Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition

Shiqing Zhang, Shiliang Zhang, *Member, IEEE*, Tiejun Huang, *Senior Member, IEEE*, Wen Gao, *Fellow, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*— Emotion recognition is challenging due to the emotional gap between emotions and audio–visual features. Motivated by the powerful feature learning ability of deep neural networks, this paper proposes to bridge the emotional gap by using a hybrid deep model, which first produces audio–visual segment features with Convolutional Neural Networks (CNNs) and 3D-CNN, then fuses audio–visual segment features in a Deep Belief Networks (DBNs). The proposed method is trained in two stages. First, CNN and 3D-CNN models pre-trained on corresponding large-scale image and video classification tasks are fine-tuned on emotion recognition tasks to learn audio and visual segment features, respectively. Second, the outputs of CNN and 3D-CNN models are combined into a fusion network built with a DBN model. The fusion network is trained to jointly learn a discriminative audio–visual segment feature representation. After average-pooling segment features learned by DBN to form a fixed-length global video feature, a linear Support Vector Machine is used for video emotion classification. Experimental results on three public audio–visual emotional databases, including the acted RML database, the acted eNTERFACE05 database, and the spontaneous BAUM-1s database, demonstrate the promising performance of the proposed method. To the best of our knowledge, this is an early work fusing audio and visual cues with CNN, 3D-CNN, and DBN for audio–visual emotion recognition.**

***Index Terms*— Emotion recognition, deep learning, convolutional neural networks, deep belief networks, multimodality fusion.**

I. Introduction

RECOGNIZING human emotions with computers is usually performed with a multimodal approach due to the inherent multimodality characteristic of human emotion expression. Speech and facial expression are two natural and effective ways of expressing emotions when human beings communicate with each other. During the last two decades, audio-visual emotion recognition integrating speech and facial expression, has attracted extensive attention owing to its promising potential applications in human-computer-interaction [1], [2]. However, recognizing human emotions with computers is still a challenging task because it is difficult to extract the best audio and visual features characterizing human emotions.

Feature extraction and multimodality fusion are two key steps for audio-visual emotion recognition. As far as feature extraction is concerned, there has been a number of works [3]–[14] focusing on extracting low-level hand-crafted features for audio-visual emotion recognition. Nevertheless, due to the emotional gap between human emotions and low-level hand-crafted features, these hand-crafted features can not sufficiently discriminate human emotions. Here, the “emotional gap” is defined as “the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective states in which the user is brought by perceiving the signal” [15]. Therefore, the “emotional gap” essentially represents the differences between emotions and the extracted affective features. To bridge the “emotional gap”, it is desirable to extract high-level audio and visual features effectively distinguishing emotions.

After feature extraction, multimodality fusion is employed to integrate audio and visual modalities for emotion recognition. Previous works [3], [8], [12], [16]–[18] focus on four typical fusion strategies: feature-level fusion, decision-level fusion, score-level fusion, and model-level fusion, respectively. Although most of existing fusion methods exhibit good performance on audio-visual emotion recognition tasks, they belong to shallow fusion models with a limited ability in jointly modeling highly non-linear correlations of multiple inputs with different statistical properties [19]. It is thus needed to design deeper fusion methods to produce a more optimized joint discriminant feature representation for audio-visual emotion recognition.

To alleviate above-mentioned two problems, the recently-emerged deep leaning [20] techniques may present a cue. Due to the large-scale available training data and the effective training schemes, deep learning techniques have exhibited powerful feature learning ability in a wide variety of domains, such as speech recognition, image processing and understanding, object detection and recognition, *etc*. Among them, two representative deep learning models are DBN [21] and CNN [22], [23], as described below.

Manuscript received December 29, 2016; revised April 24, 2017 and May 24, 2017; accepted June 20, 2017. Date of publication June 23, 2017; date of current version October 24, 2018. This work was supported in part by the National Science Foundation of China and Zhejiang Provincial National Science Foundation of China under Grant 61572050, Grant LY16F020011, Grant 91538111, Grant 61620106009, and Grant 61429201. The work of Q. Tian was supported in part by ARO under Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards by NEC Laboratories of America and Blippar. This paper was recommended by Associate Editor L. Lin.

Shiqing Zhang is with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China (e-mail: tzczsq@pku.edu.cn).

Shiliang Zhang, T. Huang, and W. Gao are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: slzhang.jdl@pku.edu.cn; tjhuang@pku.edu.cn; wgao@pku.edu.cn).

Q. Tian is with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2719043

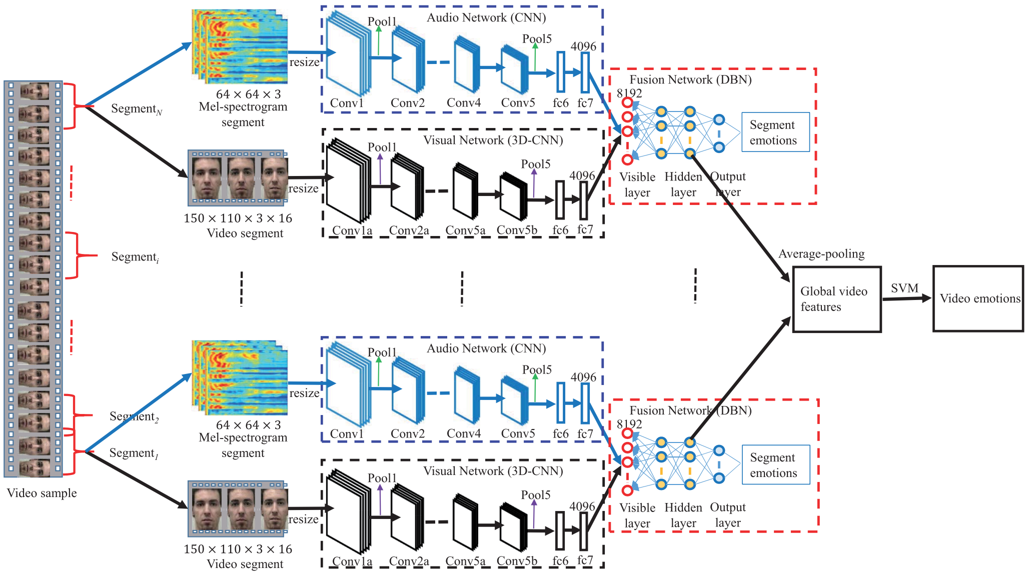


Fig. 1. The structure of our proposed hybrid deep model for audio-visual emotion recognition.

DBNs are built by stacking multiple Restricted Boltzmann Machines (RBMs) [24]. By using multiple RBMs and the greedy layer-wise training algorithm [21], DBNs can effectively learn a multi-layer generative model of input data. Based on this generative model, the distribution properties of input data can be discovered, and the hierarchical feature representations characterizing input data can be also extracted. Due to such good property, DBNs and its variant called Deep Bolzmann Machines (DBMs) [25] have been successfully utilized to learn high-level feature representations from low-level hand-crafted features for multimodal emotion recognition [26]–[28].

CNNs employ raw image data as inputs instead of hand-crafted features. CNNs are mainly composed of convolutional layers and fully connected layers, where convolutional layers learn a discriminative multi-level feature representation from raw inputs and fully connected layers can be regarded as a non-linear classifier. Due to the large-scale available training data and the effective training strategies introduced in recent works, CNNs have exhibited significant success in various vision tasks like object detection and recognition [29]–[32]. However, such CNN models are mostly applied on 2-D images and fail to capture motion cues in videos, due to the usage of 2-D spatial convolution. To address this problem, a recent work [33] has extended CNNs with deep 3-D convolution to produce a 3D-CNN model. 3D-CNNs compute feature maps from both spatial and temporal dimensions, and have exhibited promising spatial-temporal feature learning ability on video classification tasks [33].

Inspired by the powerful feature learning ability of deep models, this work proposes a hybrid deep learning framework composed by CNN, 3D-CNN, and DBN to learn a joint audio-visual feature representation for emotion classification. Fig. 1 presents the structure of our framework. It is comprised of three steps: (1) we convert the raw audio signals into a representation similar to the RGB image as the CNN input. Consequently, a deep CNN model pre-trained on large-scale ImageNet dataset can be fine-tuned on audio emotion recognition tasks to learn high-level audio segment features. (2) For multiple contiguous frames in a video segment, a deep 3D-CNN model pre-trained on the large-scale video dataset is fine-tuned to learn visual segment features for facial expression recognition. (3) The audio and visual segment features learned by CNN and 3D-CNN are integrated in a fusion network built with a deep DBN, which is trained to predict correct emotion labels of video segments. Finally, we adopt the outputs of the last hidden layer of a DBN as the audio-visual segment feature. Average-pooling is employed to aggregate all segment features to form a fixed-length global video feature. Then, a linear SVM is used for video emotion classification. Note that, a DBN is used to fuse audio-visual features, rather than to classify the emotion of the whole video sample.

Learning audio-visual features for emotion recognition is one of the critical steps in bridging the emotional gap. Previous works focus on using low-level hand-crafted features, which have been verified not discriminative enough to human emotions. In contrast, this work aims at automatically learning a joint audio-visual feature representation from raw audio and visual signals using a hybrid deep learning framework. The hybrid deep learning model converts the raw 1-D audio signals into a 2-D representation and integrates CNN, 3D-CNN, and DBN for audio-visual feature learning and fusion. It thus also presents a new method of transforming 1-D audio signals into the suitable input of CNN that conventionally processes 2-D or 3-D images. Experimental results indicate that our learned features present promising performance. The success of this work also guarantees further investigation in this direction.

The remainder of this paper is organized as follows. The related works are reviewed in Section II. Section III present our hybrid deep learning model for affective feature learning in detail. Section IV describes the experimental results. The conclusions and future work are given in Section V.

II. Related Work

An audio-visual emotion recognition system generally consists of two important steps: feature extraction and multimodality fusion. In the following parts, we review related works focusing on these two steps, respectively.

A. Feature Extraction

The widely-used audio affective features can be categorized into prosody features, voice quality features, and spectral features [34], respectively. Pitch, intensity, energy, and duration time are popular prosody features, as they are able to reflect the rhythm of spoken language. The representative voice quality features include formants, spectral energy distribution, harmonics-to-noise-ratio, and so on. Mel-frequency Cepstral Coefficient (MFCC) is the most well-known spectral features since it is used to model the human auditory perception system. Zeng *et al.* [4] extract 20 audio features including pitch, intensity, and the first four formants and their bandwidths for audio emotion recognition. Wang and Guan [3] and Wang *et al.* [16] employ pitch, intensity, and the first 13 MFCC features on audio feature extraction tasks. Similar works, which extract prosody features, voice quality features and spectral features for audio emotion recognition, can be also found in [12] and [35]–[38], respectively.

Visual feature extraction methods can be summarized into two categories according to the format of inputs that are static or dynamic [39], [40]. For static images, the well-known ones are appearance-based feature extraction methods. These methods adopt the whole-face or specific regions in a face image to describe the subtle changes of the face such as wrinkles and furrows. Among them, Gabor wavelet representation [3], [16], [37], [41], Local Binary Patterns (LBP) [42] and its variants such as Local Phase Quantization (LPQ) [12], [43] are two representative appearance-based feature extraction methods. Wang and Guan [3] and Wang *et al.* [16] adopt a Gabor filter bank of 5 scales and 8 orientations to extract high-dimensional Gabor coefficients from each facial image. In recent years, CNNs are used to extract facial features from static facial images as visual features [44]. Ding *et al.* [44] use a Long Short-Term Memory (LSTM) [45] for audio emotion recognition, and a CNN for video feature extraction, and finally fuse audio and visual modality at score-level. For dynamic image sequences representing deformations and facial muscle movements, the popular visual features are facial animation parameters or motion parameters. In [5], 34 motion parameters of head, eyebrows, eyes, and mouth are collected as visual features from each facial image in dynamic image sequences. In [46], 18 facial animation parameters are extracted with a facial feature tracking technique performed on dynamic image sequences in video.

B. Multimodality Fusion

Multimodality fusion is to integrate audio and visual modalities with different statistical properties. Existing fusion strategies [12], [16]–[18] can be summarized into four categories, *i.e.*, feature-level fusion, decision-level fusion, score-level fusion, and model-level fusion, respectively.

Feature-level fusion is the most common and straightforward way, in which all extracted features are directly concatenated into a single high-dimensional feature vector. Then, a single classifier can be trained with this high-dimensional feature vector for emotion recognition. Feature-level fusion is thus also called *Early Fusion* (EF). A substantial number of previous works [3], [5], [16], [47], [48] have testified the performance of feature-level fusion on audio-visual emotion recognition tasks. However, because it merges audio and visual features in a straightforward way, feature-level fusion can not model the complicated relationships, *e.g.*, the difference on time scales and metric levels, between audio and visual modalities.

Decision-level fusion aims to combine several unimodal emotion recognition results through an algebraic combination rule. Specifically, each input modality is modeled independently with an emotion classifier, then these unimodal recognition results are combined with certain algebraic rules such as "Max", "Min", "Sum", *etc*. Thereby, decision-level fusion is also called as *Late Fusion* (LF). Previous works Mansoorizadeh and Charkari [5], Zhalehpour *et al.* [12], Wang *et al.* [16], Schuller *et al.* [47], and Busso *et al.* [48], have adopted decision-level fusion in audio-visual emotion recognition. Nevertheless, decision-level fusion can not capture the mutual correlation among different modalities, because these modalities are assumed to be independent. Therefore, decision-level fusion does not conform to the fact that human beings show audio and visual expressions in a complementary redundant manner, rather than a mutually independent manner.

Score-level fusion, as a variant of decision-level fusion, has been recently employed for audio-visual emotion recognition [16], [18]. In [16], an equally weighted summation is adopted to the obtained class score values. The emotion category corresponding to the maximum value in this fused score vector is taken as the final predicted category. Note that, score-level fusion is implemented by combining the individual classification scores, which indicate the likelihood that a sample belongs to different classes. By contrast, decision-level fusion is performed by combining multiple predicted class labels.

Model-level fusion, as a compromise between feature-level fusion and decision-level fusion, has also been used for audio-visual emotion recognition. This method aims to obtain a joint feature representation of audio and visual modalities. Its implementation mainly depends on the used fusion model. For instance, Zeng *et al.* [4] employ a Multi-stream Fused Hidden Markov Models (MFHMM) to implement model-level fusion. This MFHMM combines bimodal information from audio and visual streams in terms of the maximum entropy principle and the maximum mutual information criterion. Lin *et al.* [8] employ an error weighted semi-coupled Hidden Markov Models (HMM) to fuse audio and visual streams for emotion recognition. In [46], a Tripled Hidden Markov Models (THMM) model is adopted to perform audio-visual emotion recognition. As for neural networks, model-level fusion is performed by first concatenating feature representations of different hidden layers of neural networks corresponding to multiple input modalities. Then, an additional hidden layer is added to learn a joint feature representation from the concatenated feature representations [26]–[28]. Limited by the shallow structure like the maximum entropy principle [4] and single hidden layer [26]–[28], existing model-level fusion methods are still not effective in modeling highly non-linear correlations between audio and visual modalities.

C. Summary

From the above-mentioned works, we can make the following two summarizations.

First, low-level hand-crafted features are widely-used for audio and visual emotion recognition. However, these hand-crafted features can not sufficiently and efficiently discriminate emotions. It is thus desirable to develop automatic feature learning algorithms to obtain high-level affective features. CNNs [23] automatically learn features from raw pixels. With raw Mel-spectrogram as inputs, CNNs may present a cue for high-level audio feature extraction. CNNs have exhibited promising performance on feature learning from static images [44]. However it can not directly capture motion cues in videos. To address this issue, a 3D-CNN [33], which computes feature maps from both spatial and temporal dimensions, may be a possible solution.

Second, Most of existing fusion methods belong to the shallow fusion method, which can not effectively model the complicated non-linear joint distribution and correlations of multiple modalities [19], *e.g.*, the feature concatenation. Therefore, it is necessary to develop deep fusion methods that leverage deep models for feature fusion. To alleviate this problem, it is hence needed to design a deep fusion model in which multiple meaningful fusion operations can be performed to learn the complicated joint audio-visual feature representation. Because each RBM in a DBN can be used to learn the joint audio-visual feature representation, it may be feasible to employ a DBN consisting of multiple layers of RBMs as a deep fusion model.

III. Proposed Method

As described in Fig. 1, our hybrid deep learning model contains two individual input streams, *i.e.*, the audio network processing audio signals with a CNN model, and the visual network processing visual data with a 3D-CNN model. The outputs of fully connected layers of these two networks are fused in a fusion network built with a DBN model.

Due to the limited amount of labeled data, we first employ the existing CNN and 3D-CNN models pre-trained on large-scale image and video classification tasks to initialize our CNN and 3D-CNN, respectively. Then, fine-tuning is conducted for these two CNN models with the labeled emotion data. To this end, we adopt the AlexNet [23] for CNN network initialization, and the C3D-Sports-1M model [33] for 3D-CNN network initialization, respectively. The AlexNet [23] has 5 convolution layers (Conv1-Conv2-Conv3-Conv4-Conv5), 3 max-pooling layers (Pool1-Pool2-Pool5), and 3 fully connected (FC) layers. The first two FC layers (fc6, fc7) consist of 4096 units and the last FC layer (fc8) has 1000 dimensions corresponding to 1000 image categories. The C3D-Sports-1M model [33] contains 8 convolution layers (Conv1a-Conv2a-···-Conv5a-Conv5b), 5 max-pooling layers (Pool1-Pool2-Pool3-Pool4-Pool5), followed by 3 FC layers. In this 3D-CNN, its fc6, fc7 also have 4096 units, and its fc8 corresponds to 487 video categories. To initialize the audio and visual networks in Fig. 1, we copy the initial network parameters from the corresponding pre-trained CNN and 3D-CNN models mentioned above. Note that the fc8 parameters in these two pre-trained models are not used.

In the followings, we describe how to generate the inputs of both CNN and 3D-CNN, and how this hybrid deep learning model is trained.

A. Generation of Network Inputs

Since emotional video samples may have different duration, we split each of them into a certain number of overlapping segments and then learn audio-visual features from each segment. This also enlarges the amount of training data for our deep models. In detail, we first extract the whole log Mel-spectrogram from audio signals. The extracted log Mel-spectrogram is computed with the output of Mel-frequency filter banks, and shows more discriminant power than MFCC for audio emotion recognition [49]. Then, we use a fixed context window to split the spectrogram into overlapping segments which are converted into the suitable input of CNN. The corresponding video segment in this context window is used as the input of 3D-CNN after preprocessing. In this way, for each video segment, we produce its Mel-spectrogram segment and video frames in the framework as illustrated in Fig. 1. In the followings, we present how these audio and visual cues are processed in detail.

1) Audio Input Generation: It is known that the 1-D spectrogram, represented by the squared magnitude of the time-varying spectral characteristics of audio signals, contains tremendous low-level acoustic information related to the speaker's emotion expression, such as energy, pitch, formants, and so on [50]. However, CNNs are commonly used to process 2-D or 3-D images in vision tasks [23]. To leverage the available CNN models and make our deep model initialization easier, it is hence intuitive to transform the 1-D spectrogram into a 2-D array as the input of CNN.

Recently, Abdel-Hamid *et al.* [51] have employed a CNN with a shallow 1-layer structure for speech recognition. Specially, the authors extract the log Mel-spectrogram from raw audio signals and reorganize it into a 2-D array as the input of CNN. Then 1-D convolution can be applied along the frequency axis. Nevertheless, audio emotion recognition is different from speech recognition [51]. First, 1-D convolution operation along the frequency axis can not capture the useful temporal information along the time axis for emotion recognition. Second, a speech segment length of 15 frames (about 165 ms) widely-used for speech recognition does not carry sufficient temporal cues for distinguishing emotion. Some previous studies also show that 250 ms is the suggested minimum segment length required for identifying emotion [52], [53].

As shown in Fig. 1, to convert the 1-D audio signals into the suitable input of CNN, we extract three channels of log Mel-spectrogram segment (*i.e.*, the $static$, $delta$ and $delta - delta$) with size 64 × 64 × 3. Specifically, for a given utterance we adopt 64 Mel-filter banks from 20 to 8000 Hz to obtain the whole log Mel-spectrogram by using a 25ms Hamming window and a 10ms overlapping. Then, a context window of 64 frames is used to divide the whole log Mel-spectrogram into audio segments with size 64 × 64. A shift size of 30 frames is used during segmentation, *i.e.*, two adjacent segments are overlapped with 30 frames. Each divided segment hence has a length of 64 frames and its time duration is 10ms × (64−1)+25ms = 655ms. In this case, the divided segment length is 2.5 times longer than the suggested minimum segment length (250 ms) for identifying emotion [52], [53]. Consequently, each divided segment conveys sufficient temporal cues for identifying emotion. The produced 2-D Mel-spectrogram segment with size 64×64 is taken as the first channel (static) among three channels of Mel-spectrogram.

After extracting the static Mel-spectrogram segment with size 64 × 64, we compute its first-order ($delta$) and second-order ($delta - delta$) frame-to-frame time derivatives. This is used to better capture the temporal information of Mel-spectrogram, *i.e.,* the feature trajectories over time, as usually done in speech recognition tasks [54].

To calculate the delta coefficients of the static 2-D Mel-spectrogram segment, the following regression formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{1}$$

where d_t is a delta coefficients of frame t computed using the static Mel-spectrogram segment coefficients c_{t+n} to c_{t-n}. The value of N represents the regression window with a typical value of 2. Then, in the same way we can calculate the $delta{-}delta$ coefficients from the obtained $delta$ coefficients. As a result, we can obtain three channels of Mel-spectrogram segment with size: 64 × 64 × 3, as illustrated in Fig. 1.

This extracted Mel-spectrogram can be regarded as the RGB image feature representation of audio data. It has two desired properties. First, we can use it to implement the 2-D convolution operation along the frequency and time axis, rather than the 1-D convolution operation. Second, as the RGB image feature representation, it is convenient to resize it into the suitable size as the input of the pre-trained CNN models. Specifically, we initialize the audio network with the AlexNet [23], which has the input size 227 × 227 × 3. Therefore, we resize the original spectrogram with size 64 × 64 × 3 into new size: 227 × 227 × 3 with bilinear interpolation. In the followings, we denote the audio input as a.

2) Visual Input Generation: After splitting the video sample into segments, we use the video segments as the 3D-CNN input. For each frame in the video segment, we run face detection, estimate the eye distance, and finally crop a RGB face image of size 150 × 110 × 3, as done in [55] and [56]. In detail, we employ the robust real-time face detector presented by Viola and Jones [57] to perform automatic face detection on each frame. From the results of automatic face detection, the centers of two eyes can be located in a typical up-right face. Then, we calculate the eye distance of facial images and normalized it to a fixed distance of 55 pixels. For a facial image, it is usually observed that its height is roughly three times longer than the eye distance, whereas its width is roughly twice. Consequently, based on the normalized eye distance, a resized RGB image of 150 × 110 × 3 is finally cropped from each frame. To conduct a fine-tuning task, the cropped facial image for each frame is resized to 227×227×3 as the input of the pre-trained 3D-CNN model. Similar resize operation is also used in a previous work [58].

To make sure each video segment has 16 frames, *i.e.*, the input size in C3D-Sports-1M model [33], we delete the first and last $\frac{L-16}{2}$ overlapping frames if a video segment has $L \geq 16$ frames. On the contrary, for $L < 16$ we repeat the first and last $\frac{16-L}{2}$ overlapping frames. It should be noted that, since we employ 64 audio frames in a context window to divide the extracted log Mel-spectrogram into audio segments, the durance of each segment is 655ms corresponding to about 20 video frames in each video segment, *i.e.*, 0.655 s × 30 frame/s. In this case, our implementation does not need to deal with the case with $L < 16$ frames. By contrast, when using 15 audio frames of Mel-spectrogram segments corresponding to about 5 video frames ($L = 5$) for experiments, we need to repeat the first 5 and last 6 overlapping frames. We denote the visual input as v.

B. Network Training

Given audio-visual data $X = \{(a_i, v_i, y_i)\}_{i=1,2,\cdots,K}$, where i is the index of the divided audio-visual segments, a_i and v_i denote the audio data and visual data, respectively, and y_i represents the class label of a segment. Note that, we use the class label of the global video sample as the class label of a segment y_i. Let $\Upsilon^A(a_i; \theta^A)$ denotes the 4096-D output of fc7 in audio network (denoted as A) with network parameters θ^A. Similarly, $\Upsilon^V(v_i; \theta^V)$ denotes the 4096-D visual feature (fc7) of the visual network (denoted as V) with network parameters θ^V. During network training, we first train the audio and visual networks respectively in the first stage, then jointly train the fusion network in the second stage.

1) Training Audio and Visual Networks: The audio and visual networks are first trained individually with a fine-tuning scheme. For the CNN and 3D-CNN, we replace their final fully connected layers, *i.e.*, fc8 layer, with two new FC layers, which correspond to the emotion categories on target audio-visual emotion recognition dataset. For instance, for 6 emotions, fc8 should produces 6 outputs. Accordingly, we predict the emotional labels with the audio and visual networks, respectively, then calculate the prediction errors and finally update the network parameters to minimize the negative log likelihood L over the training data.

On audio training data, we solve the following minimization problem to update the audio network A with back propagation:

$$\min_{W^A, \theta^A} \sum_{i=1}^{K} L(\text{softmax}(W^A \cdot \Upsilon^A(a_i; \theta^A)), y_i), \tag{2}$$

where W^A is the weight values of the softmax layer, and the softmax log-loss is calculated by

$$L(A, y) = -\sum_{j=1}^{l} y_j \log(y_j^A), \tag{3}$$

where y_j is the j-th value of the ground truth label, y_j^A represents the j-th output value of the softmax layer for the network A, and l represents the total number of class labels.

On visual training data, we solve a minimization problem similar to the audio network, since a 3D-CNN has the same minimization problem as CNNs. In this way, we can minimize the prediction error of V to update the visual network V.

During the first stage of training, we can separately update the parameters in the audio and visual networks, producing more discriminative audio and visual features, *i.e.*, $\Upsilon^A(a_i; \theta^A)$ and $\Upsilon^V(v_i; \theta^V)$. To fuse the audio and visual features, we proceed to describe the training of our fusion network.

2) Training Fusion Network: After training the audio and visual networks, we discard their fc8 layers and merge their fc7 layers into the fusion network illustrated in Fig. 1. In this way, two 4096-D features, *i.e.,* $\Upsilon^A(a_i; \theta^A)$ and $\Upsilon^V(v_i; \theta^V)$ are concatenated to constitute a 8192-D feature as the input of the fusion network $f([\Upsilon_i^A, \Upsilon_i^V]; \theta^F)$ (denoted as F) with network parameters θ^F. Here, $\Upsilon_i^A = \Upsilon^A(a_i; \theta^A)$ and $\Upsilon_i^V = \Upsilon^V(v_i; \theta^V)$.

Our fusion network is built with a deep DBN model, which aims to capture highly non-linear relationships across modalities, and form a joint discriminant feature representation for emotion classification. It contains one visible layer, two hidden layers and one output layer (*i.e.*, softmax layer), as depicted in Fig. 1. This DBN model is constructed by stacking two RBMs, each of which is a bipartite graph and its hidden nodes are able to obtain higher-order correlation of input data of visible nodes.

Following [21], we train the fusion network through two training steps. First, an unsupervised pre-training is implemented in the bottom-up manner by using a greedy layer-wise training algorithm [21]. This unsupervised pre-training aims to minimize the following reconstruction error, *i.e.*,

$$\min_{W^F, \theta^F} \sum_{i=1}^{K} C(z_i, z_i'), \tag{4}$$

where K is the number of training samples, $C(z_i, z_i')$ denotes the cross-entropy loss function between the input data z_i and the reconstructed data z_i'. Here, $C(z_i, z_i')$ is defined as

$$C(z_i, z_i') = \sum_{d=1}^{D} (-z_{i,j} \log z_{i,d}' + (1 - z_{i,d}) \log(1 - z_{i,d}')), \tag{5}$$

Second, after pre-training, each layer of RBMs is initialized. Then, a supervised fine-tuning is performed to optimize the network parameters. In detail, we take the last hidden layer output as the input of a classifier, and compute the classification error. Then, back propagation is used to readjust the network parameters.

Since the input features of DBNs are continuous values, we use a Gaussian-Bernoulli RBM with 4096 hidden nodes for its first layer, a Bernoulli-Bernoulli RBM with 2048 hidden nodes for its second layer, outputting 2048-D features for emotion classification. In this way, we get a 8192-4096-2048-C structure of DBNs, which is used to identify C emotions on target audio-visual emotional datasets. Note that, we fix the parameters in A and V during the second stage training, and update the parameters of the fusion network F to produce more accurate emotional predictions, resulting in better feature fusion results.

C. Emotion Classification

After finishing training the fusion network, a 2048-D joint feature representation can be computed on each audio-visual segment. Since each audio-visual video sample has a different number of segments, average-pooling is applied on all segment features from each video sample to form the fixed-length global video feature representation. Our experiments compared average-pooling and max-pooling, and found average-pooling performs better. Therefore, we employ average-pooling to process features extracted from segments. Based on this global video feature representation, the linear SVM classifier can be easily employed for emotion identification.

IV. Experiments

To testify the effectiveness of our proposed hybrid deep learning networks for audio-visual emotion recognition, we conduct emotion recognition experiments on three public audio-visual emotional datasets, including the acted RML dataset [3], the acted eNTERFACE05 dataset [59], and the spontaneous BAUM-1s dataset [12]. To evaluate the performance of our proposed method, we present the unimodal audio and video emotion recognition results, and then give the multimodal emotion recognition results integrating audio and video cues.

A. Datasets

RML: The RML audio-visual dataset [3] is composed of 720 video samples from 8 subjects, speaking the six different languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). It contains the six emotions: anger, disgust, fear, joy, sadness, and surprise. The audio samples are recorded with a sampling rate of 22,050 Hz with 16-bit resolution and mono channel. At least two participants who do not know the corresponding language are employed in human perception test to evaluate whether the correct emotion is expressed. A video sample is added to this dataset, when all testing subjects are able to perceive the intended emotion categories. The average duration of each video sample is around 5 seconds. The size of original video frame is 720 × 480 × 3. In our experiments, for each frame in a video segment, we crop a facial image with size 150 × 110 × 3, as described in Section III-A. Fig. 2 shows some samples of the cropped facial images on the RML dataset.

eNTERFACE05: The eNTERFACE05 [59] audio-visual acted dataset includes the six emotions, *i.e.*, anger, disgust, fear, joy, sadness, and surprise, from 43 subjects with 14 different nationalities. It contains 1290 video samples. Each audio

Fig. 2. Some samples of the cropped facial images from the RML dataset.

Fig. 3. Some samples of the cropped facial images from the eNTERFACE05 dataset.

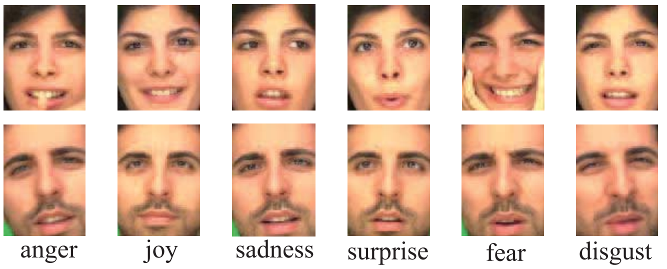


Fig. 4. Some samples of the cropped facial images from the BAUM-1s dataset.

sample is recorded with a sampling rate of 48,000 Hz with 16-bit resolution and mono channel. Each subject is asked to listen to the six successive short stories, each of which is used to induce a particular emotion. Two experts are employed to evaluate whether the reaction expresses the intended emotion in an unambiguous way. The speech utterances are pulled from video files of the subjects speaking in English. The video files are in average 3-4 seconds long. The size of original video frames is 720 × 576 × 3. Fig. 3 gives some samples of the cropped facial images on the eNTERFACE05 dataset.

BAUM-1s: The BAUM-1s [12] audio-visual spontaneous dataset contains 1222 video samples from 31 Turkish subjects. The dataset has the six basic emotions (joy, anger, sadness, disgust, fear, surprise) as well as boredom and contempt. It also contains four mental states, namely unsure, thinking, concentrating and bothered. To obtain spontaneous audio-visual expressions, emotion elicitation by watching films is employed. The size of original video frames is 720 × 576 × 3. Similar to [3], [12], and [59], this work focus on recognizing the six basic emotions, which appear in total 521 video clips. Fig. 4 gives some samples of the cropped facial images on the BAUM-1s dataset.

We divide a video sample into a certain number of audio and video segments as the input of CNN and 3D-CNN, respectively. Because multiple segments can be generated from each video sample, the amount of training data will be enlarged. For example, we generate 11,316 audio-video segments from 720 video samples on the RML dataset, generate 16,186 segments from 1290 video samples on the eNTERFACE05 dataset, and generate 6386 segments from 521 video samples on the BAUM-1s dataset, respectively.

B. Experimental Setup

For training our deep models, we use a mini-batch size of 30, and a stochastic gradient descent with stochastic momentum of 0.9. The learning rate is 0.001 for fine-tuning. The maximum number of epochs is set to 300 for CNNs, 400 for 3D-CNNs, and 100 for DBNs, respectively. For the "FC" fusion method, the dropout parameter is set to 0.3. We implement CNNs with the MatConvNet toolbox,[1] 3D-CNNs with the Caffe toolbox,[2] as well as DBNs with the DeeBNet toolbox.[3] One NVIDIA GTX TITAN X GPU with 12GB memory, is used to train these deep models. For emotion classification, we utilize the LIBSVM package,[4] to perform the SVM algorithm with the linear kernel function and one-versus-one strategy. As suggested in [60], we adopt the subject-independent Leave-One-Subject-Out (LOSO) and Leave-One-Speakers-Group-Out (LOSGO) cross-validation strategies for experiments, which are commonly used in real-world applications. In detail, on the RML and eNTERFACE05 datasets, we employ the LOSO scheme. On the BAUM-1s dataset, we adopt the LOSGO scheme with five speaker groups, as done in [12]. The average accuracy in the test-runs are finally reported to evaluate the performance of all compared methods.

C. Experimental Results and Analysis

In this section, we present experimental results of unimodality and multimodality features on the RML, eNTERFACE05, and BAUM-1s datasets, respectively.

1) Unimodality Performance: To testify the effectiveness of feature learning with deep models, we present the recognition performance of two types of features, *i.e.*, features extracted with the original AlexNet and C3D-Sports-1M models, and the learned features extracted with the fine-tuned AlexNet and C3D-Sports-1M models. For the features extracted with the original AlexNet and C3D-Sports-1M models, we directly take our generated audio and visual data as the inputs of AlexNet and C3D-Sports-1M models, producing 4096-D features from the outputs of their fc7 layers, respectively.

Table I shows the recognition performance of these features on the RML, eNTERFACE05, and BAUM-1s datasets. From the results in Table I, we can see that the learned features with fine-tuned deep models (AlexNet and C3D-Sports-1M)

[1] available at http://www.vlfeat.org/matconvnet/

[2] available at http://caffe.berkeleyvision.org/

[3] available at http://ceit.aut.ac.ir/ keyvanrad/

[4] available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/

TABLE I

Subject-Independent Unimodality Recognition Accuracy (%) on Three Datasets. $Alex_{Audio}$ and C3D$_{Visual}$ Are Audio and Visual Features Extracted by the Original Pre-Trained AlexNet and C3D-Sports-1M Models, Respectively. $Anet$ and $Vnet$ Are the Learned Features of the Fine-Tuned Audio Network (CNN) and Visual Network (3D-CNN), Respectively

Unimodality	Features	RML	eNTERFACE05	BAUM-1s
Audio	$Alex_{Audio}$	59.46	51.33	36.10
	$Anet$	66.17	78.08	42.26
Visual	$C3D_{Visual}$	53.03	48.97	41.69
	$Vnet$	68.09	54.35	50.11

TABLE II

Subject-Independent Audio Emotion Recognition Performance (%) Comparisons With Previous Works Using Hand-Crafted Features on Three Datasets. $Anet$ Is the Learned Features of the Fine-Tuned Audio Network (CNN)

Datasets	Refs.	Audio features	Accuracy
RML	Gao *et al.*,[36]	Prosody	51.04
	Elmadany *et al.*, [38]	Prosody	56.25
	Elmadany *et al.*, [37]	PNCC	58.33
	Zhang *et al.*, [62]	LLD	61.86
	Ours	$Anet$	**66.17**
eNTERFACE05	Zhalehpour *et al.*,[12]	MFCC,RASTA-PLP	72.95
	Schuller *et al.*, [35]	Prosody, MFCC	72.40
	Mansoorizadeh *et al.*, [5]	Prosody	43.00
	Bejani *et al.*, [14]	Prosody, MFCC	54.99
	Ours	$Anet$	**78.08**
BAUM-1s	Zhalehpour *et al.*,[12]	MFCC,RASTA-PLP	29.41
	Ours	$Anet$	**42.26**

TABLE III

Subject-Independent Visual Emotion Recognition Performance (%) Comparisons With Previous Works Using Hand-Crafted Features on Three Datasets.$Vnet$ Is the Learned Features of the Fine-Tuned Visual Network (3D-CNN)

Datasets	Refs.	Visual features	Accuracy
RML	Elmadany *et al.*, [37]	Gabor wavelet	64.58
	Zhang *et al.*, [62]	LBP	56.90
	Ours	$Vnet$	**68.09**
eNTERFACE05	Zhalehpour *et al.*,[12]	LPQ	42.16
	Mansoorizadeh *et al.*, [5]	Facial points	37.00
	Bejani *et al.*, [14]	QIM	39.27
	Ours	$Vnet$	**54.35**
BAUM-1s	Zhalehpour *et al.*,[12]	LPQ	45.04
	Ours	$Vnet$	**50.11**

significantly outperform the features extracted with the original pre-trained deep models. In detail, our fine-tuning strategy improves the accuracies on the RML dataset from 59.46% to 66.17% for audio features, and 53.03% to 68.09% for visual features, respectively. Similarly, on the eNTERFACE05 dataset, our method also makes an improvement from 51.33% to 78.08% for audio features, and 48.97% to 54.35% for visual features, respectively. On the BAUM-1s dataset, improvements of 6.16% for audio features, and 8.42% for visual features are achieved, respectively. The experimental results demonstrate the effectiveness of our feature learning strategy, *i.e.*, using deep model to learn emotional features. Our learned features have potential to leverage the powerful learning ability of deep models to extract more discriminative cues than the manually designed features. The experimental results also show the validity of our fine-tuning strategy. Fine-tuning allows deep models pre-trained on other domains to learn meaningful feature representations for emotion recognition.

To present the advantages of the learned features, we directly compare our performance with the reported results of previous works using hand-crafted features on these datasets. Because these compared works use the same experimental settings with ours, *i.e.*, subject-independent test-runs, we take their reported results for comparison. It is not suitable to compare our work with previous works adopting subject-dependent test-runs. Table II and Table III separately give performance comparisons of audio and visual emotion recognition between our learned features and the corresponding hand-crafted features.

From Table II, we can see that our learned audio features with CNNs outperform the hand-crafted audio features widely-used for audio emotion recognition [5], [12], [35]–[38], [61], such as prosody features, MFCC, Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP), Power Normalized Cepstral Coefficients (PNCC), and other acoustic Low-level Descriptors (LLD). This shows that our audio features learned by the fine-tuned AlexNet model is more discriminative than the hand-crafted audio features for audio emotion classification. In addition, the promising performance of our learned audio features clearly indicates that it is reasonable to employ three channels of Mel-spectrogram with size $64 \times 64 \times 3$ as the input of AlexNet. It is also interesting to know that AlexNet trained on image domain can be applied to audio feature extraction. This might be because of the powerful feature learning ability of AlexNet, *e.g.*, higher-level convolutions progressively infer semantics from larger receptive fields. The extracted Mel-spectrogram is similar to the RGB image representation. This representation makes it possible to first extract meaningful low-level time-frequency features by low-level 2-D convolutions, then infer more discriminative features by higher levels of convolutions. It is also possible that, three channels of Mel-spectrogram present emotions as certain shapes and structures, which thus can be effectively perceived by AlexNet trained on the image domain. This thus presents a new method of transforming 1-D audio signals into the suitable input of CNN that conventionally processes 2-D or 3-D images.

From Table III, it can be observed that our learned visual features with 3D-CNNs yield better performance than the compared hand-crafted features [5], [12], [14], [37], [61], such as Gabor wavelet, LBP, LPQ, facial points and Quantized Image Matrix (QIM). This demonstrates the advantages of our learned visual features produced by the fine-tuned C3D-Sports-1M model, which presents more discriminative power than the hand-crafted visual features for visual emotion recognition. The above experiments clearly show that deep model is powerful in feature learning and produces more

TABLE IV

Audio Emotion Recognition Performance (%) Comparisons Between Mel-Spectrogram Segments With Sizes 64 × 64 × 3 and 64 × 15 × 3

Mel-spectrogram	RML	eNTERFACE05	BAUM-1s
$64 \times 15 \times 3$	50.06	52.33	33.64
$64 \times 64 \times 3$	66.17	78.08	42.26

discriminative features than manually designed feature extraction models. However, deep models require a large amount of training data. This thus motivates us to transfer pre-trained models on other domain for emotional feature learning.

Table II and III show that our learned features are more discriminative to the hand-crafted features on emotion recognition tasks. However, our feature learning needs a large training set, and is easier to suffer from overfitting than the hand-crafted features. Moreover, extracting features with deep models requires more expensive computation due to the massive network parameters.

To explain why we extract Mel-spectrogram segments with a length of 64 frames rather than 15 frames widely used in speech recognition [51], we compare the performance of two types of extracted Mel-spectrograms with different length, *i.e.*, $64 \times 64 \times 3$ and $64 \times 15 \times 3$. The experimental results are summarized in Table IV. From Table IV, we can see that the extracted Mel-spectrogram with size $64 \times 64 \times 3$ clearly outperforms the other one. This indicate that the segment length of 15 frames is not suitable for audio emotion recognition. This might be because 15 frames is too short to convey sufficient information for distinguishing emotions [52], [53].

2) Multimodality Performance: To verify the effectiveness of our fusion method, we compare our method with four multimodality fusion schemes, *i.e.,* feature-level fusion, decision-level fusion, score-level fusion, as well as our recently-presented method in [61]. Note that, this work employs a deep DBN model to build the fusion network, whereas [61] uses two FC layers.

Our goal is to recognize the emotion of the global video samples. Therefore, feature fusion methods are required to aggregate features extracted on audio-visual segments into a global video feature representation. Then, a linear SVM could be used to perform emotion classification on the generated global video features. To this end, average-pooling is used, as done in Section III-C. Fig. 5 gives the structure of feature-level fusion and decision-level fusion with our two CNN models. Note that, the structure of score-level fusion is completely similar to decision-level fusion.

For decision-level fusion, six typical ensemble rules [62], [63], including "Majority vote", "Max", "Sum", "Min", "Average" and "Product" are testified. For more details about the six ensemble rules, refer to [62] and [63]. On decision-level fusion tasks, we first investigate the performance of each ensemble rule, and then find the best one, which is used to generate the reported performance. Table V presents the performance comparison of six ensemble rules on our learned features. As shown in Table V, the "Product" rule yields best performance. Therefore, in the following experiments, we only report the performance of the "Product" rule for decision-level fusion.

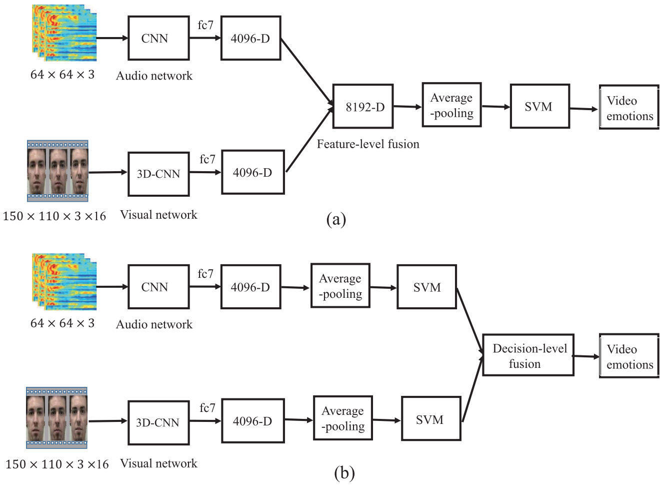


Fig. 5. The structure of feature-level fusion (a) and decision-level fusion (b) with two CNN models.

TABLE V

Multimodality Emotion Recognition Performance (%) Comparison of Six Ensemble Rules at Decision-Level Fusion With the Learned Features of *Anet* and *Vnet*

Decision-level	RML	eNTERFACE05	BAUM-1s
Majority vote	63.58	69.14	45.35
Max	72.05	80.31	48.89
Sum	72.79	80.85	50.01
Min	72.91	78.76	50.08
Average	72.79	80.85	50.01
Product	**74.60**	**81.62**	**51.73**

TABLE VI

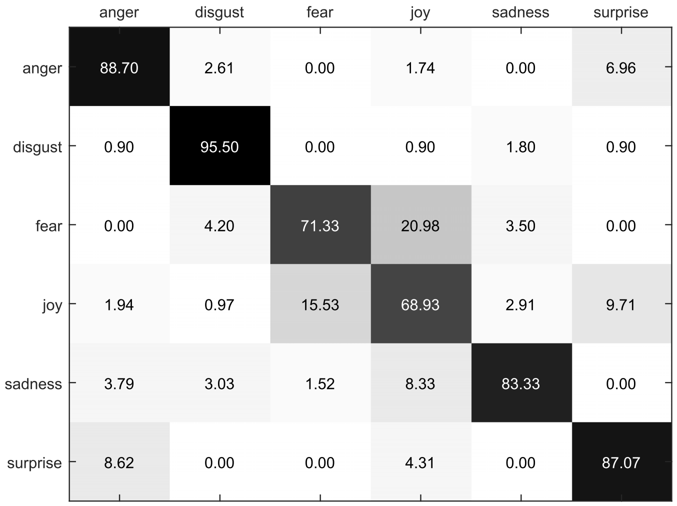
Subject-Independent Multimodality Emotion Recognition Accuracy (%) With the Learned Features of *Anet* and *Vnet*. FC Denotes the Fusion Method Built With Two FC Layers in [61], DBN Denote the Fusion Method Built With a DBN Model

Fusion method	RML	eNTERFACE05	BAUM-1s
Feature-level	74.04	81.02	51.72
Product	74.60	81.62	51.73
Score-level	73.92	80.85	51.58
FC	78.84	83.55	52.35
DBN	**80.36**	**85.97**	**54.57**

We implement score-level fusion referring to the schemes in [16]. Specifically, a equally weighted summation is adopted in terms of the obtained class score values, as described below:

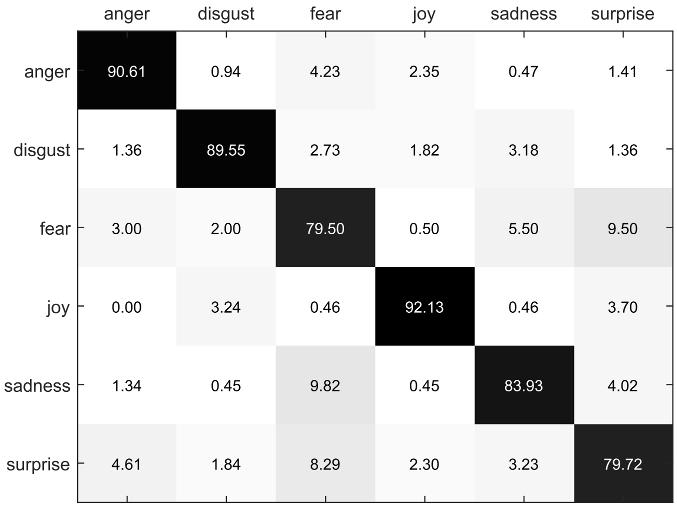
$$Score^{fusion} = 0.5Score^{audio} + 0.5Score^{visual}. \qquad (6)$$

Table VI shows the recognition performance of audio-visual modalities by using different fusion strategies. From Table VI, we can observe that the FC fusion method [61] outperforms feature-level fusion, decision-level fusion (Product), and score-level fusion. This demonstrates the advantages of the fusion network built with two FC layers. This also implies that the FC fusion network is able to learn a joint audio-visual feature



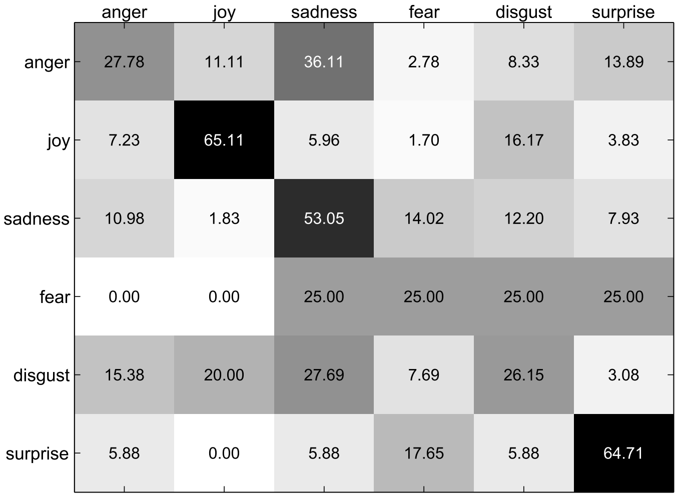
	anger	disgust	fear	joy	sadness	surprise
anger	88.70	2.61	0.00	1.74	0.00	6.96
disgust	0.90	95.50	0.00	0.90	1.80	0.90
fear	0.00	4.20	71.33	20.98	3.50	0.00
joy	1.94	0.97	15.53	68.93	2.91	9.71
sadness	3.79	3.03	1.52	8.33	83.33	0.00
surprise	8.62	0.00	0.00	4.31	0.00	87.07

Fig. 6. Confusion matrix of multimodality emotion recognition results when DBN performs best on the RML dataset.



	anger	disgust	fear	joy	sadness	surprise
anger	90.61	0.94	4.23	2.35	0.47	1.41
disgust	1.36	89.55	2.73	1.82	3.18	1.36
fear	3.00	2.00	79.50	0.50	5.50	9.50
joy	0.00	3.24	0.46	92.13	0.46	3.70
sadness	1.34	0.45	9.82	0.45	83.93	4.02
surprise	4.61	1.84	8.29	2.30	3.23	79.72

Fig. 7. Confusion matrix of multimodality emotion recognition results when DBN performs best on the eNTERFACE05 dataset.



	anger	joy	sadness	fear	disgust	surprise
anger	27.78	11.11	36.11	2.78	8.33	13.89
joy	7.23	65.11	5.96	1.70	16.17	3.83
sadness	10.98	1.83	53.05	14.02	12.20	7.93
fear	0.00	0.00	25.00	25.00	25.00	25.00
disgust	15.38	20.00	27.69	7.69	26.15	3.08
surprise	5.88	0.00	5.88	17.65	5.88	64.71

Fig. 8. Confusion matrix of multimodality emotion recognition results when DBN performs best on the BAUM-1s dataset.

representation from the outputs of two fine-tuned deep models for emotion identification through the back-propagation learning algorithms.

It also can be observed from Table VI that, our DBN fusion method outperforms the other fusion methods. Compared with feature-level fusion, decision-level fusion, and score-level fusion methods, the DBN fusion can be regarded as a deep fusion model. The comparison clearly shows the effectiveness of the deep fusion method, which shows better feature learning ability in capturing the highly non-linear relationships across modalities. This is mainly because of the use of the multi-layer structure of DBNs, in which multiple RBMs are stacked to form multiple hidden layers. A RBM is a generative model representing a probability distribution associated with input data. Each RBM in a DBN is able to learn the joint probability distribution of input audio-video data. By using multiple RBMs and the layer-wise training algorithm, DBNs can effectively learn the non-linear dependencies across modalities, and results in better fusion of audio-visual features. This finding is consistent with the one in a previous work [19]. It is desirable to intuitively visualize the learned weights of our DBN model. However, the input of our DBNs are audio-visual features, rather than the semantic images. This make the learned weights of DBNs hard to interpret intuitively.

DBNs also outperform the FC fusion method in [61], *e.g.*, 78.84% vs. 80.36% on the RML dataset, 83.55% vs. 85.97% on the eNTERFACE05 dataset, and 52.35% vs. 54.57% on the BAUM-1s dataset, respectively. This advantage might be due to the unsupervised pre-training in DBNs, which presents local optimal weights for network initialization, whereas the initial weights in the FC fusion method are randomly produced.

Fig. 6, Fig. 7 and Fig. 8 present the classification confusion matrix on three datasets, respectively. Note that, these confusion matrixes are obtained in terms of the average LOSO or LOSGO recognition results. It is interesting to find that on the RML dataset, "joy" and "fear" are more difficult to be identified than the other emotions. This might be because the audio-visual cues of 'joy" and "fear" are not distinct enough. On the eNTERFACE05 dataset, "sadness", "surprise" and "fear" are recognized with relatively lower accuracy, *i.e.*, about 80%, whereas others emotions are identified well with accuracy of about 90%. On the BAUM-1s dataset, the average classification accuracy is much lower than the ones of the other two datasets. This shows that the spontaneous emotions are more difficult to be recognized than the acted emotions.

In addition to the classification confusion matrix, we compute precision, recall and F-score to further measure the multimodality emotion recognition performance on three datasets. The experimental results are presented in Table VII, Table VIII and Table IX, respectively. The results in these three tables indicate that the three datasets show different difficulties in recognizing specific emotions. For example, it is easier to identify "disgust" on the RML dataset than the other two datasets. It is easier to identify "joy" on the eNTERFACE05 and BAUM-1s than the RML dataset.

TABLE VII

Multimodality Performance (%) Measure for Each Emotion When DBN Gives an Average Accuracy of 80.36% on the RML Dataset

Emotion	Precision	Recall	F-score
Anger	85.33	88.70	86.98
Disgust	89.71	95.50	92.51
Fear	80.71	71.33	75.73
Joy	65.53	68.93	67.19
Sadness	91.03	83.33	87.01
Surprise	83.21	87.07	85.10

TABLE VIII

Multimodality Performance (%) Measure for Each Emotion When DBN Gives an Average Accuracy of 85.97% on the eNTERFACE05 Dataset

Emotion	Precision	Recall	F-score
Anger	89.78	90.61	90.20
Disgust	91.36	89.55	90.45
Fear	75.69	79.50	77.55
Joy	92.55	92.13	92.34
Sadness	86.73	83.93	85.31
Surprise	79.95	79.72	79.84

TABLE IX

Multimodality Performance (%) Measure for Each Emotion When DBN Gives an Average Accuracy of 54.57% on the BAUM-1s Dataset

Emotion	Precision	Recall	F-score
Anger	27.06	27.78	27.41
Joy	82.03	65.11	72.60
Sadness	52.49	53.05	52.77
Fear	20.14	25.00	22.31
Disgust	25.65	26.15	25.90
Surprise	41.93	64.71	50.89

TABLE X

Subject-Independent Multimodality Emotion Recognition Accuracy (%) of Different Deep Structures in the Fusion Network

Fusion method	RML	eNTERFACE05	BAUM-1s
FC-1	**78.84**	**83.55**	**52.35**
FC-2	77.92	82.72	51.50
FC-3	75.67	81.25	50.43
DBN-1	78.50	84.16	51.61
DBN-2	**80.36**	**85.97**	**54.57**
DBN-3	80.10	85.02	53.38

TABLE XI

Multimodality Emotion Recognition Performance (%) Comparisons With State-of-the-Art Works on Three Datasets

Datasets	Refs.	Accuracy
RML	Sarvestani *et al.*, [13]	72.03
	Elmadany *et al.*, [37]	75.00
	Zhang *et al.*, [62]	74.32
	Ours	**80.36**
eNTERFACE05	Sarvestani *et al.*, [13]	70.11
	Mansoorizadeh *et al.*, [5]	71.00
	Bejani *et al.*, [14]	77.78
	Zhalehpour *et al.*, [12]	77.02
	Ours	**85.97**
BAUM-1s	Zhalehpour *et al.*, [12]	51.29
	Ours	**54.57**

D. Effect of Deep Structures in the Fusion Network

The structure of DBNs may heavily affect its performance of fusing audio-visual modalities. To evaluate the effectiveness of different deep structures, we present the performance of three DBN fusion networks: DBN-1 (8192-4096-6), DBN-2 (8192-4096-2048-6), and DBN-3 (8192-4096-2048-1024-6). Similarly, we also show the performance of three FC fusion networks corresponding to DBNs: FC-1 (8192-4096-6), FC-2 (8192-4096-2048-6), and FC-3 (8192-4096-2048-1024-6). For these fusion networks, as done in [61], a dropout layer is added before the final softmax layer corresponding to emotion categories. The dropout parameter is set to 0.3 to reduce over-fitting.

Table X presents the performance comparisons of different structures in the fusion network. From Table X, we can see that FC-1 performs best among the three FC fusion networks. This indicates that FC-1 is more effective than FC-2 and FC-3 to fuse audio and visual cues. This might be because with more layers in the FC network, the massively increasing network parameters make the FC network prone to over-fitting. For the DBN fusion network, DBN-2 slightly outperforms DBN-3, and yields substantially better performance than DBN-1. Due to using multiple RBMs and the effective layer-wise training algorithm [21], the deeper DBN models, *i.e.*, DBN-2 and DBN-3, exhibit better feature fusion ability than the 1-layer DBN-1. DBN-3 degrades the performance of DBN-2 may be because DBN-3 is deeper than DBN-2, thus involves more network parameters, which are more difficult to optimize on a small-scale training dataset.

E. Comparisons With the State-of-the-Art Results

We compare our method with some previous works on three datasets in Table XI. Note that these works also conduct subject-independent experiments, which are consistent with our experimental setting. The results in Table XI indicate that our method is very competitive to the state-of-the-art results. Specially, on the acted RML and eNTERFACE05 datasets, our method outperforms previous works [5], [12]–[14], [37] by more than about 5%. On the spontaneous BAUM-1s dataset, we improve the performance of [12] from 51.29% to 54.57%. These compared works use hand-crafted features and shallow fusion methods to integrate audio-visual modalities. This thus shows the advantages of our learned features and fusion strategy.

In addition, our method also improves our previous work [61] from 74.32% to 80.36% on the RML dataset. This is achieved by two improvements. First, compared with the CNN models in [61], 3D-CNN models in this work can extract spatial-temporal cues from video. Second, as shown in our experiments, the DBN fusion method shows better multimodal feature fusion ability than the FC fusion method in [61].

V. Conclusions and Future Work

This paper presents a new method for audio-visual emotion recognition with a hybrid deep learning framework integrating CNN, 3D-CNN and DBN. The outputs of audio and visual networks are connected with a deep DBN model to fuse audio and visual cues. To learn a joint discriminant feature representation, this network is trained in two stages: 1) the audio and visual networks are initialized with the pre-trained AlexNet and C3D-Sports-1M for fine-tuning, and 2) the DBN fusion network is trained on the target emotion classification datasets. Experimental results on the RML, eNTERFACE05, and BAUM-1s datasets show that our hybrid deep learning model jointly learns a discriminative audio-visual feature representation, which performs better than previous hand-crafted features and fusion methods on emotion recognition tasks. Its success guarantees further research in this direction.

Our work transforms 1-D audio signals into three channels of Mel-spectrogram with size $64 \times 64 \times 3$ as the suitable input of CNN. This extracted Mel-spectrogram is regarded as a RGB image feature representation. Consequently, we can conveniently resize it into the suitable size as the input of the existing CNN models pre-trained on image datasets. In this case, it is possible to fine-tune the pre-trained CNN models on target emotion datasets for audio feature extraction. Experimental results demonstrate the validity of our cross-media fine-tuning scheme.

This work employs 3D-CNNs to extract emotional features from video segments. Another commonly used solution for video feature extraction is combining CNN and LSTM [64]. A LSTM could also be a better solution for audio-visual feature fusion on video segments than average pooling. Therefore, we will investigate the performance of CNN+LSTM for facial expression recognition in our future work.

Besides, this work employs a two-stage learning strategy to train the audio-visual networks and fusion network, respectively. An end-to-end learning strategy would be more concise and has potential to further boost performance. Additionally, an end-to-end recognition system could be constructed by referring to recent LSTM based visual recognition works [64]. Therefore, end-to-end learning and recognition strategies would also be investigated in our future work.

It should also be noted that, deep models commonly contain a large number of network parameters, resulting in expensive computational cost. It is thus meaningful to investigate how to reduce the network parameters of deep models, *e.g.*, deep compression [65], to achieve real-time emotion recognition with a deep model.

This work employs face detector developed by Viola and Jones [57] for face detection in videos. It may fail in some challenging datasets like AFEW [66], where face images suffer from substantial viewpoint and illumination changes. More robust face detectors and models will be studied in our future work. Moreover, this work aims to employ deep models to identify discrete emotions, such as anger, disgust, fear, joy, sadness, and surprise. It is interesting to investigate the performance of our proposed method on dimensional emotion recognition tasks, such as the recognition of depression degree on the AVDLC dataset [67]. Text is another important modality characterizing human emotion. Therefore, it is also an important direction to employ deep models to learn audio, visual, and textural features for multimodal emotion recognition [68].

References

[1] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio–visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[2] R. Gupta *et al.*, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 33–40.

[3] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[4] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–visual affective expression recognition through multistream fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570–577, Jun. 2008.

[5] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools Appl.*, vol. 49, no. 2, pp. 277–297, 2010.

[6] M. Glodek *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction* (Lecture Notes in Computer Science), vol 6975. Berlin, Germany: Springer, 2011, pp. 359–368

[7] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr./Jun. 2012.

[8] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.

[9] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Oct. 2011.

[10] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 184–198, Apr./Jun. 2012.

[11] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2331–2352, 2017.

[12] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2016.2553038.

[13] R. R. Sarvestani and R. Boostani, "FF-SKPCCA: Kernel probabilistic canonical correlation analysis," *Appl. Intell.*, vol. 46, no. 2, pp. 438–454, 2017.

[14] M. Bejani, D. Gharavian, and N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Comput. Appl.*, vol. 24, no. 2, pp. 399–412, 2014.

[15] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006.

[16] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[18] Z. Xie and L. Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," *Int. J. Semantic Comput.*, vol. 7, no. 1, pp. 25–42, 2013.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] G. E. Hinton and R. R. Salakhutdinov, "A better way to pretrain deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 2447–2455.

[26] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 3687–3691.

[27] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[28] L. Pang and C.-W. Ngo, "Mutlimodal learning with deep Boltzmann machine for emotion prediction in user generated videos," in *Proc. 5th ACM Int. Conf. Multimedia Retr. (ICMR)*, Shanghai, China, 2015, pp. 619–622.

[29] C. Barat and C. Ducottet, "String representations and distances in deep convolutional neural networks for image classification," *Pattern Recognit.*, vol. 54, pp. 104–115, Jun. 2016.

[30] Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 54–66, 2015.

[31] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[32] W. Ouyang *et al.*, "DeepID-Net: Deformable deep convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2403–2412.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[34] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[35] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Merano, Italy, Nov./Dec. 2009, pp. 552–557.

[36] L. Gao, L. Qi, and L. Guan, "Information fusion based on kernel entropy component analysis in discriminative canonical correlation space with application to audio emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2817–2821.

[37] N. E. D. Elmadany, Y. He, and L. Guan, "Multiview emotion recognition via multi-set locality preserving canonical correlation analysis," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Montreal, QC, Canada, May 2016, pp. 590–593.

[38] N. E. D. Elmadany, Y. He, and L. Guan, "Multiview learning via deep discriminative canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2409–2413.

[39] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of Face Recognition*. London, U.K.: Springer, 2011, pp. 487–519.

[40] X. Zhao and S. Zhang, "A review on facial expression recognition: Feature extraction and classification," *IETE Tech. Rev.*, vol. 33, no. 5, pp. 505–517, 2016.

[41] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, 1998, pp. 454–459.

[42] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[43] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Santa Barbara, CA, USA, Mar. 2011, pp. 878–883.

[44] W. Ding *et al.*, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Tokyo, Japan, 2016, pp. 506–513.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[46] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1913–1920, 2008.

[47] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI)*, Aichi, Japan, 2007, pp. 30–37.

[48] C. Busso *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces (ICMI)*, PA, USA, 2004, pp. 205–211.

[49] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2225–2228.

[50] L. He, M. Lech, N. Maddage, and N. Allen, "Stress and emotion recognition using log-Gabor filter analysis of speech spectrograms," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops (ACII)*, Amsterdam, The Netherlands, 2009, pp. 1–6.

[51] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[52] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 3682–3686.

[53] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[54] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[55] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[56] X. Zhao and S. Zhang, "Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, p. 20, 2012.

[57] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[58] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 868–884.

[59] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, Atlanta, GA, USA, 2006, p. 8.

[60] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, Chiba, Japan, 2010, pp. 2794–2797.

[61] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proc. 6th ACM Int. Conf. Multimedia Retr. (ICMR)*, New York, NY, USA, 2016, pp. 281–284.

[62] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[63] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.

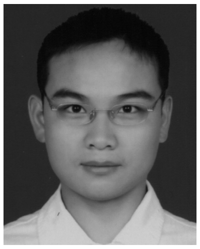
[64] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 701–716.

[65] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 2285–2294.

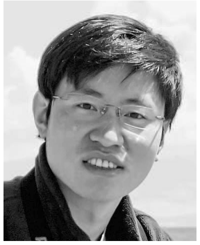
[66] A. Dhall, O. V. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Seattle, WA, USA, 2015, pp. 423–426.

[67] M. Valstar *et al.*, "AVEC 2013—The continuous audio/visual emotion and depression recognition challenge," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, Barcelona, Spain, 2013, pp. 3–10.

[68] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, to be published, doi: 10.1016/j.neucom.2016.09.117.



Shiqing Zhang received the Ph.D. degree from the School of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. He holds a post-doctoral position with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, and is also an Associate Professor with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include audio and image processing, affective computing, and pattern recognition.



Shiliang Zhang received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012. He was a Post-Doctoral Scientist with NEC Laboratories America and a Post-Doctoral Research Fellow with The University of Texas at San Antonio. He is currently a tenure-track Assistant Professor with the School of Electronic Engineering and Computer Science, Peking University. He has authored or co-authored over 30 papers in journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, *ACM Multimedia,* and ICCV. His research interests include large-scale image retrieval and computer vision for autonomous driving.

He was a recipient of the National 1000 Youth Talents Plan of China, the Outstanding Doctoral Dissertation Awards from the Chinese Academy of Sciences and Chinese Computer Federation, the President Scholarship from the Chinese Academy of Sciences, the NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He was a recipient of the Top 10% Paper Award at the IEEE MMSP 2011. His research is supported by the National 1000 Youth Talents Plan and Natural Science Foundation of China (NSFC).

Tiejun Huang (M'01–SM'12) received the bachelor's and master's degrees in computer science from Wuhan University of Technology, Wuhan, China, in 1995 and 1992, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong (Central China) University of Science and Technology, Wuhan, in 1998. He is currently a Professor with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, where he is also the Director of the Institute for Digital Media Technology. He has authored or co-authored over 100 peer-reviewed papers and three books. His research areas include video coding, image understanding, digital right management, and digital library. He is a member of the Board of Directors for Digital Media Project, the Advisory Board of the IEEE Computing Society, and the Board of the Chinese Institute of Electronics.

Wen Gao (M'92–SM'05–F'09) received the Ph.D. degree in electronics engineering from The University of Tokyo, Tokyo, Japan, in 1991.

He is currently a Professor with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China. Before joining Peking University, he was a Professor of Computer Science with Harbin Institute of Technology, Harbin, China, from 1991 to 1995 and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He has authored five books and over 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Dr. Gao serves the Editorial Board for several journals, such as IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Transactions on Autonomous Mental Development, *EURASIP Journal of Image Communications*, and *Journal of Visual Communication and Image Representation*. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.

Qi Tian (S'95–M'96–SM'03–F'16) received the B.E. degree in electronic engineering from Tsinghua University in 1992, the M.S. degree in ECE from Drexel University in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign in 2002.

He was a tenure-track Assistant Professor from 2002 to 2008 and a tenured Associate Professor from 2008 to 2012. From 2008 to 2009, he took a one-year faculty leave with the Media Computing Group, Microsoft Research Asia, as a Lead Researcher. He is currently a full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA). He has authored over 380 refereed journal and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian has co-authored a best paper in the ACM ICMR 2015, a best paper in the PCM 2013, a best paper in the MMM 2013, a best paper in the ACM ICIMCS 2012, the Top 10% Paper Award in the MMSP 2011, and a best student paper in the ICASSP 2006, and co-authored a Best Student Paper Candidate in the ICME 2015 and a Best Paper Candidate in the PCM 2007. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He was a recipient of the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is the Associate Editor of the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, the *ACM Transactions on Multimedia Computing, Communications, and Applications,* and the *Multimedia System Journal,* and is on the Editorial Board for the *Journal of Multimedia* and the *Journal of Machine Vision and Applications.* He is the Guest Editor of the IEEE Transactions on Multimedia and the *Journal of Computer Vision and Image Understanding.*