Human emotion recognition by optimally fusing facial expression and speech feature

Xusheng Wang, Xing Chen, Congjun Cao

PII: S0923-5965(20)30054-0
DOI: https://doi.org/10.1016/j.image.2020.115831
Reference: IMAGE 115831

To appear in: *Signal Processing: Image Communication*

Received date : 18 August 2019
Revised date : 21 January 2020
Accepted date : 4 March 2020

Please cite this article as: X. Wang, X. Chen and C. Cao, Human emotion recognition by optimally fusing facial expression and speech feature, *Signal Processing: Image Communication* (2020), doi: https://doi.org/10.1016/j.image.2020.115831.

Human Emotion Recognition by Optimally Fusing Facial Expression and Speech Feature

Xusheng Wang*
Xi'an University of Technology, Xi'an, China
xusheng.w@outlook.com
*Corresponding author

Xing Chen
Xi'an University of Technology, Xi'an, China

Congjun Cao
Xi'an University of Technology, Xi'an, China

Abstract: Emotion recognition is a hot research in modern intelligent systems. The technique is pervasively used in autonomous vehicles, remote medical service, and human-computer interaction (HCI). Traditional speech emotion recognition algorithms cannot be effectively generalized since both training and testing data are from the same domain, which have the same data distribution. In practice, however, speech data is acquired from different devices and recording environments. Thus, the data may differ significantly in terms of language, emotional types and tags. To solve such problem, in this work, we propose a bimodal fusion algorithm to realize speech emotion recognition, where both facial expression and speech information are optimally fused. We first combine the CNN and RNN to achieve facial emotion recognition. Subsequently, we leverage the MFCC to convert speech signal to images. Therefore, we can leverage the LSTM and CNN to recognize speech emotion. Finally, we utilize the weighted decision fusion method to fuse facial expression and speech signal to achieve speech emotion recognition. Comprehensive experimental results have demonstrated that, compared with the uni-modal emotion recognition, bimodal features-based emotion recognition achieves a better performance.

Keywords: Facial expression recognition; speech emotion recognition; bimodal fusion; feature fusion; RNN

1. Introduction

Emotion recognition plays a significant role in modern intelligent systems, such as autonomous vehicles, smart phone voice assistant, human psychological analysis, and medical services [1, 2, 3]. For example, drivers' emotion can be analyzed in real time by leveraging the speech emotion recognition system, which can judge whether the circumstance is safe or not. This can be used to warn drivers when they are in fatigue state and thus the traffic accidents can be avoided. In medical research, speech emotion recognition can be utilized to analyze emotional changes in depressive patients or autistic children, which is used as an effective tool for disease diagnosis and adjuvant treatment. Speech emotion recognition aims to effectively understand emotion status from the low-level features extracted from speech signals. It can be regarded as classification task based on speech signal sequences, which consists of emotional database compilation, speech emotional feature extraction, feature dimensionality reduction, and emotion classification/recognition. Traditional speech motion recognition techniques include hidden Markov model (HMM), artificial neural network (ANN), Gaussian mixture model (GMM), support vector machine (SVM), and K-nearest neighbor (KNN). However, the performances of these algorithms are significantly different since the corpus varies greatly. For example, the SVM and KNN based emotion recognition algorithms are generally with high certainty, whereas human emotions have strong complexity and high uncertainty. Therefore, they are deficiently effective in speech emotion recognition.

Both the physiological and psychological studies have demonstrated that facial expression and speech signal are informative for recognizing human emotion. Human beings express their different feelings by adjusting the strength of facial muscles and changing their tones. At the same time, human beings analyze the corresponding emotions by perceiving the speech signals. In addition, psychological experiments have shown that visual information will change speech perception. In this way, emotional types can be judged by visual and speech information. The database is the basis of emotional recognition research. Among all types of multi-modal emotional databases, the emotional database based on facial expression and voice is highly complete. For the early researches, Friesen et al. [4] pointed out that human facial expression can be classified into six emotional categories including happiness, anger, fear, sadness, disgust and surprise. And each emotion was related to facial expression. This research is the fundamental for emotion recognition. Combining with facial expression recognition, speech emotion recognition can achieve a better performance. Traditional RNN-based algorithms can fully exploit the contextual information to construct language models and achieve good performance in the field of emotional analysis. It is noticeable that, such recognition algorithm has the problems of gradient disappearance and explosion. To solve problems of the existing algorithms, in this paper, we propose a bimodal speech emotion recognition framework, an improved AlexNet to describe human facial expressions. Since speech signal are highly correlated at temporal-levle, we combine LSTM and CNN to recognize emotion based on human speeches.

2. Related work

Speech emotion recognition has becoming a well-known research topic in modern intelligent systems. Williams and Stevens studied the principle of speech production from a physiological point of view. When people are in the state of anger, fear or pleasure, the sympathetic nervous system will be triggered. At this time, the tone of human voices will be higher, the speed of the voice will be faster, and there will be more high-frequency energy. When people are in a sad state, the parasympathetic nervous system will be triggered, when the voice changes slowly, and contains very little high-frequency energy. Literature [6] indicated that there are three types of phonetic features that are generally considered by researchers to be closely related to human emotional expression: prosodic features, spectral features and phonetic quality features. Prosodic features, also known as super-phonological features, describe the phonological structure of language and are typical features of human natural language. Prosodic features are mainly embodied by intonation, intonation, stress and rhythm, which is one of the important features of language and emotional expression. Linear spectrum features include linear prediction coefficient (LPC), unilateral autocorrelation linear prediction coefficient and logarithmic frequency power coefficient. The performance of the three popularly-used cepstrum features is compared and analyzed by leveraging the hidden Markov model, and it is pointed out that LFPC can achieve the best effect [7]. Sound quality is a subjective evaluation index of human voice, which belongs to the true expression of human emotions. Therefore, the role of sound quality features in speech emotion recognition has also been recognized by many researchers. Commonly-used human speech quality features include formant frequency and glottic parameters.

In recent years, with the development of deep neural network, deep learning-based speech emotion recognition algorithms have achieved impressive performance. Graves et al. [1] proposed deep recurrent neural networks-based speech recognition. An end-to-end deep architecture was proposed with suitable regularization. The method achieved a test error rate of 17.7%, which can be acknowledged as a competitive precision. Abdelhamid et al. [3] proposed a deep-CNN-based algorithm for speech recognition. DNN-HMM has outperformed the conventional GMM-HMM. A limited-weight-sharing algorithm was designed to describe the speech features. Kuo and Gao [8] presented a maximum entropy-based algorithm for speech recognition, where multiple features can be optimally fused. The maximum entropy Markov model was proposed, which outperformed the traditional HMMs. Nicholson et al. [9] utilized one-class-in-one neural networks to achieve the speech emotion recognition. The trained deep model was speaker- and context-independent, which can achieve a 50% recognition rate by evaluating the eight emotional types. Jia et al. [10] proposed the data preprocessing mechanism to extract the acoustic features to achieve speech emotion recognition. The method proposed by Jia incorporated the advantages by proposing a decision tree and the random forest ensemble. Fayek et al. [11] presented a frame-based formulation to speech emotion recognition task, which was achieved by minimizing the speech processing and learning end-to-end deep model. Neumann et al. [12] proposed an attentive CNN deep architecture with multi-view learning objective function. Researchers have conducted an extensive experiments by optimally fusing different kinds of input signals, features and emotion speech. Zhang et al. [13] proposed an upgraded Supervised Locally Linear Embedding (SLLE) strategy to conduct nonlinear dimensionality reduction toward the emotional features.

3. The Proposed Method

In our work, we combine both the facial expression and speech information to achieve emotion recognition. The pipeline of our method is elaborated in Figure 1.

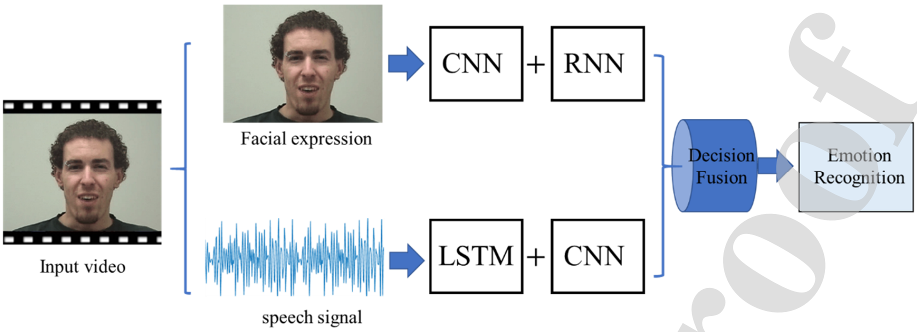

Figure 1. The pipeline of our proposed emotion recognition method.

3.1 Facial expression recognition

Facial expression recognition is significant for speech emotion recognition. Considering the basic expression of human includes anger, disgust, fear, happiness, sadness, surprise, neutrality. In order to achieve better performance during human speech emotion recognition, we leverage the VGG-Face deep model for fine-tuning. For image data, the lower level convolution neural network features are uniform and can be treated as the edge detector or color block detector. As the number of deep layers of neural networks increases, deep features will contain lots of detailed information of the existing data sets. The common deep neural network model for fine-tuning is mainly based on the ImageNet training model. The AlexNet model pre-trained on ImageNet is used to fine-tune the data set of image detection, which improves the effectiveness of image retrieval in terms of precision [14]. Literature [15] also calibrates AlexNet and VGG-16 models that have been pre-trained on the ImageNet to estimate the visual aesthetic quality.

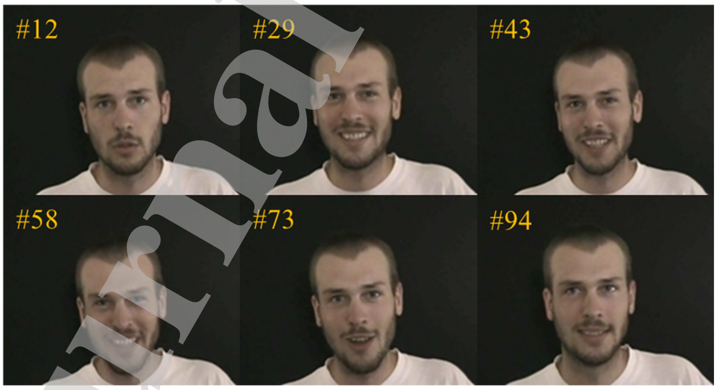

Figure 2. The example of facial expression variations. The yellow integer denotes the frame number in facial videos.

Since facial expression analysis is based on human face images, we leverage the pre-trained model based on VGG-Face [16] instead of ImageNet. VGG-Face consists of eleven blocks, where the first eight blocks are convolution operation and the latter three blocks are fully-connected operation. VGG-Face can effectively improve the accuracy in face recognition. Human expression change is a process, as shown in

Figure 2. When people are happy, facial expressions undergo a change from calm to happy and then to calm, peaking at some point in time. Many existing facial-based facial expression recognition methods are based on artificial marking of facial images. It is often subjective to distinguish the outbreak point of this emotion in a series of facial images with slight changes. Thus, we present a hybrid facial expression recognition method based on CNN and RNN architecture. The hybrid network requires a series of image sequence (such as video stream) as input. The output of each picture in the last full connection layer of convolution neural network is extracted as its corresponding feature representation. Then the facial expression feature is used as the input of the cyclic neural network. The cyclic neural network is trained to learn the distribution of input features and complete classification by using the soft Max layer. The network is shown in Figure 3.

The facial emotion recognition framework consists of two types of neural networks. The former CNN architecture aims to extract deep feature of the input image sequence. In our implementation, we leverage the features output from the fully-connection layer as deep representation. The latter RNN architecture aims to learn the relationship among the input deep representation, since RNN is good at learning time-related sequences. In our implementation, the RNN architecture only contains a hidden layer, and Dropout layer is added into the hidden layer to avoid overfitting.

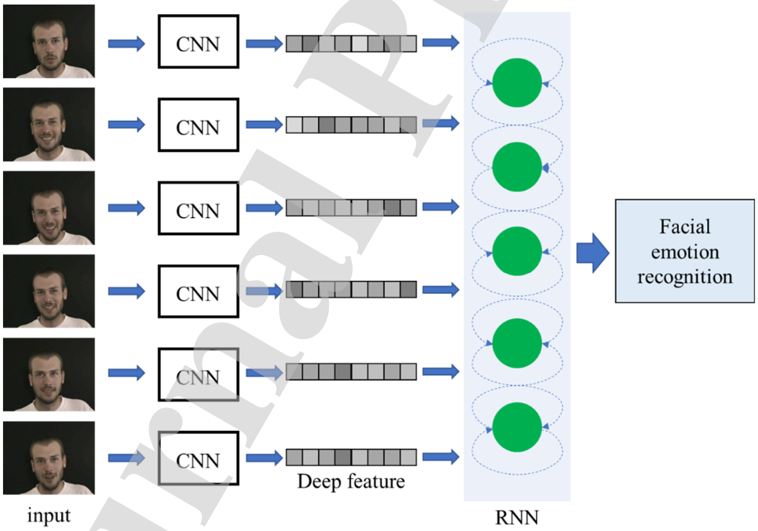

Figure 3. The overall architecture of the hybrid network for human facial emotion recognition.

3.2 Speech emotion recognition

Deep convolution neural network-based architectures have achieved impressive performance in image/video modeling, such as image recognition, classification, and image retrieval. In order to achieve speech emotion recognition, speech signal should be converted into images [17], such as the Mel Frequency Cepstrum Coefficient (MFCC). In our work, we seamlessly combine the LSTM and CNN to achieve speech emotion recognition. More specifically, we first divide the entire signal into multiple equal-length segments since the speech signal is one-dimension and temporal-domain signal, usually 20-40ms in each segment. Then, each speech signal segment is transformed into frequency-domain and processed by Mel filters. Subsequently, we leverage LSTM architecture to learn the temporal correlation of speech sequences. The features extracted by the LSTM is fed into the CNN architecture and we utilize the Softmax function to achieve speech emotion classification/retrieval. The pipeline of speech emotion recognition is shown in Figure 4.

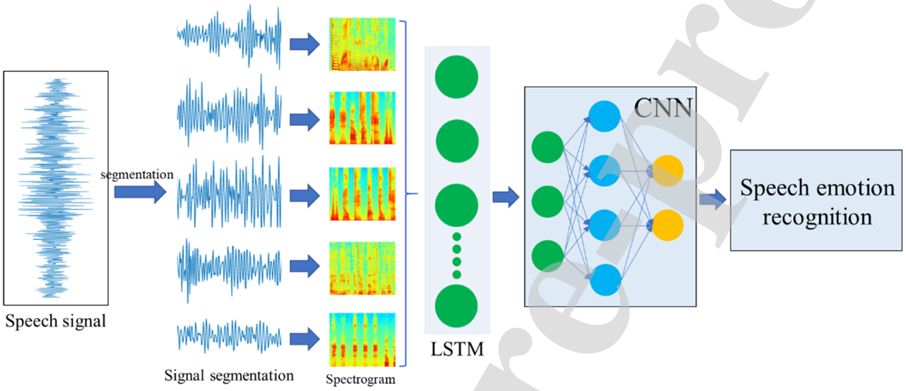

Figure 4. The pipeline of our speech emotion recognition by optimally fusing the RNN and CNN architecture.

We focus on the LSTM architecture. Conventional RNN has some limitations in learning long-time video sequences. In addition, conventional RNN can only leverage the data before the current information. Pahuja et al. [18] proposed the bidirectional recurrent neural networks (BRNN). BRNN contains two hidden layers, which process both forward and reverse sequences. Both output of the two hidden layers is considered as the output layer. However, since the gradient disappearance and explosion of conventional RNN, RNN-based algorithms cannot learn features from the random-length input sequence. Thus, we leverage the LSTM to solve this problem [19]. We define standard recurrent neuron as follows:

$$h_i^{d+1} = f\left(a_i^{d+1}\right) \qquad (1)$$

$$a_i^{d+1} = \sum_j w_{ij} x_j^{d+1} + \sum_k u_{ik} h_k^d \qquad (2)$$

where the function $f(\cdot)$ is the nonlinear activation, and h_i^d denotes the status of the i-th neuron at d-th layer. x denotes the neuron of the previous layers, w and u denote the connection weights. A recurrent neuron cell is elaborated in Figure 5.

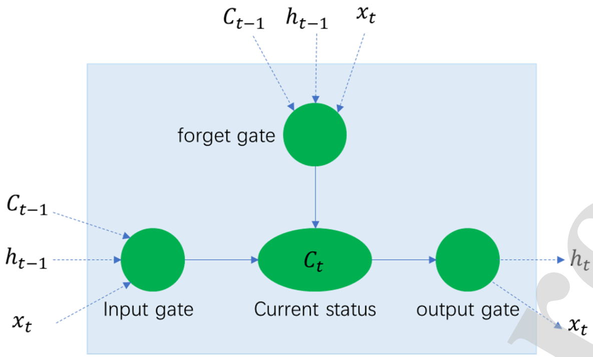

Figure 5. The structure of the recurrent neuron cell.

Considering three types of gates, formula (2) can be reorganized as follows:

$$a_i^{d+1} = c_i^{d+1} a_i^d + b_i^{d+1} \mathrm{g}\left(\sum_j w_{ij} x_j^{d+1} + \sum_k u_{ik} h_k^d \right) \quad (3)$$

where symbol c and b denote the forget/keeping gate and input gaze, respectively. As similar as function $f(\cdot)$, $g(\cdot)$ is a nonlinear activation. The net input consists of three components: signals from the previous layers, hosting cell and previous output. This can be formulated as:

$$\alpha_i^{d+1} = g\left(\sum_j w_{ij}^{\alpha} x_j^{d+1} + \sum_k v_{ik}^{\alpha} h_k^d + v_i^{\alpha} a_i^{d+i_\alpha} \right) \quad (4)$$

where α denotes a gate ranging from $\{\mathrm{b}, \mathrm{c}, \mathrm{d}\}$.

3.3 Bimodal fusion for emotion recognition

In our implementation, we integrate both the facial expression and speech signal for speech emotion recognition. We leverage the decision fusion method to recognize speech emotion. Decision fusion aims to process the category yielded by each model and leverage the specific criteria to re-distinguish. In our implementation, both the human facial expression recognition and speech emotion recognition use the Softmax function for classification. Defining the output of facial expression recognition and speech emotion recognition as $S^{face} = \{S_1^{face}, S_2^{face}, S_3^{face}, \ldots, S_k^{face}\}$ and $S^{speech} = \{S_1^{speech}, S_2^{speech}, S_3^{speech}, \ldots, S_k^{speech}\}$, where k denotes the number of human emotion categories. Then, the weighted decision fusion is calculated as:

$$\mathrm{S} = w_0 S^{face} + w_1 S^{speech} \quad (5)$$

where w_0 and w_1 denotes the two weights, and $w_0 + w_1 = 1$.

4. Experiments and analysis

In this section, we conduct experiments to verify the effectiveness of our proposed method. Our experiment is conducted on a PC equipped with an Intel Broadwell E5 CPU, Nvidia 1080ti GPU and 32GB RAM. Three human emotion datasets are utilized in our experiment including the RML [21], AFEW6.0 [20] and eNTERFACE'05 [22].

4.1 Dataset introduction

RML: RML is a bimodal dataset which contains both the facial expression and speech information. The dataset consists of 720 video samples including voice and facial emotional expression that contains the six basic expressions. The sample sampling rate of the dataset is 44100HZ, and video frame rate is 30pfs.

AFEW6.0: The dataset consists of 773 training samples, 383 validation samples, and 593 testing samples. However, since the testing samples are unavailable, we leverage the validation samples as the testing samples. The video clips from the dataset are from film or television clips. All samples of the dataset are divided into seven types of emotional expressions.

eNTERFACE'05: The dataset is comprised of 42 individuals with different nationalities as the video samples, including 1263 video clips. Among these video clips, 81% were collected from male and 19% were collected form female. The size of each frame is 720×576. This datasets contain six human emotions including anger, disgust, fear, happiness, sadness and surprise.

In our implementation, we leverage the rotation, flipping, color distortion and affine transformation for data augmentation. We use the Caffe deep learning framework to implement our method. The entire dataset is initially trained for 100 epochs with a batch size of 32. The initial learning rate is 0.01, which will be set to 0.005 after 10000 iterations. We set the weight decay and momentum to 0.0002 and 0.9, respectively. The deep emotion recognition model is trained using the Stochastic Gradient Descent (SGD) scheme.

4.2 Comparative study

4.2.1 Facial expression recognition

In our work, the testing result tend to be unchanged after training for 50 epochs. We compare the performance of our facial expression method with several counterparts, as shown in Table I.

Table I. The comparison results in facial expression recognition

	RML	AFEW6.0	eNTERFACE'05
Baseline	0.7263	0.3879	0.5932
Improved AlexNet	0.8551	0.4382	0.7484
VGG-Face	0.8712	0.4593	0.7667
CNN+RNN	0.8896	0.4830	0.7838

The baseline in Table I represents the result of the LBP feature extraction by standard feature extractor and classification by SVM. As we can see, the performance of VGG-face-based fine-tuning scheme is better than those CNN-based approaches. Three facial expression recognition models by leveraging the in-depth learning have achieved satisfactory recognition results. Among these methods, facial emotion recognition based on convolution neural network and cyclic neural network achieved the best performance on the three data sets.

4.2.2 Speech emotion recognition

Because of the small-scale of the traditional speech data, we integrate Gaussian white noise with different weights and signal-to-noise ratios onto the original speech signal. This dataset is initially trained for 100 epochs with a batch size of 32. The initial learning rate is 0.01, and the learning rate is set to 0.005 after 10000 iterations. We set the weight decay and momentum to 0.00001 and 0.9, respectively. The deep model is trained by leveraging the Stochastic Gradient Descent (SGD) algorithm. The number of hidden layer units in the long- and short-term memory network is fixed to 128. The comparison results are shown in Table II.

Table II. The comparison results in speech emotion recognition

	RML	AFEW6.0	eNTERFACE’05
Baseline	0.7600	0.3092	0.4163
CNN	0.8362	0.3506	0.4686
BLSTM	0.8123	0.3436	0.4338
LSTM+CNN	0.8546	0.3790	0.4915

Recognition results of the baseline algorithms are of speech features extracted by the OpenSMILE using SVM. Although the method of deep learning can improve the accuracy of speech emotion recognition, the accuracy increment is small. The quality of speech emotion data set is not high, and it contains more background noise. This will affect the speech recognition performance. Thus, since the speech data set is small, it is easy to have the shortcoming of over-fitting.

4.2.3 Bimodal emotion recognition

In our implementation, we conduct the decision fusion by leveraging the facial expression recognition and speech emotion recognition. Table III elaborates the weights from the three datasets.

Table III. Weight settings on the three datasets

	Facial expression recognition	speech emotion recognition
RML	0.6	0.4
AFEW6.0	0.75	0.25
eNTERFACE’05	0.8	0.2

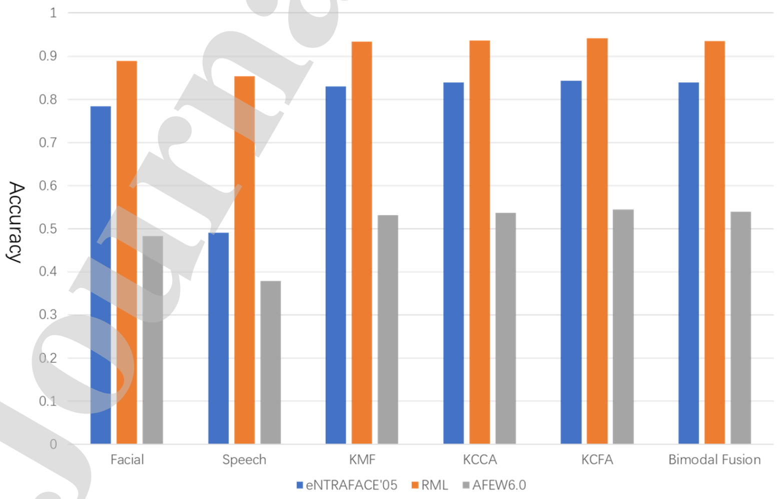

Figure 6. Performance comparison with different feature fusion algorithms.

Figure 6 reports the speech recognition performance compared with different feature fusion methods: KMF, KCCA [23], KCFA [24]. It is observable that the bimodal fusion-based emotion recognition outperforms the signal modal emotion recognition. By fusing the various facial expressions, the overall speech recognition rate can be increased by approximately 5%. Compared with various feature fusion methods in nuclear space, feature fusion by factor analysis of nuclear cross model achieves the best performance. Meanwhile, our recognition model exhibits a certain universality [25-27]. It has achieved good results on the three data sets. In our work, although the weighted decision fusion has some possible limitations compared with feature fusion, the experimental results are better.

5. Conclusions

Emotion recognition is popularly used in modern computer vision systems. In this paper, we propose a bimodal fusion-based speech emotion recognition method, where both the facial expression recognition and speech signal are seamlessly integrated. More specifically, we first combine CNN and RNN to optimally realize facial emotion recognition. Subsequently, we leverage the MFCC to convert speech signal to images. Therefore, we can employ the LSTM and CNN to understand speech emotion. Finally, we propose the weighted decision fusion algorithm to fuse facial expression and speech signal to fulfill speech emotion recognition.

Reference

[1] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics*.

[2] Shen, H., & Zhou, X. (2007). Speech recognition techniques for a sign language recognition system. *Interspeech, Conference of the International Speech Communication Association, Antwerp, Belgium, August*.

[3] Abdelhamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio Speech & Language Processing, 22*(10), 1533-1545.

[4] Ekman, P. . (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium of Motivation, 1972*. University of Nebraska, Press.

[5] Miao, Y., Gowayyed, M., & Metze, F. (2016). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. *Automatic Speech Recognition & Understanding*.

[6] Ververidis, D. , & Kotropoulos, C. . (2006). Emotional speech recognition: resources, features, and methods. *Speech Communication, 48*(9), 1162-1181.

[7] Kamaruddin, N. , & Wahab, A. . (2009). Features extraction for speech emotion. *International Conference on Software Engineering & Data Engineering*. DBLP.

[8] Kuo, H. K. J. , & Gao, Y. . (2006). Maximum entropy direct models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(3), 873-881.

[9] Nicholson, J. , Takahashi, K. , & Nakatsu, R. . (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications, 9*(4), 290-296.

[10] Rong, J. , Li, G. , & Chen, Y. P. P. . (2009). Acoustic feature selection for

automatic emotion recognition from speech. *Information Processing and Management, 45*(3), 315-328.

[11] Fayek, H. M. , Lech, M. , & Cavedon, L. . (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, S089360801730059X.

[12] Neumann, M. , & Vu, N. T. . (2017). Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech.

[13] Zhang, S. Q. , Li, L. M. , & Zhao, Z. J. . (2010). Speech emotion recognition based on an improved supervised manifold learning algorithm. *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology, 32*(11), 2724-2729.

[14] Tzelepi, M. , & Tefas, A. . (2017). Exploiting supervised learning for finetuning deep CNNs in content based image retrieval. *International Conference on Pattern Recognition*. IEEE.

[15] Wang, Y. , Li, Y. , & Porikli, F. . (2017). Finetuning Convolutional Neural Networks for visual aesthetics. *International Conference on Pattern Recognition*. IEEE.

[16] Parkhi O. M., Vedaldi A., Zisserman. A Deep Face Recognition. British Machine Vision Conference. 2015: 41.1-41.12

[17] Abdel-Hamid, O. , Mohamed, A. R. , Jiang, H. , Deng, L. , Penn, G. , & Yu, D. . (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(10), 1533-1545.

[18] Pahuja, V. , Laha, A. , Mirkin, S. , Raykar, V. , Kotlerman, L. , & Lev, G. . (2017). Joint learning of correlated sequence labelling tasks using bidirectional recurrent neural networks.

[19] Hochreiter, S. , & Schmidhuber, Jürgen. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780.

[20] Dhall A., Goecke R., Joshi J. et al. EmotiW2016: video and group-level emotion recognition challenges. ACM International Conference on Multimodal Interaction. ACM, 2016. 417-432.

[21] Wang, Y. , & Guan, L. . (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia, 10*(4), 659-668.

[22] Martin, O. , Kotsia, I. , Macq, B. , & Pitas, I. . (2006). The eNTERFACE'05 Audio-Visual Emotion Database. *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE Computer Society.

[23] HOTELLING, & H. (1936). Relations between two sets of variates. *Biometrika, 28*(3-4), 321-377.

[24] Li, D. , Dimitrova, N. , Li, M. , & Sethi, I. K. . (2003). Multimedia content processing through cross-modal association. *Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, November 2-8, 2003*. ACM.

[25] Zhang, Y. M., Tang, C., (2019) A Medical Image Threshold Segmentation Method Using Improved Swarm Optimization, *Investigación Clínica*, 60(2), 313-321.

[26] Feng, Z. S. Sun, P. Y., (2019) Medical Image Segmentation Based on GA Optimized BP Neural Network, *Investigación Clínica*, 60(1), 233-240.

[27] Wang, Q., Li, Y., & Liu, X. (2018) "The Influence of Photo Elements on EEG Signal Recognition", *Eurasip Journal on Image and Video Processing*, (1), pp. 134.

Xusheng Wang was born in Shanxi, P.R. China, in 1988. He received the PhD. degree from University of Paris-Saclay. Now, He works in Xi'an University of Technology. His research interest include computational intelligence, big data analysis and integrated circuit design.

E-mail: xusheng.w@outlook.com

Xing Chen was born in Shaanxi, P.R. China, in 1999. She received the Bachelor degree from Xi'an University of Technology, P.R. China. Now, she works as a college instructor in Xi'an University of Technology. Her research interests include big data analysis, computational intelligence.

Congjun Cao was born in October 1970. She graduated from Northwestern University with Ph.D. in computer Software and Theory. She is currently a Full professor of Xi'an University of Technology in P.R.China. Her research focuses on cross-media colour reproduction, quality control technology, and computational intelligence.

The authors stated the novelty of the manuscript and there is no conflict between us and the editor.

1�� We first combine the CNN and RNN to achieve facial emotion recognition.

2�� Subsequently, we leverage the MFCC to convert speech signal to images.

3��We utilize the weighted decision fusion method to fuse facial expression and speech signal to achieve speech emotion recognition.

4) Comprehensive experimental results have demonstrated that, compared with the uni-modal emotion recognition, bimodal features-based emotion recognition achieves a better performance.

There is no conflict of interest.