Information Fusion 46 (2019) 184–192

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Audio-visual emotion fusion (AVEF): A deep efficient weighted approach

Yaxiong Ma[a], Yixue Hao[b], Min Chen[b], Jincai Chen[*,a,b,c], Ping Lu[a,b,c], Andrej Košir[d]

[a] Wuhan National Laboratory for Optoelectronics (WNLO), Huazhong University of Science and Technology, China
[b] Embedded and Pervasive Computing (EPIC) Lab, Huazhong University of Science and Technology, China
[c] Key Laboratory of Information Storage System (School of Computer Science and Technology, Huazhong University of Science and Technology), Ministry of Education of China, China
[d] Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana 1000, Slovenia

ARTICLE INFO

Keywords:
Multi-modality emotion recognition
Deep learning
Transfer learning

ABSTRACT

The multi-modal emotion recognition lacks the explicit mapping relation between emotion state and audio and image features, so extracting the effective emotion information from the audio/visual data is always a challenging issue. In addition, the modeling of noise and data redundancy is not solved well, so that the emotion recognition model is often confronted with the problem of low efficiency. The deep neural network (DNN) performs excellently in the aspects of feature extraction and highly non-linear feature fusion, and the cross-modal noise modeling has great potential in solving the data pollution and data redundancy. Inspired by these, our paper proposes a deep weighted fusion method for audio-visual emotion recognition. Firstly, we conduct the cross-modal noise modeling for the audio and video data, which eliminates most of the data pollution in the audio channel and the data redundancy in visual channel. The noise modeling is implemented by the voice activity detection(VAD), and the data redundancy in the visual data is solved through aligning the speech area both in audio and visual data. Then, we extract the audio emotion features and visual expression features via two feature extractors. The audio emotion feature extractor, audio-net, is a 2D CNN, which accepting the image-based Mel-spectrograms as input data. On the other hand, the facial expression feature extractor, visual-net, is a 3D CNN to which facial expression image sequence is feeded. To train the two convolutional neural networks on the small data set efficiently, we adopt the strategy of transfer learning. Next, we employ the deep belief network (DBN) for highly non-linear fusion of multi-modal emotion features. We train the feature extractors and the fusion network synchronously. And finally the emotion classification is obtained by the support vector machine using the output of the fusion network. With consideration of cross-modal feature fusion, denoising and redundancy removing, our fusion method show excellent performance on the selected data set.

1. Introduction

Emotion state is controlled by human brain [1], and can be expressed through the changes in behavior and physiological features. The interaction of human beings in the daily life cannot be separated from the emotion communication. In addition, with the rapid progress of science and technology, explosion of the internet, and changes in human life style, more and more people spend a lot of time interacting with computers directly every day. Obviously, human-computer interaction is a part of our life that cannot be ignored. To gain better interactive experience in the human-computer interaction (HMI), we hope that modern human-computer interaction system can present in a more natural and friendly way. For this purpose, the computer must possess the capacity of understanding the human emotion state like humans. As we know, the physiological indices is what really matters in the generation process of emotion, which can also be used for recognition of the ‘real’ emotion. However, because physiological indices are inconvenient to access, most of researches relevant to emotion recognition concentrate on human behaviors, such as facial expression, voice, text, gesture and so on.

Among the behavior patterns of emotion expression including voice, text, and gesture, the voice and facial expression happen frequently in human emotional interaction. Also, they are what being discussed most in research community of affective computing. Since they are characterized by the homologous subject and time synchronization in most cases. That is to say, while talking with someone face to face, we can listen to his/her voice and meanwhile look at his/her facial expression in most cases. In recent years, there have been many unimodal emotion recognition researches only in view of voice [2–9] or facial expression [10–14]. Also, there are some multi-modal emotion recognition

[*] Corresponding author at: Wuhan National Laboratory for Optoelectronics (WNLO), Huazhong University of Science and Technology, China.
E-mail addresses: minchen2012@hust.edu.cn (M. Chen), jcchen@hust.edu.cn (J. Chen).

https://doi.org/10.1016/j.inffus.2018.06.003
Received 28 January 2018; Received in revised form 12 April 2018; Accepted 12 June 2018

Available online 15 June 2018

researches. These researches have proposed some methods in the aspects from feature extraction to classification algorithm. However, the recognition of human emotion state by computer is still faced with great challenges. In the unimodal emotion recognition, such as pure emotion recognition from voice or facial expression), the extraction of emotion features from the raw data is always an open-ended problem. At present, the explicit and deterministic mapping between emotion state and concrete feature does not exist. Audio-visual emotion recognition can gain more precise recognition results because audio-visual emotion recognition is more natural and there are more emotion information than unimodal emotion recognition. As a matter of fact, the similar problems also exist in the multi-modal emotion recognition, because the multi-modal emotion recognition is based on the unimodal emotion recognition. Fusing the results of audio emotion recognition and facial expression recognition to varying degrees can be helpful for audio-visual emotion recognition.

In the view of emotion features in audio data, it is reported that, the prosodic features, acoustic features and voice quality features imply comparatively abundant emotional significance. These features include pitch period, formant, and energy-related features [15–17]. In addition, the cepstrum feature represented by MFCC (Mel-Frequency Cepstrum Coefficients) [18] is often used in the research of speech emotion recognition. Eyben et al. make the detailed research on audio emotion features, and construct a concise feature set named GeMAPS, which involves 62 audio features consisting of frequency parameters, energy parameters and spectrum parameters) [4].

Comparatively, the common facial expression features can be divided into two kinds, i.e. appearance feature and geometrical feature [19,20]. The appearance feature is gained via applying image filters such as Gabor wavelet for local areas of the full face . The geometrical feature represents the LBP (local binary patterns) [21] and LBP (local binary patterns) [21] of shape and position of face components, e.g., eyebrow, eye, nose and mouth.

For the unimodal emotion recognition, it is required to learn the emotion recognition model through machine learning method after completion of feature extraction based on certain emotional corpus. The common methods include support vector machine (SVM), support vector regression (SVR), long short term memory recurrent neural network (LSTM-RNN), hidden Markov model (HMM), Gaussian mixture model (GMM), and artificial neural network (ANN), etc. (audio: [22–25]; visual: [26,27]).

In the multi-modal emotion recognition, it is required to extract the emotion features, and then fuse the emotion information to various degrees. Most of the fusion works focus on four strategies, i.e. feature fusion, decision fusion, score fusion, and model fusion. However, these methods are mostly shallow fusion, and fail to model the complex non-linear correlation between the multi-modal information, so it is necessary to design more complete fusion model [28].

To solve the problems existing in the feature extraction and multi-modal fusion better, the deep learning technology [29,30] applied to various fields can play a important role. By virtue of the available large-scale effective training data set, the deep learning technology shows the ultra-strong ability of feature learning and dimensionality reduction in the fields of image processing, speech recognition as well as natural language processing [31,32]. Among these technologies, CNN (convolutional neural network) is one of the representative technologies which plays an important role in the history of deep learning. Characterized by sparse interaction, parameter sharing, and uniformly varying representation, it shows excellent performance in feature extraction of data with specific grid structure, for example, image data. In addition, the deep belief network consisting of multi-layer restricted Boltzmann machines can be used as a deep multi-modal emotion feature fusion model.

In addition, when extracting emotion features for silent periods of audio data, the results of computing usually close to zero. This will result in noises and data pollution. Similarly, the corresponding facial expressions without voice are usually ‘static’, i.e., most of them stay the same. That can be view as a kind of redundancy. Most of the related work ignored this problem. Eyben et al. make the related research on GeMaps, but they only incorporate the average length of silence segment and voice segment and the standard deviation into 62 features simply [4]. In fact, this will lead to some noises. Han et al. use the extreme learning machine (ELM) for the voice emotion recognition based on the handcraft features, and consider the difference between silence segment and non-silence segment, but that is just not a task of multi-modal emotion recognition [24]. the direct using of mel-spectrograms of raw audio and facial expressions will result in performance loss in emotion recognition.

We firstly do voice activity detection (VAD) for original audio in order to distinguish whether an audio frame is silent, and assign emotion weights of 0 and 1 for silent frame and voice frame as well as the corresponding facial expression frames in visual data. Simply, when a certain time interval of audio/visual data is assigned a weight of zero, we discard it. And then, the mel-spectrogram and its first two order differentials are computed. Next, we encapsulate audio/visual segments following the format suitable for the corresponding feature extractors; audio-net and visual-net respectively. Finally, the whole AVEF model is trained based on the selected multi-modal emotion data set.

The remaining part of this paper includes the following sections. Section 2 introduces our method in details. Section 3 illustrates the emotion data set we used. Section 4 describes our experimental details, results and corresponding analysis. We summarize the full work in Section 5. The future work is discussed in Section 6.

2. AVEF method

As shown in Fig. 1, AVEF method consists of three stages, i.e. data preparation, feature learning, and multi-modal fusion.

(1) data preparation stage: In order to maximize the information content and meet the input demands of feature learning networks, we carry out some necessary pre-processing for raw visual and audio data in the selected corpora. (2) feature learning stage: Feature learning stage is composed of two convolutional neural networks (CNNs), i.e. audio-net used for learning the emotion features in the voice, and visual-net for facial expression images. (3) multi-modal fusion stage:In multi-modal fusion stage, we employ a deep belief network (DBN) to fuse the audio emotion features and visual emotion features for an audio/visual segment, fuse emotion multi-modal features of all segments in each certain video flip via average pooling, and further get the final emotion estimation by a using a support vector machine (SVM) model.

In Sections 2.1–2.2, we will discuss the details of these three stages respectively.

2.1. Data preparation

Among all data sets, the length of emotion video samples is generally different. It means that the duration of audio data and the quantity of face frame in every video clip are usually not the same. Thus, we should not regard the entire video clip as the basic unit for emotion analysis. Furthermore, handling a full video clip also can lead poor real-time performance in real-world application. Therefore, we need segment audio/visual data with an appropriate time duration. In previous researches, some segmenting schemes are used, such 255 ms, 655 ms. And the scheme of 655 ms is being proved with better performance in multi-modal emotion recognition [33]. Therefore, we also use 655 ms audio/visual segments. As is shown in Fig. 1, the audio-network needs inputting the image formatted mel-spectrograms (including the first and the second differential) of the audio segment. The visual-network needs inputting the facial expression image sequence of facial expression in the video clips.

Fig. 2 describes data preparation stage with an example sample. The

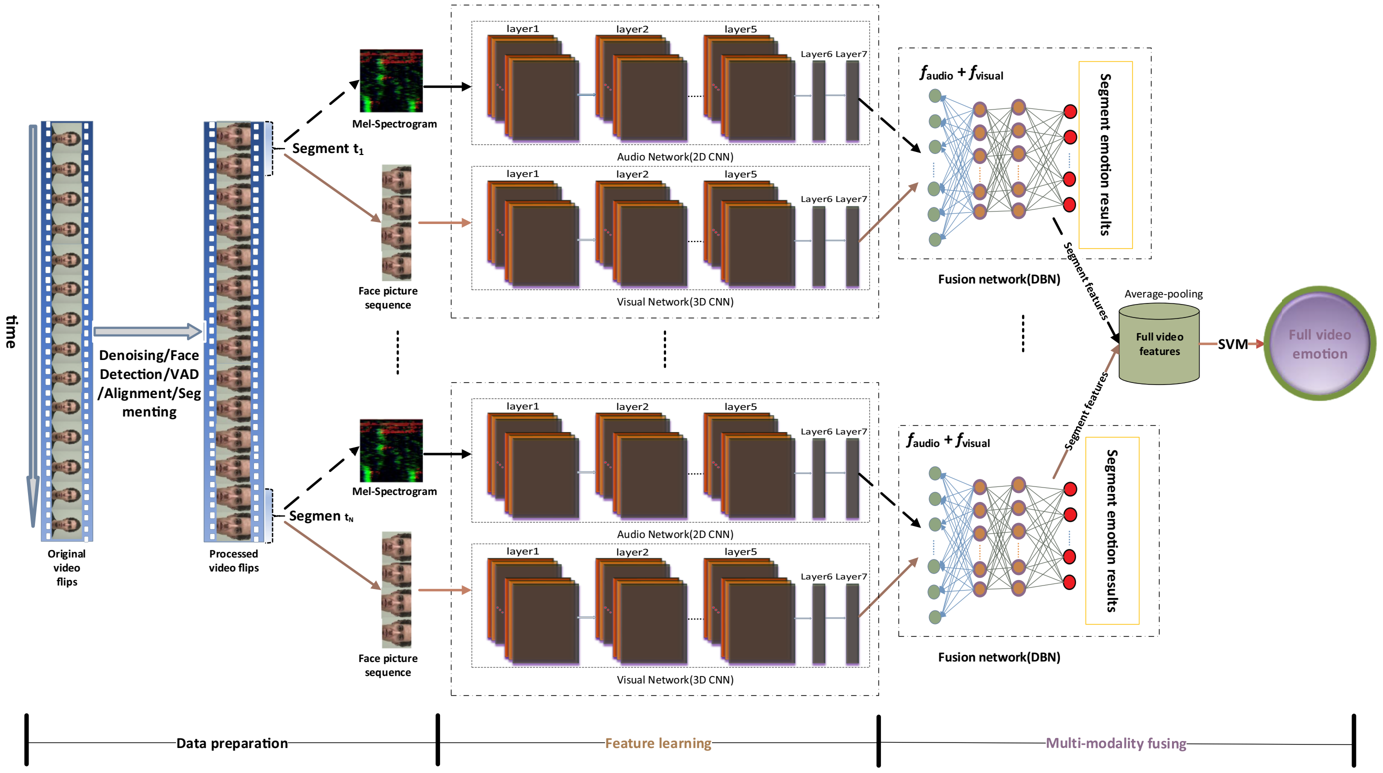

Fig. 1. AVEF method: AVEF method consists of three stages; datapreparation, feature learning and multi-modality fusion.

example sample comes from enterface05 data set, and is labeled as happiness. Rows 1–3 represent its audio wave, mel-spectrograms and visual sequence before data preparation stage. After the process including denoising, face detection, voice activity detection (VAD) and segmenting, they become what in rows 4–6. The original video has a time duration of 2.7 s. After data preparation, it is shortened into 1.6 s. Sections 2.1.1 and 2.1.2 discuss the preparation of audio and visual data in details.

2.1.1. *Audio data*

For the convenience of the subsequent feature calculation, it is very necessary to make some preprocessing for the voice sequence, involving sampling rate unification and channel conversion. Firstly, it is required to convert the voice sequence into the monophonic signal. The specific voice sequence X is generally composed of one or two channels, i.e. $X = [X_1;...;X_{Ch}]$. *Ch* refers to the number of channels, generally being 1 or 2. In addition, every X_i, $i \in \{1, ..., Ch\}$ is a vector with the same dimensionality. We need the voice sequence with only one channel in order to achieve higher computational efficiency, so we should make the following preprocessing of $X^S(n) = \frac{1}{Ch}\sum x_c(n)$, $c \in \{1, \ldots, Ch\}$. And then, X^S will go through a process pipeline including pre-emphasis, VAD, segmenting, and Mel-spectrogram computing and imaging. The whole pipeline is as follows:

Pre-emphasis : It is required to evaluate mel-spectrogram on the basis of original voice sequence, so we firstly make the pre-emphasis processing for the original audio sequence, which is equivalent to a high-pass filter and results in decreasing noises with lower frequency without leading to obvious voice distortion. Supposing the voice sequence is $x(n)$ before pre-emphasis and the voice sequence is after pre-emphasis $x_p(n) = x(n) - \mu{*}x(n-1)$. Here $\mu \in [0, 1]$, 1 standing for the strongest and 0 the weakest. We adopt 0.97 here. The pre-emphasis can eliminate the sub-bass effect [34].

VAD & Weight assigning : The silence segment will result in data pollution to a large extent. The calculation of Mel-spectrograms actually deems a voice frame as the atomic unit, so the calculation result of silent frames will be close to 0. Equivalently, a lot of nearly identical data are labeled differently, leading to serious data pollution in the training data. Also the same pollution exists in the test data. Segbroeck et al. put forward a comparatively robust VAD method [35]. Following their method, we use a sliding window to extract the MRCG (Multi-Resolution Cochleagram) feature of voice sequence, and perform VAD process by DNN. Based on the distinguishing results, the emotion weight is assigned for silence frames and non-silence frames, 0 and 1 respectively. We make the simple test on Emo-DB [36], and the results prove that VAD can improve the global accuracy of emotion recognition by 19.7% relatively. VAD results prove that the portion of voice frames in Emo-DB is 70.3%. Of course, VAD results could be changeable with the various related parameters.

Segmenting : We divide the continuous voice sequence into several segments. To meet the input demands of audio-network, the number of frames in a segment is set as 64 to reach a segment with 655 ms. In addition, to keep the quasi-steady state between segments, we will have 30-frame overlap between the adjacent segments. The framing parameters are as follows, with the frame length of 25 ms and frame shift of 10 ms. So we get 655 ms = 25 + 63 × 10 Every frame uses the Hamming window for smoothing.

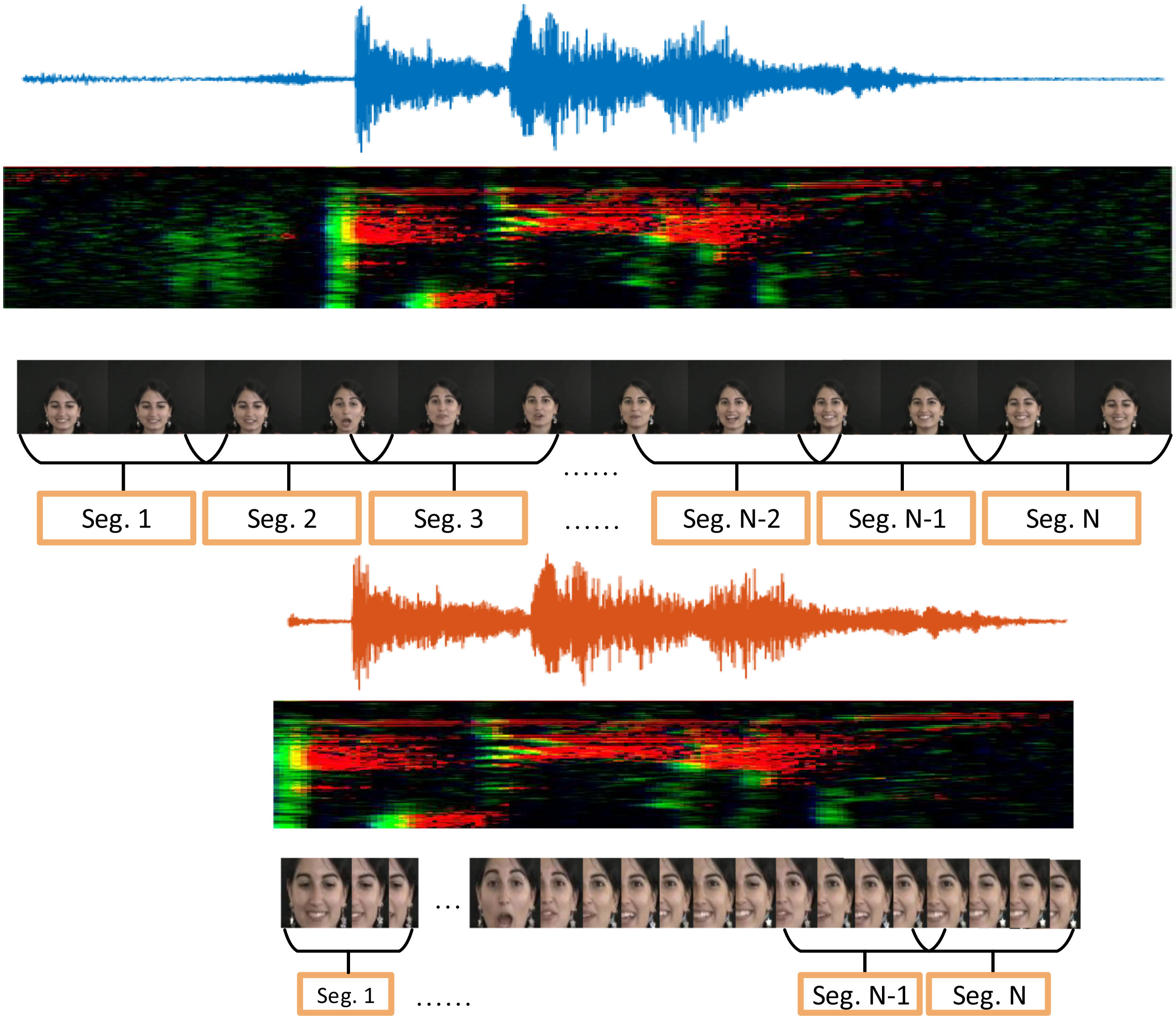

Fig. 2. Rows 1–3: audio wave, Mel-spectrogram and picture sequence in the raw video; rows 4–6: audio wave, Mel-spectrogram and facial expression sequence in the video clips after data preparation.

After pre-emphasis, VAD & Weight Assigning and Segmenting. It is necessary to calculate the short-time Mel-spectrogram with 64 Mel band bins as well as its first and second order differentials form 64 × 64 × 3 array. Finally they are combined into a matrix of 64 × 64 × 3.

Mel-spectrogram of every frame is calculated as described in Eq. (2.1).

$$MelSpectrogram_{frame_t} \approx \log(melbank(64)*(abs(rfft(frame_t))^2 + 0.01) \tag{2.1}$$

Where $frame_t$ represents the t voice frame, and $melbank(64)$ stands for converting hertz frequency within 0 8000 Hz into Mel scale applying 64 Mel bins, $rfft$ represents the fast Fourier transform. abs is the amplitude of a complex number, plus 0.01 means to prevent taking the logarithm for 0.

The dimension of $MelSpectrogram_{frame_t}$ obtained from the above process is 64 × 1. Then we compute the first and second order differential of Mel-spectrogram of all frames in each segment and obtain a 64 × 3 matrix. Extending the calculation results can obtain the feature matrix of 64 × 64 × 3 for each audio segment. As for the calculation mode of first order differential, we adopt the common method in the voice recognition task [37], as described in Eq. (2.2).

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.2}$$

Where $N = 2$. The second order differential of mel-spectrogram is computed in the same way.

Now extended Mel-spectrograms are obtained, with the dimension of 64 × 64 × 3. Such data can be regarded as an image representation of audio data as is shown in Fig. 3(a). Additionally, Fig. 3(b)–(d) are static Mel-spectrogram, the first and the second differentials respectively. In such images, the width direction represents time, the height direction represents frequency in Mel scale, and the pixel represents the corresponding amplitude. Obviously, Mel-spectrogram is much more different from liner spectrogram that has a clear harmonic structure. However, we can still find useful patterns in Fig. 3(a)–(d). (b) has a clear structure in width direction, while (c) and (d) have a clear structure in height direction. (a) is the combination of (b)–(d). The following contents of this paper will denote the image of extended Mel-spectrograms as a, to be deemed as the input of audio-network.

2.1.2. Visual data

Generally, in the audio-visual emotion data set widely used, the video is composed of image frames involving faces. But in most cases,

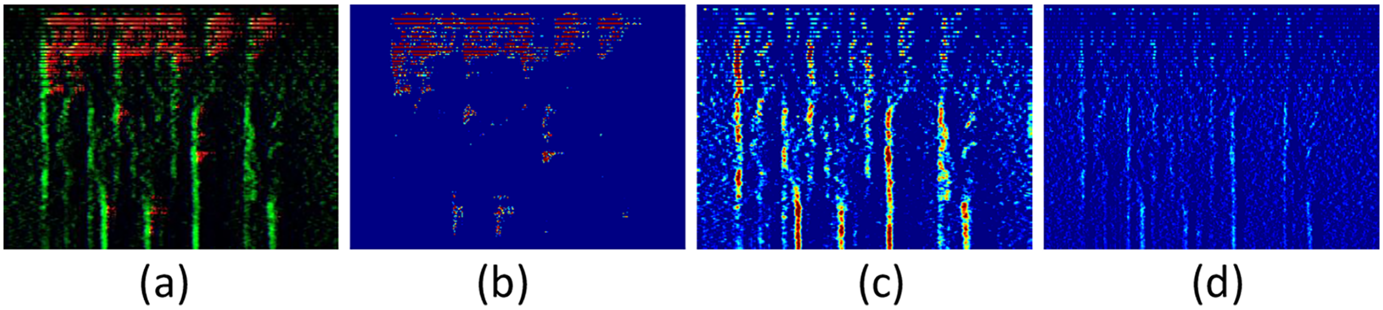

Fig. 3. Mel-spectrograms: (a) mel-spectrogram including its first and second order differentials (each of them is a color channel); (b) static mel-spectrogram; (c) first order differentials of (b); (c) second order differentials of (b).

these image frames have large-area background, either being irrelevant to the theme we concentrate on, or misleading us when doing feature learning. To reduce such side effect and promote the ultimate model performance as much as possible, it is required to carry out necessary pre-processing. Thus, conduct face detection via detecting four key positions of left eye, right eye, nose and mouth, and obtain the human face through an extending for these key points.

Now we have done most of the data preparation except segmenting.

The above processing segments the voice data, with the length of 655 ms per segment. We mark the time range $[T_{begin}, T_{end}]$ in the original video, and then encapsulate the image sequences within this time interval into a corresponding visual segment. For example, the time interval of some voice segment is [100 ms, 755 ms], then in the video of 30 frames/s, the corresponding number of images should be $int(0.655 \times 30) = 20$. The number of images required for the input of visual-network is 16, so we remove the first two and the last two face images in every visual segment, to ensure that the visual segment has continuous 16 face frames. In the other case, if the number of frames in a visual segment is less than 16 under smaller video frame rate, supposing it has only 12 images, it is required to simply repeat the first two and the last two frames, and the ultimate visual segment is the continuous 16 face frames. Supposing the face image we get is $173 \times 173 \times 3$, the face expression sequence will be a tensor of $16 \times 173 \times 173 \times 3$. Of course, this tensor will be resized through with bilinear interpolation to suit visual-network if necessary. Corresponding to the voice data, we denote the tensor standing for the sequence of facial expressions as v, which will be regarded as the input of visual-network.

2.2. AVEF model learning

At the data preparation stage, we carry out the processing of video including pre-emphasis, VAD, face detection, and segmentation. Each audio/visual segment is marked with the same emotion label.

As shown in Fig. 1, the learning model includes two connected stages, i.e. feature learning stage and multi-modal fusion stage. The feature learning stage includes the audio-network and visual-network. The multi-modal fusion stage is composed of segment fusion network and global fusion model. Audio-network is a 2D convolutional neural network and visual network is a 3D convolutional neural network. The segment fusion network is a deep belief network (DBN), and the global fusion model is a multi-class support vector machine(SVM).

The output features of audio-network and visual-network are feeded into the DBN with a certain portion like 1:1 to learn the segment-based emotion features. The output of deep belief network is deemed as the multi-modal emotion features of a audio-visual segment, and then average pooling processing is conducted for the multi-modal emotion features of all segments in a full video. Finally, the SVM deems the average pooling result as the input for the emotion analysis to get final recognition result.

In practice, we choose a strategy of transfer learning for tackling the lack of labeled data [38,39]. In related works, similar strategy was also used. For example, Zhang etc. use ImageNet-based AlextNet [40] and C3D-Sports-1M model [41] to initialize audio-network and visual-network respectively [33]. However, there is a large domain gap in these two transferring. Firstly, ImageNet consists of massive images which are much different from Mel-spectrogram; and secondly, the data set used for training C3D-Sports-1M model is also much different from human facial expressions. Therefore, to cope with these domain gaps, we do transfer learning just among the multi-modal emotion data set. For example, If we want to train a model in RML data set, we first initialize audio-network, visual-network and the deep belief network via Enterface05 data set and BAUM-1s data set, And then, keep the shallow layers of each network and train the deep layers from scratch in the target data set RML. It is the similar case in other target data sets.

In the remained of this section, we discuss these four models in details.

2.2.1. Audio-network

The audio-network is a 2D CNN for audio emotion feature extracting. The architecture of audio-net we choose is Alexnet [40]. It has 8 layers, involving 5 convolutional layers (Conv1-Conv2-Conv3-Conv4-Conv5) , 3 fully connected layers (fc6-fc7 -fc8) and 3 max-pooling layers (pool1-pool2-pool5).

We denote the audio-network model as $A(a; \theta^A)$, where a is the input variable, and θ^A represents the hyper-parameter set of audio-network. It is noted that we modify the number of neurons in the last layer (fc8) from 1000 to C. C indicates the number of categories to be classified. For discrete model of six emotions, C is 6 since the data set we use consist of 6 different emotion states; for binary valence-arousal dimensional emotion model, C equals 2 because we acquire a binary scheme (low and high for arousal, positive and negative for valence).

The output of the last layer (fc8) represents the probabilistic estimation for classification as described in Eq. (2.3).

$$\hat{y}_i^A = A(a_i; \theta^A), \ i \in \{1, 2, ..., K\} \tag{2.3}$$

K is the number of samples. $\hat{y}_i^A$ is a vector of $C \times 1$. Each element corresponds to the prediction result of audio segment sample a_i as the probability of a certain emotion. Thus:

$$\sum_{j=1}^{C} \hat{y}_i^A(j) = 1 \tag{2.4}$$

Based on pre-trained Alexnet (as above, modified fc8 layer), we make the further fine tuning on the target emotion audio data set. The overall fine tuning process is to use the back propagation algorithm combined with the Stochastic Gradient Decreasing to adjust the audio-network parameters θ^A. Actually is to solve the following optimization problems as described in (2.5).

$$\underset{\theta^A}{\text{argmin}} \sum_{i=1}^{K} L(\hat{y}_i^A, y_i) \tag{2.5}$$

Table 1
Dataset.

	A/S	Fs_audio	Channel_audio	Fps_visual	Size_visual frame	Language	Subject
RML	A	22,050	Stereo	30	368 × 240 × 3	6	8
Enterface05	A	48,000	Stereo	25	720 × 576 × 3	1	43
BAUM-1s	S	48,000	Stereo	30	854 × 480 × 3	1	31

L is the cross entropy between the estimation result and the true distribution as described in (2.6), where i is the index of samples, j is the index for the category a given sample belonging to, y_i stands for the ground truth probable distribution of the ith sample, and $\hat{y}_i^A$ for the predicted one. Both y_i and $\hat{y}_i^A$ is a vector of $C \times 1$. y_i is with the form of one-hot, while $\hat{y}_i^A$ is the result of softmax.

$$L(\hat{y}_i^A, y_i) = -\sum_{j=1}^{C} y_i(j)\log(\hat{y}_i^A(j)) \tag{2.6}$$

2.2.2. Visual-network

The visual-network is a 3D CNN for facial expression feature extracting. The architecture of audio-network we choose is C3D-Sports-1 model [41]. It has to 8 layers, involving 8 convolutional layers (Conv1a-Conv2a-Conv3a-Conv3b-Conv4a-Conv4b-Conv5a), 3 fully connected layers (fc6-fc7 -fc8) and 5 max-pooling layers (pool1-pool2-pool3-pool4-pool5).

We denote the visual-net as $V(v; \theta^V)$, where v is the input variable, and θ^V represents the hyper-parameter set of visual-network. We modify the number of neurons in the last full connected layer (fc8) from 487 to C. The output of the last layer (fc8) represents the probabilistic estimation for the classification.

The theoretical description of modeling and optimization of visual-network is similar to audio-network as demonstrated in (2.3)–(2.6).

2.2.3. Segment-based fusion

We use the deep belief network (DBN) to complete segment-based fusion for audio-visual emotion features from audio-network and visual-network. DBN can learn highly non-linear relation in multi-modal emotion features. As shown in Fig. 1, the deep belief network is composed of an explicit layer, two implicit layers and an output softmax layer. The implicit layer is used for feature extracting, and the explicit layer is used for receiving the input. The input is from fc7 layer of audio-net(f_{audio}) and visual-network(f_{visual}), and these two feature vectors form a feature of 8192, so the input layer of deep belief network includes 8192 neurons. Two implicit layers respectively include 4096 neurons and 2048 neurons. The number of neurons in the output layer is equal to the total number of emotion categories which is denoted as C above. Therefore, the architecture of DBN is 8192-4096-2048-C. For a global optimization in segment-based fusion, we train audio-network, visual-network and deep belief network as a whole model. The method of training the deep belief network abides by Hinton et al., i.e. firstly using the greedy learning algorithm layer by layer for the pre-training of network, as a non-supervised mode. Then utilize the output features of audio-network and visual-network to conduct the supervised training.

2.2.4. Video-based fusion

In Section 2.2.3, the deep belief network (DBN) conducts deep non-linear fusion for emotion features learned by audio-network and visual-network, and stores multi-modal emotion features with the dimensionality of 2048 in the last implicit layer. Finally, we build a SVM model, and complete the last fusion for the full video.

As shown in Fig. 1, we firstly conduct the average pooling for the features of all segments in the second implicit layer of DBN, and then train a multi-class support vector machine(SVM) model by the pooling results to complete the final classification. In fact, more than one classifier can be used at this stage, such as neutral networks, extreme learning machine, and SVM. The average performance of these three methods is almost the same for the data set size and feature dimensions at this stage, but SVM is faster, more stable and easier to implement. Specifically, the final data set size is within one thousand and the feature dimension is 2048. So SVM is selected. The kernel function we used is polynomial kernel function.

3. Dataset used

The audio-visual emotion data sets we select to evaluate AVEF method are RML data set [42], Enterface05 data set [43] and BAUM-1s data set [44]. We put forward the multi-modal emotion fusion network in three public audio C video multi-modal emotion data sets, including RML data set (performance), enterface05 data set (performance), and BAUM-1s data set (natural).

Table 1shows three data sets we use. A/S stands for whether a data set is acted or spontaneous. *fsaudio* is the sample rate of the audio data. *fpsvideo* is the measure of video via *fps*(*framepersecond*). *sizevisualframe* is size of visual frame. *language* is the number of languages contained in a data set. *speakers* stands for the number of participants in the data set. Detail description is as follow:

RML data set : RML data set includes 720 clips of videos. There are 8 different speakers, and 6 languages including English, Chinese Mandarin, Urdu (Pakistan), Punjabi (India), and Italian. It includes 6 basic emotions(anger, disgust, fear, happiness, sadness, and surprise). The audio sample rate is 22,050 Hz. All emotions are acted by the participants. The video frame rate is 30 fps, and the image size is 720*576*3.

Enterface05 data set : Enterface05 data set is composed of 1290 video clips. There are 43 different speakers who all speak English. It includes 6 basic emotions as same with RML data set. The audio sample rate is 48,000 Hz. All emotions are emotions acted by the participants. The video frame rate is 25 fps, and the image size is 368 × 240 × 3.

BAUM-1s data set : BAUM-1s data set includes 1222 segments of video clips. There are 31 different speakers who all speak Turkish. In addition to 6 basic emotions included in RML and Enterface05, it also includes emotions categories such as boredom and contempt, as well as 4 mental states of uncertainty, thinking, and concentrating, etc. To keep pace with the previous two data sets, we choose 521 video samples involving the above 6 basic emotions. The audio sampling rate of the data set is 48,000 Hz. The video frame rate is 30 fps, and the image size is 854 × 480 × 3. All emotions are naturally expressed by the participants under stimulation.

4. Experiments and analysis

In this section, experiments based on the three public data set mentioned above are explored. We will discuss the emotion model, experimental setup, and the results and analysis.

4.1. Emotion model used

In practice, we use the two popular emotion models: discrete emotion model and binary dimensional emotion model. In our case, the discrete emotion model is composed of six emotion categories including **anger, disgust, fear, happiness, sadness** and **surprise**. And then, discrete emotion categories are mapped into binary arousal labels and binary valence labels as shown in Table 2. The mapping method is similar to [4].

4.2. Experiment setup

In AVEF model, audio-network, visual-network and segment fusion network are completed based on TensorFlow.[1] The final emotion classification applies the libsvm toolkit [45] using polynomial kernel function. In pre-training stage, audio-network and visual-network are trained separately. And then they are trained incorporated with DBN as a whole model. For discrete emotion model, six different emotion labels are used. And for binary dimensional emotion model, the labels are positive/negative and strong/weak respectively. The detail of experiment setup is shown in Table 3.

All related training set and testing set are in a ratio about 7:3. To ensure that the speakers in training set are not in the corresponding test set, i.e., the subject-independent strategy, we use the LOSO strategy for data set with fewer speakers, RML data set for example. Here LOSO means **Leave One Speaker Out**. Comparatively, for Enterface05 data set and BAUM-1s data set with more speakers, we adopt **LOSGO(Leave One Speaker Group Out)** strategy.

4.3. Results and analysis

In this section, we demonstrate the experimental results on the three public data set and analyse the corresponding results to give an illustration of the classifying performance. Also, the time complexity of the whole system is analysed.

4.3.1. Classifying performance

Multiple comparison experiments have been conducted, including unimodal emotion recognition on discrete model, unimodal emotion recognition on binary dimensional model, multi-modal emotion recognition on discrete model, and multi-modal emotion recognition on binary dimensional model. In each of these four experiments, we make comparison for scheme with (***case 2***) or without (***case 1***) cross-modal pollution decreasing and redundancy reduction as described in Section 2.1. Global accuracies of each experiment are shown in Tables 4 − 7 respectively.

Table 4 is the result of global accuracies in unimodal emotion recognition experiments when discrete emotion model is used.

Table 5 is the result of global accuracies in unimodal emotion recognition experiments when binary arousal-valence dimensional emotion model is used.

Table 6 is the result of global accuracies in multi-modal emotion recognition experiments when discrete emotion model is used. It is worth noting that, after cross-modal denoising and removing redundancy, the scale of the data set is decreased: RML datasets are changed to 72% of the total number of samples in raw data, for Enterface05 data set is 67%, and BAUM-1s 76%.

Table 7 is the result of global accuracies in multi-modal emotion recognition experiments when binary arousal-valence dimensional emotion model is used.

Table 2
Mapping of emotion categories to binary arousal labels (low/high) and binary valence labels (negative/positive).

arousal	**low**	**disgust, sadness**
	high	**anger, happiness, fear, surprise**
valence	**negative**	**anger, fear, disgust, sadness**
	positive	**happiness, surprise**

Table 3
Experiment setup.

Hard Ware Platform	CPU	Intel(R) Core(TM) i7-5820K CPU @ 3.30 GHz		
	Memory	64GB		
	GPU	NVIDIA GTX TITAN XP (12GB memory)		
Training setup		**Audio-network**	**Visual-network**	**the whole model**
	Batch Size	30	30	10
	Number of Epochs	500	500	500
	Dropout Parameter		0.3	
	Stochastic Momentum		0.9	
	Learning Rate		0.001	

As shown in Tables 4–7, the performance of case 2 outperforms that of case 1 since cross-modal pollution decreasing and redundancy reduction. Besides, the recognition performance of the proposed AVEF method is also better than other works that utilize handcrafted features [42–44].

The merits above are average results over all emotion categories. In addition, to explore the recognizability of each individual emotion, we give the corresponding confusing matrices covering all emotion categories, as is shown in Table 8–10. All rows represent real labels, and all columns represent predicted labels.

Table 8–10 show that: emotions with higher intensity seem to be easier to identify, such as anger and happiness. Moreover, emotions with similar intensity are easily misclassified, for example, anger and happiness, fear and disgust.

4.3.2. Time complexity

Under our experimental configuration, it takes nearly one day to train a multi-modal emotion recognition model. Of course, the real time cost will depend on the actual hardware configuration and specific experimental parameter settings. Once the model deployment is completed, it will give the corresponding result of emotion recognition when feeding a video clip. In our case, when the data preparation is done, i.e., the images of facial expression and the corresponding Mel-spectrogram (including the differentials), a video clip with duration of 5 s can be processed within 1 s. It just looks fine and satisfy the real-time requirements. However, the data preparation process actually takes a lot time, because it is completed via MATLAB 2017. In fact, the data preparation process often takes 10 times of the time duration of a video clip. Therefore, the real-time performance of the whole system needs a lot of efforts.

5. Summary

We propose the deep weighted fusion architecture for multi-modal emotion recognition in this paper. Firstly, we conduct the cross-modal noise modeling for the multi-modal data, and eliminate most of the data pollution in audio data and most of the data redundancy in visual data.

The AVEF model includes four parts, i.e., audio-network, visual-network segment fusion model and global fusion model. The audio-net

[1] https://www.tensorflow.org/.

Table 4
Unimodal classification for six discrete emotions.

	RML_audio	RML_visual	Enterface05_audio	Enterface05_visual	BAUM-1s_audio	BAUM-1s_visual
case1	68.23	71.18	80.36	55.26	39.48	52.41
case2	71.26	73.88	81.41	58.19	42.38	54.69

Table 5
Unimodal classification of binary dimensional emotion model.

	RML_audio	RML_visual	Enterface05_audio	Enterface05_visual	BAUM-1s_audio	BAUM-1s_visual
case1	(79.2,81.3)	(82.5,85.4)	(88.3,87.1)	(83.7,81.6)	(68.2,66.2)	(76.3,75.8)
case2	(81.5,83.6)	(85.1,84.1)	(89.4,88.1)	(84.6,83.9)	(73.1,74.5)	(80.1,77.8)

Table 6
Multi-modal classification for six discrete emotions.

	RML	Enterface05	BAUM-1s
case1	80.46	83.94	57.61
case2	82.38	85.69	59.17

Table 7
Multi-modal classification for binary dimensional emotion model.

	RML	Enterface05	BAUM-1s
case1	(83.1,87.9)	(91.2,88.6)	(77.3,79.2)
case2	(86.6,90.1)	(92.3,91.8)	(80.5,82.3)

Table 8
Confusion matrix of multi-modal classification for six discrete emotions on RML data set.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
anger	**91.13**	0.00	1.13	5.52	1.37	0.85
disgust	1.22	**79.45**	6.92	3.35	7.92	1.14
fear	1.32	1.01	**77.90**	1.62	13.5	4.65
happiness	7.65	2.38	1.54	**86.7**	1.27	0.46
sadness	2.84	3.64	12.55	3.64	**76.12**	1.21
surprise	2.35	6.32	7.99	1.34	1.40	**80.60**

Table 9
Confusion matrix of multi-modal classification for six discrete emotions on enterface05 data set.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
anger	**90.25**	0.56	0.31	5.52	1.16	2.20
disgust	1.28	**82.36**	9.30	0.95	5.37	0.74
fear	1.24	6.51	**80.22**	1.28	8.75	2.00
happiness	5.62	1.86	1.13	**89.01**	1.25	1.13
sadness	1.58	6.75	2.68	1.18	**84.95**	2.86
surprise	5.63	2.29	3.08	6.82	1.56	**80.62**

Table 10
Confusion matrix of multi-modal classification for six discrete emotions on BAUM-1s data set.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
anger	**65.32**	3.35	4.21	15.32	4.98	6.82
disgust	5.66	**56.11**	10.58	6.89	13.11	7.65
fear	6.52	10.85	**58.62**	7.95	10.32	5.74
happiness	12.96	6.23	5.85	**60.25**	7.62	7.09
sadness	7.59	10.33	14.65	8.96	**53.18**	5.29
surprise	11.33	5.68	6.01	15.59	2.39	**59.00**

and visual-net are respectively 2D CNN and 3DCNN. They are used as emotion feature extractors. Then we use DBN for highly non-linear fusion of the emotion features learned by the above two feature extractors, and finally carry out the emotion classification by a support vector machine.

Experiment results show that: (1) CNN-based feature extraction outperforms traditional handcraft features in emotion recognition task; (2) CNN and DBN based feature extraction and highly nonlinear feature fusion scheme can effectively improve the efficiency of emotional feature fusion in audio-video multi-modal emotion recognition. (3) The method that conducting VAD for video streams, segmenting the voice segments and aligning them into corresponding facial expression sequence, can effectively reduce audio data pollution and visual data redundancy, and hence improve the performance of emotion recognition. (4) Using transfer learning on corpus in closer domains can effectively solve the problem of insufficient data in large deep network and can also speed up the training process.

6. Future work

In our current work, CNN is used as feature extractor. 3D CNN can effectively extract spacial and latent time memory in continuous image frames, but 2D CNN can only extract spacial invariant features in image of Mel-spectrogram. We will further study the combined use of 2D CNN and LSTM for feature learning and emotion recognition of audio sequence.

In addition, though we employ transfer learning strategy on audio-visual emotion data sets, there is still a large culture gap. To cope with this problem, we will explore more efficient methods of audio-visual emotion recognition from two aspects: (1). developing larger data set on a certain cultural backgrounds; (2). trying emotion recognition model with more adaptive ability when faced with cultural gap.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61672246, No. 61272068 and the Fundamental Research Funds for the Central Universities under Grant HUST:2016YXMS018, and by Slovenian Research Agency under grant “Algorithms and optimization methods in telecommunications” no P2-0246. In addition, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. Yaxiong would also like to express his gratitude to the helpful discussion from Dr. Aharon Satt. Hubei Provincial Key Project under grant 2017CFA051, the Applied Basic Research Program through Wuhan Science and Technology Bureau under Grant 2017010201010118, and the National Natural Science Foundation of China (Grant No. 61572220).

References

[1] H.D. Garfinkel, S.N. Critchley, Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on:anterior insular cortex mediates bodily sensibility and social anxiety by Terasawa et al. (2012), Soc. Cogn. Affect Neurosci. 8 (2013) 231–234, http://dx.doi.org/10.1093/scan/nss140.

[2] R. Fernandez, R. Picard, Analysis and classification of stress categories from drivers' speech, 2000, M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 513.

[3] J. Healey, J. Seger, R. Picard, Quantifying driver stress: developing a system for collecting and processing bio-metric signals in natural situations, Biomed. Sci. Instrum. 35 (1999) 193–198.

[4] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The Geneva minimalistic acoustic parameter set (geMAPS) for voice research and affective computing, IEEE Trans. Affect Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[5] P. Patel, A. Chaudhari, R. Kale, M.A. Pund, Emotion recognition from speech with Gaussian mixture models & via boosted GMM, Int. J. Res. Sci.Eng. 3 (2017).

[6] B.W. Schuller, Intelligent-Audio-Analysis, Springer, 2013.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2016), pp. 5200–5204.

[8] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features, and methods, Speech Commun. 48 (9) (2006) 1162–1181.

[9] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect Comput. 6 (1) (2015) 69–75, http://dx.doi.org/10.1109/TAFFC.2015.2392101.

[10] M.S. Bartlett, G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: development and applications to human computer interaction, 2003 Conference on Computer Vision and Pattern Recognition Workshop, vol. 5, (2003). 53–53. doi:10.1109/CVPRW.2003.10057 .

[11] A.T. Lopes, E. de Aguiar, A.F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit. 61 (Supplement C) (2017) 610–628, http://dx.doi.org/10.1016/j.patcog.2016.07.026. www.sciencedirect.com/science/article/pii/S0031320316301753 .

[12] P.K. Manglik, U. Misra, Prashant, H.B. Maringanti, Facial expression recognition, 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), vol. 3, (2004), pp. 2220–2224. vol.3. 10.1109/ICSMC.2004.1400658 .

[13] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816, http://dx.doi.org/10.1016/j.imavis.2008.08.005. www.sciencedirect.com/science/article/pii/S0262885608001844 .

[14] A. Wood, M. Rychlowska, S. Korb, P. Niedenthal, Fashioning the face: sensorimotor simulation contributes to facial expression recognition, Trends Cogn. Sci. 20 (3) (2016) 227–240, http://dx.doi.org/10.1016/j.tics.2015.12.010. www.sciencedirect.com/science/article/pii/S1364661316000164 .

[15] K.R. Scherer, Vocal communication of emotion: a review of research paradigms, Speech Commun. 40 (1) (2003) 227–256, http://dx.doi.org/10.1016/S0167-6393(02)00084-5. www.sciencedirect.com/science/article/pii/S0167639302000845 .

[16] C. Lee, S. Lui, C. So, Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition, J. Acoust. Soc. Am. 135 (4) (2014), http://dx.doi.org/10.1121/1.4878044. 2422–2422.

[17] C.-H. Wu, J.-F. Yeh, Z.-J. Chuang, Emotion Perception and Recognition from Speech, Springer: London, London, 2009, pp. 93–110, http://dx.doi.org/10.1007/978-1-84800-306-4_6.

[18] W. Han, C.-F. Chan, C.-S. Choy, K.-P. Pun, An efficient MFCC extraction method in speech recognition, 2006 IEEE International Symposium on Circuits and Systems, (2006), p. 4pp., http://dx.doi.org/10.1109/ISCAS.2006.1692543.

[19] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275, http://dx.doi.org/10.1016/S0031-3203(02)00052-3. www.sciencedirect.com/science/article/pii/S0031320302000523 .

[20] Y. Tian, T. Kanade, J.F. Cohn, Facial Expression Recognition, Springer: London, London, 2011, pp. 487–519, http://dx.doi.org/10.1007/978-0-85729-932-1_19.

[21] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928, http://dx.doi.org/10.1109/TPAMI.2007.1110.

[22] H. Hu, M.X. Xu, W. Wu, GMM supervector based SVM with spectral features for speech emotion recognition, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 4, (2007), http://dx.doi.org/10.1109/ICASSP.2007.366937. IV–413–IV–416.

[23] V. Garg, H. Kumar, R. Sinha, Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers, 2013 National Conference on Communications (NCC), (2013), pp. 1–5, http://dx.doi.org/10.1109/NCC.2013.6487987.

[24] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, 2014. www.isca-speech.org/archive/interspeech_2014/i14_0223.html.

[25] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis. Comput. 31 (2) (2013) 153–163, http://dx.doi.org/10.1016/j.imavis.2012.03.001. Affect Analysis In Continuous Input, http://www.sciencedirect.com/science/article/pii/S0262885612000285 .

[26] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816, http://dx.doi.org/10.1016/j.imavis.2008.08.005. www.sciencedirect.com/science/article/pii/S0262885608001844 .

[27] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2014).

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, the 28th International Conference on Machine Learning (ICML), 2011.

[29] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org .

[30] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436, http://dx.doi.org/10.1038/nature14539.

[31] D. Yu, L. Deng, Automatic Speech Recognition: A Deep Learning approach, Springer, 2014.

[32] Y. Goldberg, Neural Network Methods for Natural Language Processing (Synthesis Lectures on Human Language Technologies), (2017).

[33] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio-visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. PP (99) (2017). 1–1. 10.1109/TCSVT.2017.2719043 .

[34] S. Young, M.G. Gunnar Evermann, D.K. Thomas Hain, G.M. Xunying Liu, D.O. Julian Odell, V.V. Dan Povey, P. Woodland, HTK book (2009).

[35] M.V Segbroeck, T. Andreas, S.S. Narayanan, A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice, INTERSPEECH, (2013).

[36] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, Proceedings of Interspeech, Lissabon, (2005), pp. 1517–1520.

[37] X. Huang, A. Acero, H.W. Hon, Spoken Language Processing: A Guide to Theory, algorithm, and System Development, Prentice Hall PTR, 2001.

[38] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.

[39] L. Shao, F. Zhu, X. Li, Transfer learning for visual categorization: a survey, IEEE Trans. Neural Netw. Learn. Syst. 26 (5) (2015) 1019–1034, http://dx.doi.org/10.1109/TNNLS.2014.2330900.

[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf .

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, The IEEE International Conference on Computer Vision (ICCV), (2015).

[42] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals*, IEEE Trans. Multim. 10 (5) (2008) 936–946, http://dx.doi.org/10.1109/TMM.2008.927665.

[43] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 audio-visual emotion database, 22nd International Conference on Data Engineering Workshops (ICDEW'06), (2006). pp. 8–8. 10.1109/ICDEW.2006.145 .

[44] S. Zhalehpour, O. Onder, Z. Akhtar, C.E. Erdem, BAUM-1: a spontaneous audio-visual face database of affective and mental states, IEEE Trans. Affect Comput. 8 (3) (2017) 300–313, http://dx.doi.org/10.1109/TAFFC.2016.2553038.

[45] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm .