A Gesture Recognition Approach Using Multimodal Neural Network

Xiaoyu Song, Hong Chen*, Qing Wang
College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
norman_chen@263.net

Abstract. Gesture recognition based on visual modal often encounters the problem of reduced recognition rate in some extreme environments such as in a dim or near-skinned background. When human beings make judgments, they will integrate various modal information. There should also be some connections between human gestures and speech. Based on this, we propose a multimodal gesture recognition network. We use 3D CNN to extract visual features, GRU to extract speech features, and fuse them at late stage to make the final judgment. At the same time, we use a two-stage structure, a shallow network as detector and a deep network as classifier to reduce the memory usage and energy consumption. We make a gesture dataset recorded in a dim environment, named DarkGesture. In this dataset, people say the gesture’s name when they make a gesture. Then, the network proposed in this paper is compared with the single-modal recognition network based on DarkGesture. The results show that the multi-modal recognition network proposed in this paper has better recognition effect.

1. Introduction
As a natural and convenient means of interaction, non-contact gesture recognition based on computer vision plays an increasingly important role in human-computer interaction applications. Many current gesture recognition methods mainly use images collected by cameras for processing. This method is generally restricted by factors such as dim lighting, background complexity, and near-skin environment.

For a long time, the input of gesture recognition has only one modality of video image, and the input of speech recognition only has one modality of audio. Nowadays, the improvement of accuracy has been very gentle, reaching a bottleneck stage. On the other hand, more and more people notice that there is complementary information between multiple modalities. For example, the visual characteristics of the mouth shape change during speech recognition can significantly improve the effect of speech recognition in noisy environments. In the affect emotion analysis, text information, facial information and audio information are combined to eliminate ambiguity and uncertainty. The concept of information complementarity between multiple modalities also brings new ideas to the field of gesture recognition. For example, in [1], it is believed that there is connection and complementarity between the gestures and speech. Therefore, adding speech modality as supplementary information to improve the robustness of recognition is of great research value.

In this paper, we propose a multi-modal gesture recognition network based on audio-visual. It consists of two parts, a shallow detector to identify whether gestures appear in the scene and a deep classifier for classification. In this network, 3D CNN is used to extract visual features and GRU is used to extract speech features. The detector is the switch of the classifier. The classifier is turned on
only when the detector detects a gesture. Once the detector cannot detect a gesture, the classifier is
turned off. The reason why the detector and classifier architecture are introduced is that in gesture
recognition for human-computer interaction, most of the time, there is no gesture in the scene, so it is
not necessary to maintain the high performance of the network at all times. Figure 1 shows the
pipeline of the proposed approach.

![Figure 1](image)

**Figure 1. Illustration of the proposed pipeline for gesture recognition.** The detector and classifier both move along the input stream with a sliding window. The detector and classifier each of them maintains their own queue. When the score in the detector queue exceeds a threshold, the classifier is turned on, and the classifier make the final conclusion according to the scores in its queue.

In our experiments, we used C3D and GRU to extract the features of their respective modalities for fusion. In fact, the feature extraction network can be replaced at will. The reason for choosing C3D and GRU is to ensure the accuracy of identification, reduce network parameters, reduce energy consumption and delay.

2. Related Work

In terms of vision-based gesture recognition, there are various CNNs used to extract temporal features. In [2], video frames are input to 2D CNN as multiple channels. In [3], the author used 2D CNN to extract shallow features firstly, and then input these features into an LSTM network to further extract temporal features. Although 2D CNN is very effective in extracting spatial features, due to these structural limitations, 2D CNN alone cannot extract timing features well, and LSTM will cause a large amount of calculation.

C3D has a good compromise between the above problems. In [4], deep C3D is used as a feature extraction network and achieved significant results. In this paper, we will also use C3D to extract visual features.

For the extraction of speech features, the combination of MFCC and HMM has achieved great success for a long time, but with the continuous development of DNN, it has surpassed HMM in effect. However, the end-to-end model has the problem of being difficult to train, so extracting the MFCC features and then inputting them to the DNN is the current mainstream practice. Both [5] and [6] have adopted the method of post-processing by DNN after feature extraction and achieved the best results. In this paper, we will first extract FBank features and then input them into GRU to extract speech features.
3. Model Architecture

Figure 2 shows the basic structure of the detector in this work. The structure of the classifier is the same as that of the detector, but there are some differences in the depth of the model and the output. The classifier is deeper, and the final classification result is output.

![Figure 2. Illustration of the structure of the detector.](image)

For the detector, its only role is to recognize the presence of a gesture, then turn on the classifier, and then close the classifier when the gesture is over. Because our structure is very dependent on the speed of the detector, it must be high-speed. At the same time, it should not miss any gesture, so it should also have the characteristics of high recall. In this work, we fully connect the output of a 10-layers ResNet-C3D and a single-layer GRU as a detector. In order to improve the recall rate, weighted-cross entropy loss is used for training and positive samples have higher weights. In addition, the detector also maintains a queue of length \( n \). The queue stores the score value of the softmax layer output of the detector, and each time the detector’s final output is based on the average of the previous score in the queue.

The classifier requires high accuracy. We use the ResNeXt-C3D of the 101 layers and the output of the GRU of the two layers as the classifier. The classifier's network processes the data stream only when the detector opens the classifier. The classifier also maintains a queue of length \( m (m > n) \). The queue stores the previous score results of the classifier, and the final result is also made according to the average of the previous scores. Unlike the detector, the classifier will only output results when the score of a certain class exceeds the threshold or the detector closes the classifier.

Figure 3 shows a schematic diagram of the units of the networks proposed in this paper.

![Figure 3. Illustration of the units of the networks.](image)
4. Data

We use the EgoGesture[7] dataset to train a visual feature extraction network. EgoGesture is a large-scale data set for first-person interactive gesture classification. It collects 83 static and dynamic gestures from 6 different indoor and outdoor scenes, splits the data into 3: 1: 1, and generates 1,239 respectively. The training set, 411 validation set and 431 test set, have 14416, 4768, and 4977 gesture samples, respectively. Then use Olivia[8] to train the speech feature extraction network. Olivia consists of spoken examples of 1544 keywords.

Finally, the homemade dataset DarkGesture is used to train for our multimodal network, including 20 dynamic gestures and 55 videos, which are also split into 3: 1: 1. Unlike the EgoGesture dataset, the videos in DarkGesture are Recorded under very dim conditions, and people say its name when making a gesture. Figure 4 shows an example of data set.

![Illustration examples of a data set.](image)

5. Experiment and result

Firstly, for visual feature extraction networks, thanks to the ResNeXt for calssifier and ResNet for detector pre-trained models provided by [9], they have achieved good results on EgoGesture. On the Olivia dataset for GRU, the learning rate is initially 0.02, parameters are optimized with stochastic gradient descent for 50 epochs and a minibatch size of 128.

Then, we use stochastic gradient descent with Nesterov momentum = 0.9, damping factor = 0.9, and weight decay = 0.001 as optimizer for dector and classifier on DarkGesture 50 epochs. The learning rate is started with 0.01, and divided by 10 at 10th and 25th epochs, and training is completed after 5 more epochs for detector and classifier. The difference is that the detector uses weighted-cross entropy loss function and the classifier uses cross entropy loss function.

Finally, we make two comparison experiments. The control group of each experiment is a visual classification model, and the experimental group is the multi-modal recognition method proposed in this paper. First, the discrete samples (the test set after segmenting the DarkGesture video) were tested and compared with the effects of the detector and the classifier, and then the comparison of the real-time gesture recognition result in the continuous video was also tested and compared. Table 1, 2, 3 shows the results of our experiments.

We use Levenshtein distance as our evaluation metric for real-time experiments. The Levenshtein distance is a metric that measures distance between sequences by counting the number of item-level changes (insertion, deletion, or substitutions) to transform one sequence into the other. For our case, one video and the gestures in this video correspond to a sequence and the items in this sequence, respectively.

| Model     | Input         | Precision | Recall  |
|-----------|---------------|-----------|---------|
| ResNet-10 | 16-frames     | 78.95     | 80.36   |
Table 2. Classifier’s classification accuracy scores on the test set of DarkGesture dataset.

| Model                | Input                  | Precision |
|----------------------|------------------------|-----------|
| GRU                  | 1s                     | 87.91     |
| ResNet-10 + GRU      | 16-frames + 1s\textsuperscript{c} | 93.37     |

Table 3. Real-Time Classification Results on DarkGesture dataset.

| Modality     | Queue Size | Levenshtein Precision |
|--------------|------------|-----------------------|
| Video        | 4 / 16     | 75.60                 |
| Audio        | 4 / 16     | 85.72                 |
| Video + Audio| 4 / 16     | 90.47                 |

6. Conclusion
In this paper, we propose a gesture recognition network that fuses visual and speech modalities. The proposed network fuses the features of the two modalities at a late stage. And in the real-time recognition phase, a two-stage architecture with a separate detector and classifier is adopted, which greatly reduces memory usage and energy consumption without losing accuracy.

Because of the complement of the speech modality, the model proposed in this paper has achieved a significant performance improvement over the single-modal recognition method in the DarkGesture dataset, which lacks lighting. Due to the limitations of the data set and experimental hardware, we have not achieved a very satisfactory accuracy rate, but we believe that the experiments in this paper can prove that the complementarity between multi-modalities will become an important way to improve the robustness of gesture recognition in the future.

References
[1] Wagner, P. , Malisz, Z. , & Kopp, S. . (2014). Gesture and speech in interaction: an overview. Speech Communication, 57, 209-232.
[2] Karpathy, A. , Toderici, G. , Shetty, S. , Leung, T. , & Fei-Fei, L. . (2014). Large-Scale Video Classification with Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[3] Donahue, J. , Hendricks, L. A. , Guadarrama, S. , Rohrbach, M. , Venugopalan, S. , & Saenko, K. , et al. (2014). Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Software Engineering.
[4] Tran, D. , Bourdev, L. , Fergus, R. , Torresani, L. , & Paluri, M. . (2016). Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE.
[5] Arik, S. O. , Klieg, M. , Child, R. , Hestness, J. , Gibiansky, A. , & Fougner, C. , et al. (2017). Convolutional recurrent neural networks for small-footprint keyword spotting.
[6] G. Chen, C. Parada and G. Heigold, (2014). "Small-footprint keyword spotting using deep neural networks," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 4087-4091.
[7] Y. Zhang, C. Cao, J. Cheng and H. Lu, (2018). "EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition," IEEE Transactions on Multimedia (T-MM), Vol. 20, No. 5, pp. 1038-1050.

[8] Lengerich, C., & Hannun, A. (2016). An end-to-end architecture for keyword spotting and voice activity detection.

[9] Ö. Köpükü, A. Gunduz, N. Kose and G. Rigoll, "Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1-8.