Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition

Hengshun Zhou1, Debin Meng2, Yuanyuan Zhang1, Xiaojiang Peng1,2, Jun Du†, Kai Wang2, Yu Qiao2

1National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP), University of Science and Technology of China, P.R. China
2ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

ABSTRACT

The audio-video based emotion recognition aims to classify a given video into basic emotions. In this paper, we describe our approaches in EmotiW 2019, which mainly explores emotion features and feature fusion strategies for audio and visual modality. For emotion features, we explore audio feature with both speech-spectrogram and Log Mel-spectrogram and evaluate several facial features with different CNN models and different emotion pretrained strategies. For fusion strategies, we explore intra-modal and cross-modal fusion methods, such as designing attention mechanisms to highlight important emotion feature, exploring feature concatenation and factorized bilinear pooling (FBP) for cross-modal feature fusion. With careful evaluation, we obtain 65.5% on the AFEW validation set and 62.48% on the test set and rank third in the challenge.

CCS CONCEPTS
• Computer systems organization → Embedded systems; Redundancy; Robotics; • Networks → Network reliability.

KEYWORDS
Emotion Recognition; Attention Mechanism; Deep learning; Affective Computing; Convolutional Neural Networks

ACM Reference Format:

1 INTRODUCTION

Emotion recognition (ER) has attracted increasing attention in academia and industry due to its wide range of applications such as human-computer interaction [7], clinical diagnosis [19], and cognitive science [14]. Although great progress in the face and video analysis has been made [4, 23, 26–29], audio-video emotion recognition in the wild remains a challenging problem due to the expression suffers from the large pose, illumination variance, occlusion, motion blur, etc.

Audio-Video emotion recognition can be summarized as a simple pipeline shown in Fig 1, which includes four parts, namely Video preprocessing, Feature Extraction, Feature Fusion, and Classifier. Specifically, video preprocessing refers to extract the spectrogram of the audio, the faces or landmarks of video. Feature extraction and feature fusion respectively extracts emotion features from the audio or visual signal and fuses emotion features into compact feature vectors, which are subsequently fed into a classifier for prediction.

Reviewing the methods of Audio-Video emotion recognition, we find that some methods emphasize feature extraction and other methods emphasize feature fusion. Yao et al [31] construct Holonet as discriminative feature extraction, which combines residual structure [12] and CReLU [22] to increase network depth and maintain efficiency. The EmotiW 2017 winner team [13] gets robust feature extraction with Supervised Scoring Ensemble (SSE) which adds supervision to intermediate layers and shallow layers. Since SSE only uses high-level representations, Fan et al [8] further improve SSE by utilizing middle feature maps to provide more discriminative features. These methods mainly use average pooling to obtain video-level representation from frame-level.

Many feature fusion strategies have been used in previous EmotiW challenges. [9, 18, 25] extract CNN-based frame features and use LSTM[10] or BLSTM[11] to fuse them. [1, 15, 17] use Statistical encoding module to aggregate frame features which compute the mean, variance, minimum, and
maximum of the frame feature vectors. However, these methods ignore the importance of frames. Besides, all previous methods mainly apply score averaging or feature concatenation for audio-video fusion, which ignores the correlation between the features from different modalities.

In this paper, we exploit three types of intra-modal fusion methods, namely self-attention, relation-attention, and transformer [24]. They are used to learn weights for frame features to highlight important frames. For cross-modal fusion, we explore feature concatenation and factorized bilinear pooling (FBP) [32]. Besides, we evaluate different emotion features, including convolutional neural networks (CNN) for audio information with both speech-spectrogram and Log Mel-spectrogram and several facial features with different CNN models and different emotion pretrained strategies. Finally, we obtain 62.48% and rank third in the challenge.

Our contributions and finds can be summarized as follows.

- We experimentally show that better face recognition CNN models and choosing suitable emotion datasets to further pretrain the face CNN models is important.
- We design three kinds of attention mechanisms for visual and audio feature fusion.
- We apply a Factorized Bilinear Pooling (FBP) for cross-modal feature fusion.

2 THE PROPOSED METHOD

We develop our ER system based on the pipeline of Video preprocessing-Feature Extraction-Feature Fusion-Classifier.

Video preprocessing

Face detection and alignment. We apply face detection and alignment by Dlib toolbox 1. We extend the face bounding box with a ratio of 30% and then resize the cropped faces to scale of 224 × 224. We do not apply face detection and alignment for AffectNet dataset, due to the face bounding box had been provided. For AFEW dataset, If no face is detected in the picture, the entire frame is passed to the network.

Audio processing and Spectrogram calculation. For each audio, the speech spectrogram and log Mel-spectrogram extraction process is consistent with [32] and [3] respectively. For speech spectrogram, we use the Hamming window with 40 msec window size and 10 msec shift. Finally, the 200-dimensional low-frequency part of the spectrogram is used as the input to the audio modality. As for log Mel-spectrogram, we calculate its deltas and delta-deltas.

Feature Extraction

Visual Features. We apply three CNN backbones to extract facial emotion features, namely VGGFace, ResNet18, and IR50 [4]. The dimensions are 4096, 512, and 512, respectively.

Audio Feature. We extract the feature maps of the audio from the last Pooling layer of AlexNet. The size of a 3-dimensional feature map is $H \times W \times C$, where the $H(W)$ is the height(width) of the feature map, and $C$ is the number of the channel of the feature map. The feature maps are then split into $n$ vectors($n = H \times W$). Each vector is C-dimensional.

Intra-modal Feature Fusion

We apply the attention-based strategies for intra-modal feature fusion. It converts a variable number of emotion features(from audio or visual modality) into a fixed-dimension feature. We explore three attention methods, namely Self-attention, Relation-attention, and Transformer-attention. Formally, we denote a number of emotion features as $\{f_1, \cdots, f_n\}$.

Self-attention. We apply 1-dimensional Fully-Connected(FC) layer $W_0^{d \times 1}$ and a sigmoid function $\sigma$ for each emotion feature, the weight of the $i$-th feature $f_i^T$ is defined by:

$$\alpha_i = \sigma(f_i^T \cdot W_0^{d \times 1})$$ (1)
We formulate the attention weight as follows:

\[ f_i = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i} \]  

(2)

**Relation-attention.** This attention module was designed to learn weights from the relationship between features. After the self-attention, features are aggregated into a single vector \( f_i \). Since \( f_i \) inherently contains global representation of these features, we use the sample concatenation of individual features and global representation \([f_i : f_i]\) to model the global-local relation. Similar to the Self-attention module, with individual emotion features, we apply 1-dimensional FC layer \( W_1^{d \times 1} \) and a sigmoid function \( \sigma \). The relation-attention weight of the \( i \)-th feature \([f_i : f_i]^T\) is formulated as follows:

\[ \beta_i = \sigma([f_i : f_i]^T \cdot W_1^{d \times 1}) \]  

(3)

With Self-attention and Relation-attention weights, all the emotion features was convert into a new feature as follows:

\[ f_i = \frac{\sum_{i=0}^{n} \alpha_i \beta_i [f_i : f_i]}{\sum_{i=0}^{n} \alpha_i \beta_i} \]  

(4)

**Transformer-attention.** Inspired by the works in[32] and [30], we formulate the attention weight as follows:

\[ f_i' = W_{md}^{m \times d} \cdot f_i + b \]  

(5)

\[ y_i = \exp(u_{d \times 1} \cdot \tanh(f_i')) \]  

(6)

To reduce the dimension of the feature \( f_i \), we use a \( w \times d \)-dimensional FC layer \( W_2^{m \times d} \) in Eq.(5). Then the weight of the \( i \)-th feature \( f_i \) is processed by a 1-dimensional FC layer \( u^i \), \( \exp() \) and \( \tanh() \) function in Eq.(6).

With these transformer-attention weights, we aggregate all the emotion features into a single feature \( f_t \) as follows:

\[ f_t = \frac{\sum_{i=1}^{n} y_i f_i}{\sum_{i=1}^{n} y_i} \]  

(7)

**Cross-modal Feature Fusion**

We apply **Factorized Bilinear Pooling**(FBP) for cross-modal feature fusion. Given two features in different modalities, i.e. the audio feature vector \( a \in \mathbb{R}^m \) for a spectrogram and visual feature \( \mathbf{v} \in \mathbb{R}^n \) for frame sequence, the simplest cross-modal bilinear model is defined as follows:

\[ z_i = a^T W_i v \]  

(8)

where \( W \in \mathbb{R}^{m \times n} \) is a projection matrix, \( z_i \in \mathbb{R} \) is the output of the bilinear model. We use the Eq.(9) to obtain the output feature \( z = [z_1, \cdots, z_o] \). The formula derivation from formula Eq.(8) to Eq.(9) was discribed in the paper[32].

\[ z = [z_1, \cdots, z_o] = \text{SumPooling}(\tilde{U}^T a \circ \tilde{V}^T v, k) \]  

(9)

The implementation of Eq(9) is illustrated in Fig2, where \( \tilde{U}^T a \) and \( \tilde{V}^T v \) are implemented by feeding feature \( a \) and \( \mathbf{v} \) to FC layers, respectively, and the function SumPooling\((x, k)\) applies sum pooling with non-overlapped windows to \( x \). Besides, Dropout is adopted to prevent over-fitting. The L2-normalization \((x \leftarrow x/\|x\|)\) is used to normalize the energy of \( z \) to avoid the dramatical variation of the output magnitude, due to the introduced element-wise multiplication.

### 3 EXPERIMENTS

**Dataset**

In this work we use four emotion datasets to train our models, i.e. AffectNet[20], RAF-DB[16], FER+[2], AFEW[5, 6].

The human-annotated part of AffectNet dataset contains 287,651 training images and 4,000 test images, which are annotated with both emotion labels and arousal valence values. Only emotion labels are used in this task.

The RAF-DB dataset consists of 15,339 images labeled with 7-class basic emotion and 3,954 labeled with 12-class compound emotion. Only images labeled with basic emotion are used in this study.

The FER+ dataset contains 28,709 training, 3,589 validation and 3,589 test images. We combine its training data with validation data for the training split and evaluate the model performance on the test data.

The AFEW contains 773 train, 383 val and 653 test samples, which are collected from movies and TV serials with spontaneous expressions, various poses, and illuminations.

**Exploration of Emotion Features**

We explore emotion features in two perspectives, namely CNN backbones and pretraining emotion datasets.

For the choice of the CNN model, we compare IR50[4], ResNet18[12], and VGGFace[21] in the Table 1, where the former two models are pretrained on MS-Celeb-1M dataset and the last one on VGGFace dataset. We find that the large CNN, IR50, is superior to the other two models.

---

Figure 2: Our factorized bilinear pooling(FBP) module.
We use the well-trained IR50 model to extract features and only train softmax classifier using these features. The IR50 models pre-trained on FER+, RAF-DB, and AffectNet achieve 50.13%, 51.436%, and 53.78%, respectively. Therefore, we choose the IR50 model pretrain on AffectNet as our visual features in the following fusion experiments.

Table 1: Exploration of CNN models and pretrained emotion datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>FER+</th>
<th>RAF-DB</th>
<th>AffectNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGFace</td>
<td>88.84%</td>
<td>86.93%</td>
<td>51.425%</td>
</tr>
<tr>
<td>ResNet18</td>
<td>88.65%</td>
<td>86.96%</td>
<td>52.075%</td>
</tr>
<tr>
<td>IR50</td>
<td>89.257%</td>
<td>89.075%</td>
<td>53.925%</td>
</tr>
</tbody>
</table>

Exploration of Fusion Strategies

We explore three intra-modal attention strategies with the FBP cross-modal fusion. We use speech spectrogram for audio CNN, which obtains 38% on AFEW validation set individually. In the Table 2, we find the FBP improves performance for all the intra-modal fusion methods. Transformer attention for intra-modal fusion is the best for FBP.

Table 2: Evaluation of intra-modal fusion methods.

<table>
<thead>
<tr>
<th>Audio</th>
<th>Visual</th>
<th>Self</th>
<th>Relation</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>54.6%</td>
<td>56.9%</td>
<td>60.3%</td>
<td></td>
</tr>
<tr>
<td>Relation</td>
<td>54.0%</td>
<td>57.2%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Transformer</td>
<td>54.8%</td>
<td>58%</td>
<td>61.1%</td>
<td></td>
</tr>
</tbody>
</table>

We also use log Mel-spectrogram for audio CNN, which obtains a little better performance, but the final results are very similar after intra- and cross-modal fusion. Besides, the concatenation of audio and visual vectors gets 58% accuracy in AFEW validation set with transformer attention. This is 3% lower than FBP which shows the effectiveness of FBP.

Feature Enhancement

In the Table 3, the Basic Features means that we only extract one feature vector for each frame. Besides, we apply 5 kinds of feature enhancement strategies as presented in Table 3. Specifically, for feature F-Mean, we first obtain 18 transformation frames by using three rotations, three scales, and flipping for a frame. After that, we compute the features of these 18 transformation frames and average these 18 features as the feature F-Mean. For the feature F-MeanStd, we compute the average feature and feature standard deviation of these 18 features. We then concatenate the average feature and the standard deviation as F-MeanStd. For the feature F-normFFT, we first compute the Fast Fourier transform(FFT) of the Basic Feature, and then normalize the feature and concatenate the real and imaginary parts as F-normFFT. For the feature F-AR-Mean, A means that the features are extracted by the models pre-trained on Affectnet, and R by the models pre-trained on RAF-DB, we concatenate these two mean features of two different pretrained models as F-AR-Mean.

Table 3: Evaluation of five feature enhancement strategies. The default setting is Rotation ∈ [−15°, 15°], scale ∈ [0.75, 1, 1.25]

<table>
<thead>
<tr>
<th>Visual Feature</th>
<th>Augmentation details</th>
<th>AFEW Val acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Feature</td>
<td>---</td>
<td>61.1%</td>
</tr>
<tr>
<td>Basic Feature</td>
<td>---</td>
<td>58.5%</td>
</tr>
<tr>
<td>F-Mean</td>
<td>default setting</td>
<td>62.14%</td>
</tr>
<tr>
<td>F-MeanStd</td>
<td>default setting</td>
<td>63.7%</td>
</tr>
<tr>
<td>F-MeanStd-2</td>
<td>Rotation ∈ [−15°, 15°], scale ∈ [0.75, 1, 1.25]</td>
<td>62.4%</td>
</tr>
<tr>
<td>F-NormFFT</td>
<td>Normalized FFT</td>
<td>61.35%</td>
</tr>
<tr>
<td>F-AR-Mean</td>
<td>default setting</td>
<td>62.92%</td>
</tr>
<tr>
<td>FG-Net</td>
<td>---</td>
<td>59%</td>
</tr>
</tbody>
</table>

Table 3 shows that the five feature enhancement methods further improve the performance of FBP where the feature F-MeanStd achieves the best result on the validation set.

Table 4: Submission results of different model combinations.

<table>
<thead>
<tr>
<th>Sub</th>
<th>Val</th>
<th>Test</th>
<th>Fusion detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>---</td>
<td>62.481%</td>
<td>4 FG-Net-1</td>
</tr>
<tr>
<td>(2)</td>
<td>---</td>
<td>59.112%</td>
<td>2 F-MeabStd-2 + 2 F-AR-Mean</td>
</tr>
<tr>
<td>(3)</td>
<td>---</td>
<td>54.518%</td>
<td>4 FG-Net-2</td>
</tr>
<tr>
<td>(4)</td>
<td>64.5%</td>
<td>61.41%</td>
<td>4 F-MeanStd</td>
</tr>
<tr>
<td>(5)</td>
<td>65.5%</td>
<td>62.328%</td>
<td>F-Mean + F-MeanStd + F-NormFFT + F-MeanStd-2 + F-AR-Mean</td>
</tr>
</tbody>
</table>

Results On EmotiW2019

In the Table 4. The first three submitted models are trained on the training and validation set of AFEW, and the last two models are trained on the training set of AFEW. We find that it is difficult to choose models and fuse models if combining the validation set with the training set. We adopt class weight in all submissions, which means that we re-weight the predicted scores by the square root of the sample numbers([0.15, 0.097, 0.129, 0.185, 0.138, 0.082, 0.215]).

4 CONCLUSIONS

In this paper, we exploit three types of intra-modal fusion methods, namely self-attention, relation-attention, and transformer. They are mainly used to highlight important emotion
feature. For the fusion of audio and visual information, we explore feature concatenation and factorized bilinear pooling (FBP). Besides, we evaluate different emotion features, including an audio feature with both speech-spectrogram and Log Mel-spectrogram and several facial features with different CNN models and different emotion pretrained strategies. With careful evaluation, we obtain 62.48% and rank third in the EmotiW 2019 Challenge.

5 ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (U1613211), Shenzhen Basic Research Program (JCYJ20170818164704758), the Joint Lab of CAS-HK.

REFERENCES


