Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild

Cheng Lu
School of Information Science and Engineering
Southeast University, Nanjing, China
cheng.lu@seu.edu.cn

Wenming Zheng*
Key Laboratory of Child Development and Learning Science of Ministry of Education
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
wenming_zheng@seu.edu.cn

Chaolong Li
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
lichaolong@seu.edu.cn

Chuangao Tang
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
tcg2016@seu.edu.cn

Suyuan Liu
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
syl@seu.edu.cn

Simeng Yan
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
220174274@seu.edu.cn

Yuan Zong
School of Biological Sciences and Medical Engineering
Southeast University, Nanjing, China
xhzongyuan@seu.edu.cn

ABSTRACT

The difficulty of emotion recognition in the wild (EmotiW) is how to train a robust model to deal with diverse scenarios and anomalies. The Audio-video Sub-challenge in EmotiW contains audio-video short clips with several emotional labels and the task is to distinguish which label the video belongs to. For the better emotion recognition in videos, we propose a multiple spatio-temporal feature fusion (MSFF) framework, which can more accurately depict emotional information in spatial and temporal dimensions by two mutually complementary sources, including the facial image and audio. The framework is consisted of two parts: the facial image model and the audio model. With respect to the facial image model, three different architectures of spatial-temporal neural networks are employed to extract discriminative features about different emotions in facial expression images. Firstly, the high-level spatial features are obtained by the pre-trained convolutional neural networks (CNN), including VGG-Face and ResNet-50 which are all fed with the images generated by each video. Then, the features of all frames are sequentially input to the Bi-directional Long Short-Term Memory (BLSTM) so as to capture dynamic variations of facial appearance textures in a video. In addition to the structure of CNN-RNN, another spatio-temporal network, namely deep 3-Dimensional Convolutional Neural Networks (3D CNN) by extending the 2D convolution kernel to 3D, is also applied to attain evolving emotional information encoded in multiple adjacent frames. For the audio model, the spectrogram images of speech generated by preprocessing audio, are also modeled in a VGG-BLSTM framework to characterize the affective fluctuation more efficiently. Finally, a fusion strategy with the score matrices of different spatio-temporal networks gained from the above framework is proposed to boost the performance of emotion recognition complementally. Extensive experiments show that the overall accuracy of our proposed MSFF is 60.64%, which achieves a large improvement compared with the baseline and outperform the result of champion team in 2017.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Emotion Recognition, Spatio-Temporal Information, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), 3D Convolutional Neural Networks (3D CNN)

ACM Reference Format:
Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. 2018. Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild. In *ICMI '18: 2018 Int'l*

*The corresponding author.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264992

Conference on Multimodal Interaction, Oct. 16–20, 2018, Boulder, CO, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3242969.3264992

1 INTRODUCTION

Emotional communication between humans and machines is a major technical problem of artificial intelligence, hence emotion recognition plays an essential role in human-computer interaction [2]. The emotion recognition in the wild (EmotiW) challenge [4], which is most close to the interaction in real scenario, has attracted a lot of attentions from the competitors for six years since 2013. The audio-video sub-challenge is one of the main tasks in EmotiW 2018. For this task, the main purpose is to automatically assign an emotional label to a given video clip from seven main emotions (angry, disgust, fear, happy, sad, surprise and neutral). All competitors train a robust model through the released training and validation dataset so as to get the final recognition result on the test dataset. The dataset of audio-video [5] is consisted of many video clips in which the facial expressions are always contaminated by abnormal conditions such as the variation of illumination, occlusion, rotation, perspective, or scale change. These anomalies make it too difficult to accurately extract the features of the emotional information from facial images in video clips. Therefore, how to obtain more abundant and effective emotional information from the training data for robust model is a pivotal issue in EmotiW challenge.

Benefiting from the significant advantages in the feature representation of image by convolutional neural networks (CNN) [16], a lot of computer vision tasks have been made breakthroughs, especially in image classification. Many researches indicate that, compared to traditional methods, CNN can obtain hierarchical representation in images including low/middle/high-level features. These hierarchial features can more accurately characterize the details of images in spatial dimension than the hand-crafted features. However, for the video analysis tasks, it is not enough to just consider the appearance feature of images in spatial dimension, and the motion information of video is also significant. The spatial feature generated by pure CNN methods are not directly suitable for videos since the lack of motion modeling which encoding in the adjacent frames of videos. Hence, in video tasks, it in no enough to only model image appearance texture well for the spatial features. The dynamic variation of motion information in video, which can be captured by temporal representation. Recurrent Neural Networks (RNN) [17] provide an attractive framework for propagating information over a sequence using a continuous valued hidden layer representation to capture the temporal features, and are very effective for the sequence tasks, e.g. action recognition, speech recognition, natural language processing (NLP). Furthermore, a particular type of RNN, namely Long Short-Term Memory (LSTM) [12], is proposed to more effectively solve the weakness of long-term memory ability in RNN. For the EmotiW challenge, both spatial and temporal features are all considered to obtain better performance for the audio-video emotion recognition. Kahou et al. [6] combined CNN with RNN to model the spatio-temporal evolution of visual information in EmotiW2015. Yan et al. [24] employed a CNN and bidirectional RNN architecture to learn facial appearance texture in EmotiW2106. Meanwhile, Fan et al. [9] utilized LSTM with deep convolutional networks to train a model spatially and temporally for video sequences.

Different from composite structure of CNN-RNN, 3-Dimension Convolutional Neural Networks (3D CNN) [21] uses 3-dimension convolution operations instead of 2-dimension convolution in network structure. This architecture generates multiple channels of information from adjacent video frames, meanwhile it performs convolution and sub-sampling separately in each channel. The final feature representation is obtained by combining information from all channels. Currently, Convolutional 3D (C3D) network is widely applied in dealing with various video analysis tasks [14]. Fan et al. [9] also utilized C3D as the part of proposed hybrid network to encode the appearance and motion information in facial expression and got the champion in EmotiW2016. Vielzeuf et al. [22] proposed a novel descriptor combining C3D and LSTM as a segment of temporal multimodal fusion in EmotiW2017.

As well as facial image, audio in the video clips is also significant for the expression of emotional information. Some competitors [9, 15, 24] considered this factor and added it to their model combined with SVM classifier by the OpenSMILE [8] features of audio. Basically, nearly all audio of video clips is speech in EmotiW challenge. Furthermore, speech carries wealthy information about a person's identity, age, mood, health, etc. Indeed, speech effectively reflects the emotional state of the human from some characteristics of pitch, tone, and prosodic feature [7] . Therefore, the speech extracted from audio is an important aspect for the video emotion recognition.

Although several spatial-temporal models have been applied in EmotiW challenge by many competitors [9, 22, 24], they are almost taken as a part of overall model and always fused with other single type model (only spatial, temporal or others). Not only that, audio models are almost considered by using openSMILE feature, and they do not adequately describe the dynamic spatio-temporal information of emotions in audio. Therefore, in this paper, we propose a multiple spatio-temporal feature fusion (MSFF) framework to capture the dynamic variation of emotional information in two modality (facial image and audio) obtained from video clips.

The proposed framework is shown in Fig. 1. Firstly, referring to the facial images in video clips, three different types of neural networks, namely Resnet-50 pre-trained on ImageNet [3] and VGG-Face [18], are applied to extract diverse-level features containing the facial appearance texture in spatial dimension. Then, the features in spatial dimension are fed into a Bidirectional LSTM to capture the variation of emotional information for the better emotion recognition. Also, the complementary emotional features in spatial-temporal dimension are obtained via a C3D framework for model fusion. Furthermore, the segments of spectrogram images, which derived from the speech in video clips, are employed as the input of a VGG19 and Bidirectional LSTM architecture to generate the spatio-temporal features of another modality. Finally, we fuse the classification scores generated from different kinds of spatio-temporal neural networks architectures to promote the performance of emotion recognition. Benefiting from the effective representation via the spatial-temporal features for emotional information in video clips, our MSFF approach achieves a recognition accuracy of 60.64% which surpasses the baseline [1] and outperforms the winner of EmitiW2017 [13].

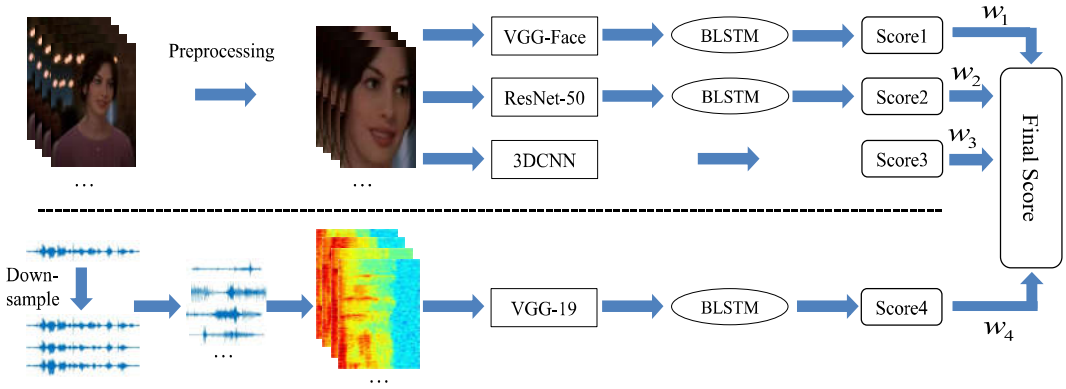

Figure 1: Proposed MSFF framework. The part in top of the figure is the facial image model, while the speech model locates in the bottom part of firgure. All Scores from these models are finally fused to predict the emotion.

The remainder of this paper is organized as follows. The facial appearance texture learning of the proposed framework is expounded in Section 2. In Section 3, audio model is illustrated. Then, a score fusion strategy is proposed for emotion recognition in Section 4. And Section 5 presents experiments to validate the proposed framework. Finally, Section 6 concludes the paper.

2 THE PROPOSED METHOD

The framework of proposed Spatio-temporal information fusion is shown in Fig. 1. The overview of the system has three parts: the facial image model, the audio model and the fusion strategy. These two different modalities play complementary roles in feature extraction of emotional information. The three parts are described in detail below.

2.1 The Facial Image Model

After splitting the video into frames, a large number of images are generated. For these images, face detection and alignment are applied to get the facial images, and the details of process are described in section of experiment. For facial images, facial appearance texture is a essential factor for the emotion recognition. Hence, two different architectures of neural networks, which are CNN-RNN and 3D CNN, are utilized to extract accurate spatial-temporal features about facial images in video. The architecture of CNN-RNN combines feature representation of CNN in the spatial dimension with dynamic information capture of RNN in temporal dimension. We uses two different depths CNN, namely the pre-trained VGG-Face and ResNet50 to extract various high-level features of facial appearance texture. And a special RNN, Bidirectional LSTM is also utilized to capture the variation of emotions in video. The other architecture of facial image model is 3D CNN. Different from the combination of CNN and RNN, 3D CNN is a special CNN to deal with video tasks, which directly obtain the spatio-temporal representation of video by the 3-dimension operations in convolution and pooling. These various high-level features and different types of spatio-temporal neural networks represent the facial appearance texture in video more accurately and complementally.

2.1.1 VGG-Face Network. VGG-Face model [20] is a special type of CNN with 16 layers, which is developed by Visual Geometry Group in Oxford University, and is evaluated on the Labeled Faces in the Wild and the YouTube Faces dataset [18]. Since its appearance, it has been turned out an excellent performance in face recognition. Therefore, the VGG-Face is adept to capture the discriminative features of facial images, which are beneficial to represent the facial expression by reason of the similarity between facial expression images and facial images.

Each video clip corresponds to lots of facial expression images after the data pre-processing. The cropped facial expression images are all resized to a standard size as the input of VGG-Face network. For the sake of robustness of the model, face expression images are augmented each time before sent to the network, such as flipping, mirroring, panning, random cropping, etc.

In order to obtain better recognition results, we fine-tune the pre-trained model in facial expression images from video clips, rather than training the VGG network from scratch. We fine-tune the network layers of VGG-Face from the last fully connected layer forward to the first convolutional layer on the processed facial expression database, and get the highest accuracy in fine-tuning the three fc layers. The reason has been given by Yosinski et al. [25] in Cornell University. They researched the transfer ability of neural network, which revealed that the first few layers have learned the general features. While, as the network level deepens, the latter network is more focused on learning task-specific features. Therefore, based on the VGG-Face model pre-trained on the face dataset, we freeze the weights of all convolutional layers and only update that in the fully connected layers in VGG-Face network by the reprocessing facial images.

2.1.2 ResNet-50 Network. A natural attribute of deep networks is the integration of low/mid/high-level features and classifiers in an end-to-end multi-layer fashion. The more stacked layers, the richer level features that can be extracted with more abstract and semantic information. With the increase of network depth, however, a notorious obstacle is vanishing or exploding in gradient, which hamper convergence from the beginning. He et al. [11] proposed

the deep residual nets (ResNet), which introduced the residual representation module and overcame the convergence of network.

ResNet-50 with 50 layers is applied to obtain richer level features than VGG nets in our experiment. Because of concerns on the training time, a deeper bottleneck architecture is also designed instead of the building block. Benefiting from lots of 1×1 convolutional kernels in the bottleneck block, the scale of parameters in network has been greatly reduced so that networks converge better.

As well as the above VGG-Face, we utilize the ResNet-50 model pre-trained on the ImageNet dataset [3], and also fine-tune the fully connected layers on the facial expression images.

2.1.3 Bidirectional LSTM. Both VGG-Face and ResNet-50 model are adept at extracting the features of facial appearance texture in spatial dimension, however, they all lack the ability to capture dynamic changes in temporal dimension. RNN [17] has an important advantage of capturing context information during the mapping process between input and output sequences. Unfortunately, the range of contextual information that standard RNNs can access is limited, hence the influence of the input in hidden layers declines continuously as the network continues to recur.

LSTM [12] is a variation of RNN, which is set an enhanced component called long short-term memory in RNN. The core of LSTM is the state of cell, therefore three gates (input gate, forget gate and output gate) are utilized to update the cell state such that the problem of long-term dependency in RNN could be effectively solved . For better characterization of the variation about emotion in facial appearance texture, a bidirectional LSTM (BLSTM) [10]is applied to combine the spatial features with temporal features. The inputs of BLSTM are one-direction frame sequences from each video clip and the reverse order of the sequence. Therefore, positive and negative sequences about facial emotion features in each video are generated, and these sequences are fed into the BLSTM. These features constituting the sequences are obtained from the last fully connected layer in VGG-Face or ResNet-50, and the dimension is 4096.

2.1.4 3D convolutional Neural Network. The image features extracted by CNN have a great performance in computer vision tasks, however, image based deep features are not directly suitable for videos due to the lack of motion modeling. Hence, in order to recover this defect, the usual approach is the hybrid of CNN and RNN where the information of spatial and temporal dimensions about video are considered simultaneously. Nevertheless, learning spatio-temporal features not only includes the framework of CNN-RNN, also has a simple and effective approach, namely C3D [21].

Compared with 2D CNN, C3D utilizes the operation of 3D convolution and 3D pooling to model the temporal information in videos. It is notable that the size of convolution kernel is $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$, and the pooling is $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$, especially the pool1 has kernel size of $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$. As the continual convolution and pooling, the information in temporal dimension is transferred to increased channels. Each fully connected layer is a vector of 4096 dimension.

In our experiments, we use the 3D CNN with similar architecture of VGG-16 except the 3D convolution and 3D pooling kernels, and resize the facial images to a proper size as the input. For the sake of better recognition accuracy, the fc layers are also fine-tuned as well as VGG-Face and ResNet-50 models.

2.2 The Audio Model

Audio is a significant part of video information, especially in EmotiW challenge, which expresses emotional information independently and can be regarded as the complement for the contaminated facial expression images. We extract the useful speech from video clips by filtering the futile part and removing the background noise. The resulting speech implies abundant emotional information, which is very useful for emotion recognition in video. For speech emotion recognition, the openSMILE features [8] are commonly utilized, while these features can not accurately characteristic the spatio-temporal information of speech since they are the combination of diverse speech features.

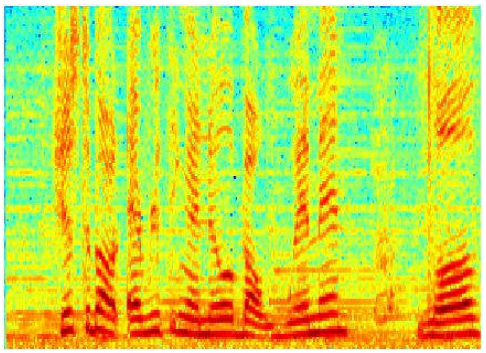

Figure 2: Spectrogram of the speech.

In order to extract the emotional features precisely, we use spectrogram image [19] by the short-time Fourier transform(STFT) to represent the signal variation in time and frequency domain , which is shown in Fig. 2. The voiceprint features are shown by many horizontal stripes, which contain rich information of different emotions, such as pitch frequency, format, spectrum envelope and so on [19].

In audio model, firstly, the speech is down-sampled at different sizes of sample rates to augment the data. Then, each speech of new sample rates is segmented into lots of frames with a overlap rate. Finally, the segment images are all resized to standard size as the input of audio model.

Therefore, we not only utilize the strong advantage in representation of image feature by CNN, also preferably capture the spatiotemporal information than openSMILE features through splitting a speech into segments as with dividing a video into frames. Eventually, the spectrogram images of segments for each speech are fed into a framework of VGG-19 and BLSTM. These two networks are introduced in the facial image model, where the VGG-19 is a VGG network with 19 layers and pre-trained on the speech emotion database.

2.3 The Fusion Strategy

In the previous sections, we describe the two models of our MSFF framework in details, which are the facial image model and audio model. For these models, spatio-temporal features of facial

appearance texture and speech are well described. In order to more adequately integrate the two modality, we employ a decision fusion approach to promote the performance of emotion recognition complementally. Specifically, each model will generate a score matrix in the phase of prediction, which donating the corresponding probability of each class in seven emotions. For the better recognition performance, a proper weight is assigned to evaluate the contribution of each model. Hence, the final score matrix $\hat{S}$ is formulated as follows,

$$\hat{S} = \sum_{i=1}^{m} \omega_i \cdot S_i \tag{1}$$

where ω_i and S_i are the weight and score of the ith model respectively. To specify the weight of each model, a grid search strategy is adopted to obtain the proper weight for the fusion optimization on the validation dataset. Obviously, the weight for each model in the fusion framework corresponds the contribution to performance of video emotion recognition.

3 EXPERIMENT

In this section, the details of experiment are provided and the results are also shown and discussed.

3.1 Data Preparation

The dataset of Audio-video Sub-challenge in EmotionW2018 [5] is collected from movies and TV reality shows, which contains 773, 383, 653 video clips in training, validation, testing database respectively. As a matter of fact, the majority frames from video clips have been contaminated by the abnormal conditions such as illumination, occlusion, scale change and so on. Therefore, it is hard to utilize these images for the emotion recognition directly.

For the sake of effective facial images, we must preprocess the image framed from the video clips. Specifically, MTCNN [26] is firstly applied to detect the face in images which contains many background objects and other irrelevant information, and the positions of 5 landmark points in face image are also obtained. Then, we employ the identification block of the SeetaFace recognition system [23] to normalize the face image with 5 landmark points and get the pure facial images with the size of 256×256. Finally, We crop the proper image size from pure face images as needed to input the facial image model, where the input image sizes of VGG-Face, ResNet50 and C3D are 224×224, 224×224, 112×112, respectively.

Meanwhile, the audio is also obtained from video clips. Removing the irrelevant audio and background noise, clear speech is ready to down-sample four scales of sample rates, which are 1.0, 0.75, 0.5, 0.25 respectively. Then, these speech is segmented into the size of 300ms with the frame overlap rate of 50%. After such processing, each speech consists of lots of segments like video frames, hence we can utilized the spatio-temporal network to implement audio model as what we do in video tasks. Finally, the segments of each speech are transformed the spectrogram images by SIFT and resized to 224×224 for the input of speech model.

3.2 Parameter Setting

The MSFF framework includes facial image model and audio model. The facial image model is consisted of two architectures spatio-temporal networks: CNN-RNN (VGG-Face + BLSTM, ResNet-50 + BLSTM) and C3D, while the audio model contains one CNN-RNN (VGG-19 + BLSTM). For all types of CNN-RNN model in proposed framework, they are all fed with video consequences where the length of each video frames is set to 8, while C3D uses 16 frames as the input of network. Notably, although both two models in the MSFF framework use the video frames as the input of spatio-temporal networks, different data augmentation methods are adopted to vary dataset. In facial image model, we utilize the data augmentation methods containing flipping, mirroring, panning, random cropping, while Gaussian noise and salt-pepper noise are added to the spectrogram images in audio model. After the data augmentation, we resize the image from the scale of 256×256 to 224×224 and 112×112 as the input of RNN-CNN and C3D, respectively. The 4096 dimensions features from CNN feed into BLSTM with two layers where the number of hidden nodes is 256.

In addition, the batch sizes of the input video clips are all set to 8, and the learning rates are 0.01 with the decay coefficient of 0.95 in both facial image model and audio model,.

3.3 Result and Discussion

For the audio-video sub-challenge, we propose different types of spatio-temporal networks to extract more discriminative features for the better performance of video emotion recognition. In order to verify the effect of these saptio-temporal features, the pure CNN architecture, VGG-Face and ResNet50 are compared with the spatio-temporal networks, VGG-Face + BLSTM and ResNet50 + BLSTM. The results on the validation dataset reveal that the recognition rate of VGG-Face, ResNet50, VGG-Face + BLSTM, ResNet50 + BLSTM are 50.32%, 48.61%, 53.91%, 49.31% respectively. Compared with spatial features, it is obviously that the spatio-temporal features are more effective for the video emotion recognition.

For the better architectures of spatio-temporal network , we select four types of networks to compare their performance, which contain VGG-Face+BL-STM, ResNet-50+BLSTM, C3D, VGG-19+BL-STM. The performance of them on the validation set is shown in Table. 1. The results reveal that the validation of VGG-Face+BLSTM model has best accuracy of 53.91%, since this model is pre-trained on the facial image datasets such that it can capture the spatio-temporal information from facial images in video clips when the challenge datasets are utilized to fine-tune this model. The audio

Table 1: Recognition accuracy of each model on the validation and test datasets.

Models	Validation(%)	Test(%)
VGG-Face+BLSTM	53.91%	-
ResNet-50+BLSTM	49.31%	-
C3D	39.36%	-
VGG-19+BLSTM	30.81%	-
Fusion	56.05%	60.64%

Table 2: Recognition accuracy of all submissions

Submission	Fusion Models	Validation(%)	Test(%)
1	{VGG-16+BLSTM, ResNet-50+BLSTM}	48.19	49.46
2	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D, VGG-19+BLSTM}	56.05	**60.64**
3	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D, VGG-19+BLSTM}*	-	59.57
4	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D, openSMILE-feature+SVM}	59.42	59.26
5	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D, openSMILE-feature+SVM}*	-	58.19
6	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D, VGG-19+BLSTM, openSMILE-feature+SVM}	57.89	58.81
7	{VGG-Face+BLSTM, VGG-Face2+BLSTM, ResNet-50+BLSTM, C3D, VGG-19+BLSTM, openSMILE-feature+SVM}	58.50	60.18
-	{VGG-Face+BLSTM, ResNet-50+BLSTM, C3D}	54.85	-
-	{VGG-Face+BLSTM, ResNet-50+BLSTM, VGG-19+BLSTM}	55.13	-

model consists of VGG-19 and BLSTM. Although we have preprocessed the speech, they still have some flaws, such as shortage of speech length, lack of semantic information, etc. Hence, the performance of speech model is not good enough. Then, we use the proposed fusion strategy to obtain the validation accuracy of 56.05% and test accuracy of 60.64% which is our best result in all submissions. These results also reveal that our fusion strategy indeed boosts the performance of emotion recognition.

The results of fusion networks are all listed in the Table. 2. The first seven results are submitted, while the last two times are the unsubmitted results, hence there are no test dataset accuracy. The $*$ indicates that models are trained with both the training and validation datasets so that validation accuracy are not provided. Hence, we employ 3-fold cross validation to tune our models including the configuration of proposed models and the fusion weights of scores. The result of second submission and last two fusion networks show that both audio-visual fusion and different face models fusion boost the performance of video emotion recognition. Notably, in Table 2, the VGG-Face2 model is a updating version of VGG-Face with the architecture of VGG-16 pre-trained in the new face datasets. In order to compare the hand-crafted feature with the spatiotemporal features of speech, the openSMILE features of 384 dimensions are also extracted to fusion with the facial image model. The results reveal that the spatiotemporal features outperform the openSMILE features in the fusion recognition accuracy. In the last two submissions, the openSMIL+SVM and VGG-Face2+BLSTM models are respectively added to the fusion framework, and finally we achiever close result of 60.18% with our best result of 60.64%. The fusion weights of best result obtained on the validation dataset are 0.6190, 0.0239, 0.0476, 0.3095, corresponding to networks in the Table. 2.

According to the confusion matrix of the best result shown in Fig. 3, three emotions of angry, happy and neutral are effortless to recognize , while and the surprise and disgust are hardly to discriminate, especially for the disgust, our fusion model has not yet been able to classify it. Due to the serious imbalance of the samples in the competition data set, the numbers of samples in disgust and surprise are less than other emotions. Consequently, the class imbalance makes the model be overwhelmed by those majority classes and thus degrades the performance.

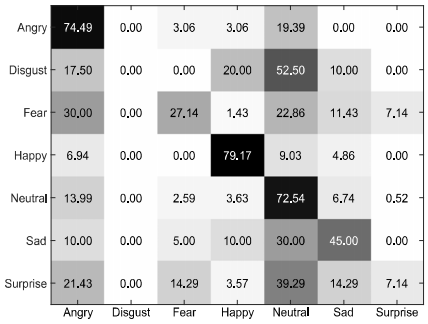

Figure 3: Confusion matrix of the 8th submission.

4 CONCLUSION

In this paper, we presented a multiple spatio-temporal feature fusion framework to deal with the video emotion recognition in the wild in EmotiW2018. Emotions in video clips are characterized from two aspects including facial appearance textures, and speech. The proposed facial image model effectively model the dynamic changes of facial textures in emotions. For speech signals in the video, the spectrogram images are extracted and then fed into a specified VGG-19 and BLSTM network to capture the spatio-temporal features. Finally, we combine the previous scores together by considering their complementarity in the emption recognition. The results reported in the challenge indicate that our proposed MSFF framework is more promising on the task of emotion recognition in the wild.

5 ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China under Grants 2015CB351704, the National Natural Science Foundation of China under Grants 61572009, the Jiangsu Provincial Key Research and Development Program under Grant BE2016616.

REFERENCES

[1] Roland Goecke Abhinav Dhall, Amanjot Kaur and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction(in press). (2018).

[2] Roddy Cowie, Ellen Douglascowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2002. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (2002), 32–80.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[4] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary. In *ACM on International Conference on Multimodal Interaction*. 371–372.

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[6] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 467–474.

[7] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[9] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[10] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 553–560.

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 221–231.

[15] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. 2017. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598* (2017).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[17] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association,InterSpeech*.

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition. In *BMVC*, Vol. 1. 6.

[19] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. 1995. Speech recognition with primarily temporal cues. *Science* 270, 5234 (1995), 303–304.

[20] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[22] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 569–576.

[23] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan, and Xilin Chen. 2017. Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing* 221 (2017), 138–145.

[24] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, Yuan Zong, and Ning Sun. 2016. Multi-clue fusion for emotion recognition in the wild. In *ACM International Conference on Multimodal Interaction*. 458–463.

[25] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[26] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.