Intel Core i5-6200 CPU @ 2.30GHZ × 8,16GB RAM and NVIDIA Tesla M40 GPU. Input videos are sampled into 32 frames and each frame is resized into 128×171. Moreover, the video clips are random-cropped into 112×112. To learn the weights of the architecture, we parallel train them by using the Adam optimizer with a batch size of 6 based on two Tesla M40 GPU. The initial learning rate is set to 0.0001 and the training process is stopped after 12 epochs. Weight decay is set to 0.00005. In addition, Batch Normalization layer is implemented after each convolutional layer to improve the training efficiency. Furthermore, the optical flow videos are generated by us with pyflow [17], a python wrapper for dense optical flow.

In DEVISIGN-D, every vocabulary has 12 videos: 8 videos for training and other 4 videos for testing. Therefore, in order to keep away from overfitting, the data is augmented by three strategies like image mirroring, different sampling unification approaches and random crop. We also use temporal jitter [7] to augment the dataset.

As no pre-trained models on other datasets are utilized in our training, a novel fine-tuning approach for DEVISIGN-D has been implemented by us. To be specific, we train on uniform sampling videos, then fine-tuning on the key-frame based sampling videos.

Several strategies are implemented to help us evaluate the proposed approach:

Strategy 0: To obtain effective segment interval, we try to compare several intervals' results. Since some videos only contain 33-35 frames, we select 16 as maximum segment interval to ensure at least two video segments. To analyze the tendency, we choose 4, 8 and 16.

Strategy 1: To test our novel sampling methods, we train the RGB-based networks of different sampling strategies on the scratch, respectively.

Strategy 2: To test our cross-sampling method, train uniform-sampling (U-S) based networks on the scratch at first, and then finetune other sampling approaches' models based on the models of uniform sampling-based models in three modalities, individually.

Strategy 3: To enhance the diversity of our data, we use both the optical flow-based sampling (OF-S) videos and the skeleton-based sampling (S-S) videos as the input, we first train skeleton-based sampling RGB videos on the scratch, then finetune the models of the synthetic strategies on the models of skeleton-based sampling RGB videos in each modality, respectively.

Strategy 4: We also compare several models based on the same training tricks as Strategy 3, such as C3D, C3D + ConvLSTM +SPP, R3D18 and R3D34.

Table 2. Sign language recognition accuracy for different strategies of different modalities. To analyze the results more evidently, we list top-1 accuracy and top-5 accuracy.

	Modality	Method	Top-1	Top-5
Strategy 0	RGB	Interval-4	42.86	69.56
	RGB	Interval-8	44.06	70.51
	RGB	Interval-16	44.28	72.09
Strategy 1	RGB	U-S	36.80	63.75
	RGB	OF-S	39.47	68.46
	RGB	S-S	44.28	72.09
Strategy 2	RGB	OF-S	49.05	74.00
	Depth	OF-S	47.31	72.95
	Flow	OF-S	39.47	65.82
	RGB	S-S	49.90	75.50
	Depth	S-S	49.65	75.05
	Flow	S-S	41.92	70.56
Strategy 3	RGB	OF+S	**52.22**	**77.39**
	Depth	OF+S	**53.07**	**77.31**
	Flow	OF+S	**49.48**	**74.49**

3.3. Learning spatiotemporal features based on R(2+1)D-18 model

Table 2 shows the results of four strategies on different modalities. The findings demonstrate that our sampling strategies are beneficial to sign language recognition, and cross-sampling finetuning might be a quite useful approach to prevent overfitting when no pretrained models are available.

Observing the results about three intervals, we find that it seems the final results is proportional to increasing interval. This is due to that if each clip only contains a small number of frames, the difference of importance between clips will be very small, which is hard to select key frames by importance. Finally, interval is set to 16.

3.3.1. Key Frame Sampling

By training the networks on the scratch, we can deduce that our two strategies must contain more valuable information than the normal sampling mechanism. S-S and OF-S results outperform the uniform-sampling result in strategy 1, e.g. S-S achieves 7.48% improvement in top-1 acc and 8.34% improvement in top-5 acc. OF-S achieves 2.67% relative improvement in top-1 acc and 4.71% relative improvement in top-5 acc.

3.3.2. Small sample

With only small sample, it is quite easy to overfit even we have done some data augmentation tricks. Therefore, finetuning must be utilized to prevent over-fitting. However, due to our distinct data distribution of our processed videos, it is difficult to find any other common datasets which have similar data distribution. Fortunately, we find that taking cross-sampling

strategy's model as the pretrained model can solve this problem coincidentally. From Table 2, we surprisingly find that the cross-sampling strategies could help prevent the overfitting and improve the final classification results, e.g. OF-S achieves 12.75% improvements in RGB modality.

3.3.3. Model Selection

We also compare several models performance based on the same training tricks, such as C3D, CLSTM+SPP, R3D18 and R3D34, depicted as Table 3.

Table 3. Isolated Sign Language recognition accuracy for different models.

Net	Params	Method	Top-1	Top-5
C3D[6]	113.6M	OF+S	37.04%	67.04%
CLSTM+SPP[13]	29.9M	OF+S	37.23%	62.27%
R3D-18[9]	33.4M	OF+S	48.78%	74.59%
R3D-34[9]	56.5M	OF+S	48.10%	71.89%
***R(2+1)D*[9]**	**33.3M**	**OF+S**	**52.22%**	**77.39%**
Maximum	99.9M	OF+S	59.43%	/
Averge	99.9M	OF+S	59.43%	/
Concatenate	**99.9M**	**OF+S**	**61.51%**	/

Table 4. The evaluation results of previous state-of-the-art and our methods on DEVISIGN-D.

Method	**Accuracy**
HMM[5]	43.5%
ARMA[5]	46.5%
fastDTW[5]	49.0%
GCM[18]	57.0%
Proposed Method	**61.51%**

The results reveal that comparing with previous high performance methods related to gesture recognition, Resnet-3D performs better. Also, it reveals that even deeper like R3D-34, the final accuracy isn't higher than R3D18. Therefore, it is suitable to choose R(2+1)D-18 layers in our small-scale dataset.

3.3.4. Multimodal

From strategy 3, we have learnt that when using several strategies together to enhance the diversity of the training data, the overall performance will be improved a lot.

After that, we do some work about different strategies for multimodal fusion since it can further improve the final classification results, such as score fusion and feature fusion. Two score fusion methods are examined for the deep architecture in Table 3, such as average fusion and maximum fusion. Furthermore, we also concatenate the features from different modalities to obtain a single vector sent to a linear SVM classifier, the results of above methods are depicted in Table 3. From Table 3, we find that in our experiments, feature-level fusion as concatenating can increase the accuracy by at least 8.441%, while late fusion as maximum fusion can achieve better results about 6.356% improvement, so we choose concatenating fusion scheme at last.

3.4. Comparisons of state-of-the-art

The performance of our deep-learning based method is compared with the previous state-of-the-art methods about hand-craft features like [2]. From Table 3 and Table 4, we find that even the single modality's accuracy can overperform many previous multimodal results.

3.5. Validation Analysis

As we can see in the Figure. 5, the results reveal that double hand words are recognized with higher accuracy than single hand words. It seems that single hand words might be more challenging in this task, so we take a clear look at the database and find that many completely wrongly predicted single hand words are static gestures like '0'-'9' and 'A'-'Z'. Almost 52.8% static single words are classified completely wrong. It reveals that spatiotemporal extractors may not be an effective approach for these static sign words who have small number of useful and distinct frames and many irrelevant frames.

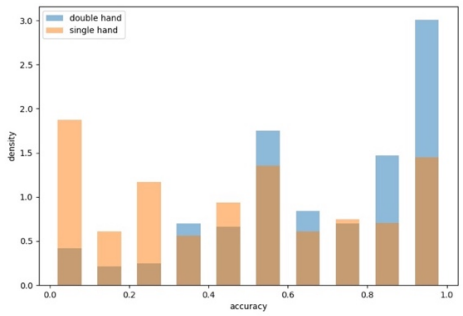

Figure 5. Per-class validation accuracy of our methods.

4. CONCLUSIONS

This paper presents an effective architecture based on R(2+1)D for isolated sign language recognition under small sample. The results demonstrate that our key-frame sampling methods outperform common uniform sampling in sign language recognition, and cross- sampling will be an effective mechanism to be used as a finetuning method even when no suitable pretrained models exist. However, the results might be better if we consider to fuse skeleton features or we consider to utilize attention mechanism to improve the feature extraction.

ACKNOWLEDGEMENTS

This work was supported by the National Science and Technology Innovation Training Program (No. SZDG2018002); The National Science Foundation of China (No. 61901227); The National Science Foundation of the Jiangsu Higher Education Institutions of China (No. 19KJB510049).

REFERENCES

[1] Magariños M, Milo M, Varela-Nieto I. Aging, neurogenesis and neuroinflammation in hearing loss and protection[J]. Frontiers in aging neuroscience, 2015, 7: 138.

[2] Kamal S M, Chen Y, Li S, et al. Technical Approaches to Chinese Sign Language Processing: A Review[J]. IEEE Access, 2019, 7: 96926-96935.

[3] Wan J, Zhao Y, Zhou S, et al. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 56-64.

[4] Liu L, Shao L. Learning discriminative representations from RGB-D video data[C]. Twenty-Third International Joint Conference on Artificial Intelligence. 2013.

[5] Chai X, Wanga H, Zhoub M, et al. DEVISIGN: Dataset and Evaluation for 3D Sign Language Recognition[R]. Beijing, Tech. Rep, 2015.

[6] Li Y, Miao Q, Tian K, et al. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model[C]. 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 25-30.

[7] Zhu G, Zhang L, Shen P, et al. Multimodal gesture recognition using 3-D convolution and convolutional LSTM[J]. IEEE Access, 2017, 5: 4517-4524.

[8] Miao Q, Li Y, Ouyang W, et al. Multimodal gesture recognition based on the resc3d network[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 3047-3055.

[9] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.

[10] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[11] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.

[12] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[13] Zhang L, Zhu G, Shen P, et al. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 3120-3128.

[14] Zhu G, Zhang L, Shen P, et al. Multimodal gesture recognition using 3-D convolution and convolutional LSTM[J]. IEEE Access, 2017, 5: 4517-4524.

[15] Wang H, Wang P, Song Z, et al. Large-scale multimodal gesture recognition using heterogeneous networks[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 3129-3137.

[16] Gasparini F, Schettini R. Skin segmentation using multiple thresholding[C]. Internet Imaging VII. International Society for Optics and Photonics, 2006, 6061: 6061.

[17] Pathak D, Girshick R, Dollár P, et al. Learning features by watching objects move[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2701-2710.

[18] Wang H, Chai X, Hong X, et al. Isolated sign language recognition with grassmann covariance matrices[J]. ACM Transactions on Accessible Computing (TACCESS), 2016, 8(4): 14.