DOI: 10.3969/j. issn. 1673-5188. 2017. S2. 004
http://kns.cnki.net/kcms/detail/34.1294.TN.20180105.1633.002.html, published online January 5, 2017

Multimodal Emotion Recognition with Transfer Learning of Deep Neural Network

HUANG Jian[1,2], LI Ya[1], TAO Jianhua[1,2,3], and YI Jiangyan[1,2]
(1. National Laboratory of Pattern Recognition, Beijing 100190, China;
2. School of Artificial Intelligence, University of Chinese Academy of Science, Beijing 100190, China;
3. CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract

Due to the lack of large-scale emotion databases, it is hard to obtain comparable improvement in multimodal emotion recognition of the deep neural network by deep learning, which has made great progress in other areas. We use transfer learning to improve its performance with pre-trained models on large-scale data. Audio is encoded using deep speech recognition networks with 500 hours' speech and video is encoded using convolutional neural networks with over 110,000 images. The extracted audio and visual features are fed into Long Short-Term Memory to train models respectively. Logistic regression and ensemble method are performed in decision level fusion. The experiment results indicate that 1) audio features extracted from deep speech recognition networks achieve better performance than handcrafted audio features; 2) the visual emotion recognition obtains better performance than audio emotion recognition; 3) the ensemble method gets better performance than logistic regression and prior knowledge from micro-F1 value further improves the performance and robustness, achieving accuracy of 67.00% for "happy", 54.90% for "angry", and 51.69% for "sad".

Keywords

deep neutral network; ensemble method; multimodal emotion recognition; transfer learning

1 Introduction

Emotion recognition has won more and more attention due to its key role in artificial intelligence and human-computer interaction [1]. It is hard to understand people's purpose without emotional intelligence. When interacting with others, we express our emotion in multiple ways, e.g., speech, facial expressions, and body languages. We can identify an emotional state more accurately by observing multimodal information together. Many studies [2], [3] have verified that the complement of different modalities can promote the performance and robustness of emotion recognition.

In the literature of emotion recognition, many handcrafted features are explored. In audio emotion recognition, the spectral, prosody and voice quality features of audio have different contributions [4], and various acoustic low-level descriptors (LLDs) are proposed. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [5] and the INTERSPEECH 2014 Computational Paralinguistics ChallengE (ComParE) [6] are employed widely [7]–[9]. Various handcrafted features, such as LBP from Three Orthogonal Planes (LBP-TOP), Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG), Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP), [10], [11], have played important roles in visual emotion recognition.

In recent years, Deep Neural Networks (DNN), which can extract distinguishable features from raw data, has gained wide acceptance. Han et al. [12] utilize a DNN to construct audio emotional utterance-level features from segment-level probability distributions to boost its performance. Mao et al. [13] use Convolutional Neural Networks (CNN) to introduce a feature learning framework, which disentangles affect-salient features from other noisy factors such as speakers and language in audio emotion recognition. Besides, deep CNN has made great progress in visual emotion recognition. Samira et al. [14] utilize CNN hybrid Recurrent Neural Networks (RNN) to model the spatio-temporal evolution of facial features, which is one of the strongest cues for emotion recognition. Kim et al. [15] construct the hierarchical structure with multiple deep CNNs and achieve over 60% accuracy for static facial emotion recognition.

This work is is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379), the National Key Research & Development Plan of China (No. 2017YFB1002804) and the Major Program for the National Social Science Fund of China (13&ZD189).

However, due to the lack of large-scale emotion databases, it is hard to obtain comparable improvement in multimodal emotion recognition by deep learning which has made great progress in other areas. Transfer learning can relieve this tough problem by applying the previously-acquired knowledge from other similar problems. Transfer learning exploits the knowledge learned from data in auxiliary domains to facilitate predictive modeling in the current domain [16]. The strategy for transfer learning of deep neutral network is that the trained parameters of the source task model act as the initial parameters of the new target task model. If the source task model has been trained with adequate data, only the last layer needs to be retrained for the target task model [17]. Besides, any layer of the target task model can be fine-tuned to be adaptive if necessary [18].

To exert the power of DNN in multimodal emotion recognition, we extract high level audio-visual features with transfer learning in this paper. As mentioned above, DNN has been applied for audio emotion recognition [12], [13]. However, it cannot achieve great performance of DNN due to the small scale of emotion databases. Transfer learning is also applied for audio emotion recognition [17], which adopts speaker recognition networks as the source task. We utilize another strategy that the trained network of the source task acts as a feature extractor to extract more intrinsic and robust features for the target task. The deep speech recognition networks trained with large audio corpus are used to extract high level audio features for audio emotion recognition model. The same strategy is applied for visual emotion recognition.

As to the multimodal fusion, feature level fusion and decision level fusion strategies are widely utilized to combine all the modalities together [19]. Feature level fusion is a straight strategy, which extracts audio and visual features separately and then concatenates them into feature vector for final emotion recognition [3], [14]. Nevertheless, the alignment of audio and video with different sampling rate results in difficulty and inconsistency in multimodal fusion. Moreover, the failure of feature extraction from any modality will bring about noise in the multimodal feature vector and finally result in the low accuracy of emotion recognition. To overcome this problem, Yu and Zhang [20] ensemble multiple stat-of-the-art face detectors to improve the face detection accuracy. On the contrary, decision level fusion assumes each modality is independent and builds separate emotion recognition models. In the end, the emotion recognition results of different modalities are fused together. Kahou et al. [14] use a weighted sum of the class probabilities estimated by the modality-specific classifiers to get resulting score for each class. Kim et al. [15] adopt exponentially-weighted decision fusion strategy to optimize the weight of different modalities.

In this paper, we adopt two strategies of decision level fusion, logistic regression (LR) and ensemble method. They both act on the posterior probability value from all the modalities. LR regards them as intermediate features to train recognition models, whereas the ensemble method [17] is to choose the better ones as final results for each test sample. We also combine these two strategies to raise the accuracy of multimodal emotion recognition. Then, we propose the modified ensemble method with prior knowledge, which can further improve its performance and robustness.

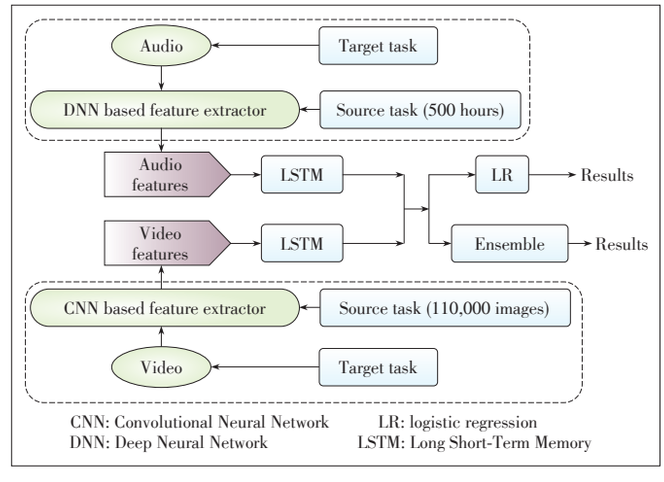
In the following, Section 2 briefly introduces the proposed method. Section 3 presents the database and feature set adopted in the experiments. Section 4 describes the details of the entire experiment and results. Section 5 concludes this paper.

2 Proposed Method

In this paper, we train deep speech recognition networks with large audio corpus and CNN model with large image databases. The trained models act as feature extractors with transfer learning. That is, audio information is encoded with the DNN based feature extractor and video information is encoded with the CNN based feature extractor. The extracted audio and visual features are then fed into Long Short-Term Memory (LSTM) to train emotion recognition models separately. Finally, LR and the ensemble method are performed to fuse audio and video emotion recognition results in the decision level fusion. **Fig. 1** gives the overview of our method. The top and the bottom dotted boxes in Fig. 1 represent the DNN feature extractor and the CNN feature extractor respectively.

2.1 LSTM Network Structure

LSTM is an extension of RNN, which avoids the gradient disappearance and gradient explosion problem. In a standard RNN, given the input sequence $m=(m_1,...,m_T)$ and the hidden vectors $h=(h_1,...,h_T)$, the output vectors $y=(y_1,...y_T)$ are com-



▲ **Figure 1. Framework of the proposed method. The audio features from the DNN feature extractor and video features from the CNN feature extractor are fed into LSTM, and then logistic regression and the ensemble method are utilized in decision level fusion.**

puted as:

$$h_t = f_{act}\left(W_{mh} m_t + W_{hh} h_{t-1} + b_h\right), \tag{1}$$

$$y_t = f_{out}(W_{hy} h_t + b_y), \tag{2}$$

where f_{act} is the activation of the hidden layer, and f_{out} is the activation of the output layer. The weight matrix W_{zz} and the bias vectors b_z are identified by the subscript $z \in \{m,h,y\}$. For example, W_{mh} is the input-hidden matrix.

LSTM replaces f_{act} with an elaborately designed memory block including three multiplicative units (the input, the output, and forget gates), which are better at storing and accessing information. What's more, LSTM can learn long-term dynamic information well. Exactly, emotion is a temporally expression event which can be better inferred by LSTM network structure. Therefore, LSTM is used for training emotion recognition models.

2.2 Decision Level Fusion

We utilize two different decision level fusion strategies to enhance the performance of multimodal emotion recognition. Firstly, LR is performed based on the combination of posterior probability values from different modalities. Secondly, we consider the ensemble method with confidence, which chooses a better one as the final result among different predicted results based on the difference of posterior probability. When fusing predicted results from different networks for each test sample, Huang et al. [17] point out that the difference between the top two posterior probabilities could be an indicator of confidence of the category with the top entry value. In other words, larger difference of posterior probability makes predicted results more reliable.

In this paper, we adopt the ensemble method with confidence to fuse audio and visual emotion recognition results. Suppose $y^v_{1,m}$, $y^v_{2,n}$ denote the top two video posterior probabilities and $y^a_{1,i}$, $y^a_{2,j}$ denote the top two audio posterior probabilities, where m, n, i, j, are labels. The output of ensemble follows this strategy [17]:

$$l = \begin{cases} m, & (y^v_{1,m} - y^v_{2,n}) \geqslant (y^a_{1,i} - y^a_{2,j}) \\ i, & \text{otherwise} \end{cases}. \tag{3}$$

Actually, one modality may have better performance than another. For instance, visual emotion recognition performs better than audio emotion recognition [10], [11]. However, wrong predicted results may interfere the correct predicted results during multimodal decision level fusion using the ensemble method. Therefore, we propose a modified ensemble method with prior knowledge, and the idea behind this method is to favor the one with better performance to maintain its great performance when fusing with other modalities. In particular, we calculate the micro-F1 value [21] in (4).

$$M_{micro} = M\left(\sum_{\lambda=1}^{m} tp_\lambda, \sum_{\lambda=1}^{m} fp_\lambda, \sum_{\lambda=1}^{m} tn_\lambda, \sum_{\lambda=1}^{m} fn_\lambda\right), \tag{4}$$

where λ denotes one sample, m denotes the number of samples, M denotes the calculation of F1-score and tp_λ, fp_λ, tn_λ, fn_λ denote true positives, false positives, true negatives, and false negatives of sample λ. The difference of the micro-F1 values of two different predicted results is regarded as prior knowledge to lay stress on the one having higher micro-F1 value. Suppose the network of $y^v_{1,m}$ has higher micro-F1 value, the proposed ensemble method with prior knowledge from micro-F1 value follows this strategy:

$$l = \begin{cases} m, & (y^v_{1,m} - y^v_{2,n}) + d \geqslant (y^a_{1,i} - y^a_{2,j}), \\ i, & \text{otherwise} \end{cases} \tag{5}$$

where d is the positive difference of micro-F1 values. Equation (5) indicates that we still choose the predicted results having smaller difference of the posterior probabilities within the range of difference of the micro-F1 values.

3 Database and Feature Set

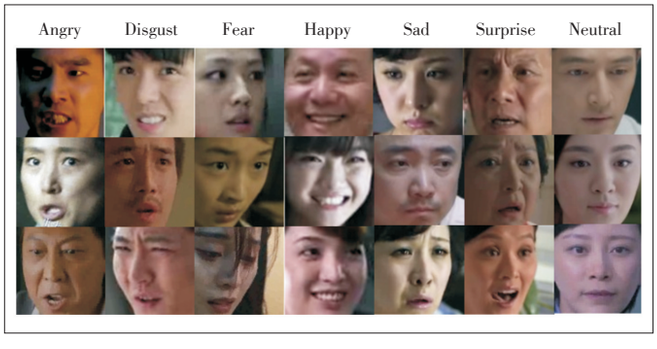
In this paper, we utilize Chinese Natural Audio-Visual Emotion Database (CHEAVD) [22], which is collected from movies, talk shows and TV shows and is composed of audio-video short clips. Seven emotional categories, namely angry, disgust, fear, happy, sad, surprise, and neutral, are chosen and their distributions are shown in **Table 1**. Some examples of different types of facial expression images are shown in **Fig. 2**. The database is divided into three sets: training, validation, and test, approximating the proportion of 7:1:2.

3.1 Audio Features

To evaluate the performance of audio features extracted

▼**Table 1. Number of samples for seven emotional categories**

Category	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Number	462	310	56	440	460	190	300



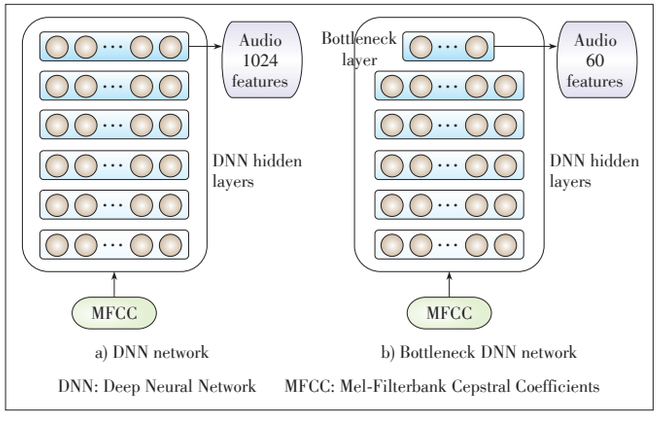
▲**Figure 2. Different types of facial expression images.**

from deep speech recognition networks, we also extract handcrafted audio features for audio emotion recognition. Specifically, we add the first dimension of the Mel-Filterbank Cepstral Coefficients (MFCC 0), the first order and second derivatives of all the low-level descriptors (LLDs) to the ComParE [6]. The resulting features 147 LLDs [9] for every frame are extracted by openSMILE.

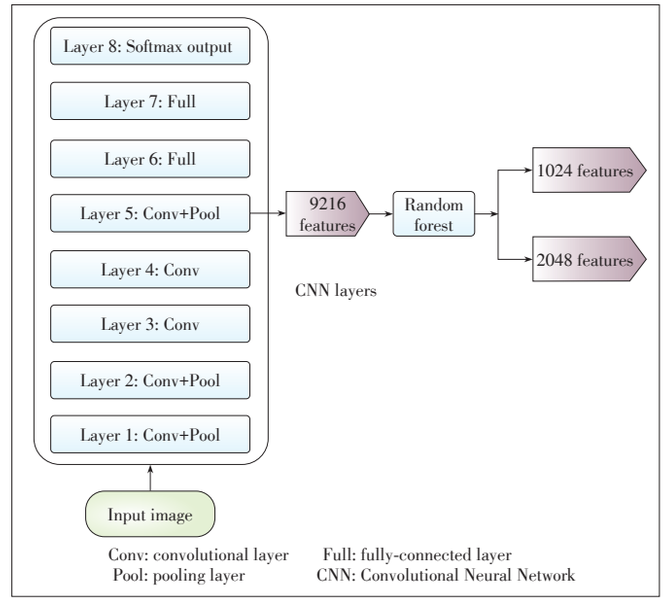
On the other hand, we extract high level audio features from pre-trained deep speech recognition networks. Bottleneck features were designed for automatic speech recognition [23] and since then have been integrated into many top-performing ASR systems. We extract emotion bottleneck features from a DNN network (**Fig. 3a**) and a bottleneck DNN network (Fig. 3b) for comparison. The DNN network has six hidden layers and each layer has 1024 nodes. The last layer of bottleneck DNN network has 60 nodes. These two networks are trained with 500 hours spontaneous and accented Mandarin speech corpus using 13-dimensional MFCC features plus their 1st and 2nd order derivatives. The trained deep speech recognition networks act as DNN based feature extractors with transfer learning. We can obtain high level audio features from the output of emotion audio given by the last hidden layer.

3.2 Visual Feature Extraction

We utilize pre-trained CNN network to extract high level visual features in visual emotion recognition. We train an effective deep CNN with Celebrity Faces in the Wild (CFW) [24] and Facescurb [25] dataset, which have over 110,000 face images from 1032 people for training in all. The CNN architecture contains three fully connected layers and five convolutional layers as shown in **Fig. 4**. The trained CNN model acts as CNN based feature extractor with transfer learning. We utilize the tracking algorithm and toolkit [26] to detect human face from video clips. The detected face images are scaled into 100×100 pixels and fed into trained CNN network to extract the 9216-dimensional features from the 5th pooling layer. Finally, the random forest is employed for feature selection and two different dimensional features, namely 1024 and 2048, are kept for comparison. Fig. 3 shows the diagram of visual features extraction.



▲Figure 3. Diagram of audio features extraction.



▲Figure 4. Diagram of visual features extraction.

4 Experiments

4.1 Audio and Visual Emotion Recognition

After obtaining audio and visual features, LSTM is utilized to train audio and visual emotion recognition models separately. The audio network includes one hidden layer, one LSTM layer and one softmax regression layer. The hidden layer converts audio features into proper dimensional features for the LSTM layer followed by the softmax layer. The hyper-parameters are chosen by the prediction accuracy in the validation set. The hidden layer and LSTM layer both have 128 nodes and the weight decay in the softmax regression layer is 0.0001. We use dropout after LSTM with the rate 0.5. The visual network architecture is the same as the audio network architecture except different dimensional input. The experiment results are shown in **Table 2**, where a60 and a1024 denote 60 bottleneck acoustic features and 1024 acoustic features from DNN, while v1024 and v2048 denote 1024 visual features and 2048 visual features from CNN.

▼Table 2. Recognition accuracy for audio and video modalities

Features	Accuracy
Handcrafted audio features (147)	0.31
a60	0.40
a1024	0.36
v1024	0.45
v2048	0.46

Table 2 indicates the audio features extracted from DNN achieve better performance than handcrafted audio features significantly. Especially, 60 bottleneck acoustic features extracted from the bottleneck DNN network achieve superior performance than 147 handcrafted audio features, which indicates the DNN extracted audio features have more suitable characteristics to be trained with LSTM than handcrafted audio features. Therefore, we only adopt the DNN audio features in the multimodal decision level fusion. Also, Table 2 indicates visual emotion recognition obtains better performance than audio emotion recognition, which is consistent with Wu's work [10] and Sun's work [11].

4.2 Multimodal Decision Level Fusion

To obtain better emotion recognition from both of the audio and video modalities, we utilize two decision level fusion strategies. Firstly, LR is utilized to fuse the posterior probability results of audio and visual emotion recognition. As a result, we obtain predicted results and preserve corresponding posterior probability results from LR which will be used in the following steps. Secondly, the ensemble method explained by (3) is performed. The experiment results are shown in **Table 3**, where LR and ensemble denote logistic regression and the ensemble method respectively.

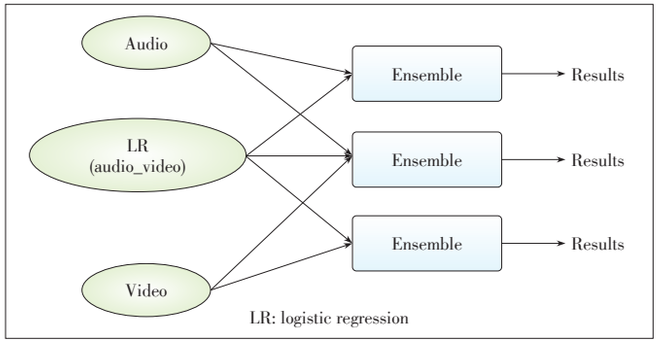
Compared with Table 2, Table 3 indicates the multimodal decision level fusion can improve the performance obviously. Moreover, the ensemble method gets better performance than LR. In the following experiments, we only utilize the ensemble method. In addition, the results of LR act as the new model results for decision level fusion. **Fig. 5** shows that the results of LR are fused with corresponding audio and video results using the ensemble method, which combines logistic regression and the ensemble method. The experiment results based on Fig. 5 are shown in **Table 4**, where LR (a60_v1024) denotes the results of logistic regression from a60 and v1024, and so forth.

From Table 4, we can observe that the performance has been improved compared with Table 3 as a whole, especially when fusing the results of logistic regression with corresponding audio results. This strategy takes advantage of logistic regression and the ensemble method together, which can promote the performance of multimodal emotion recognition. We also fuse the results of logistic regression with corresponding audio and video results together. However, it is observed from Table 4 that the performance of third column has not been improved compared with the first column.

▼**Table 3. Recognition accuracies for multimodal decision level fusion (LR and the ensemble method)**

Accuracy	v1024		v2048	
	LR	Ensemble	LR	Ensemble
a60	0.48	0.49	0.49	0.49
a1024	0.47	0.47	0.47	0.48

LR: logistic regression



▲**Figure 5. Fusion between the results of LR with corresponding audio and video results using the ensemble method.**

4.3 The Ensemble Method with Prior Knowledge

Although the modality with good performance has correct predicted results for some test samples, wrong predicted results of other modalities may be chosen as final results due to its larger difference of posterior probability in multimodal fusion using the ensemble method. Therefore, we modify the ensemble method in such way that prior knowledge from the micro-F1 value is introduced to favor the modality with better performance to maintain its great performance when fusing with other modalities. We conducted experiments described by (5). The value of d is obtained from the validation set and applied in the test set. The experiment results on the basis of Table 4 are shown in **Table 5**, the value in brackets is d.

Compared with Table 4, Table 5 shows better and more robust performance for different feature combinations and obtains the best accuracy of 0.52, which verifies the effectiveness of proposed ensemble method. The confusion matrix of the above best result is shown in **Fig. 6**. The confusion matrix indi-

▼**Table 4. Recognition accuracy of fusion between the results of LR with corresponding audio and video results using the ensemble method**

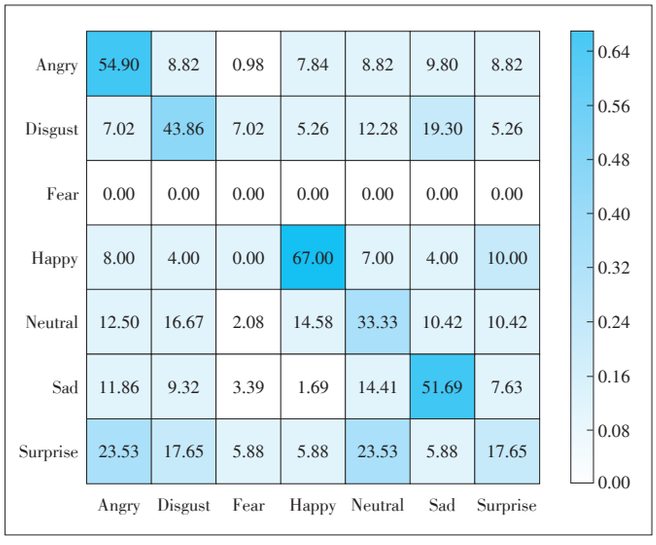
Accuracy	Audio	Video	Audio+Video
LR (a60_v1024)	0.51	0.47	0.51
LR (a60_v2048)	0.51	0.46	0.49
LR (a1024_v1024)	0.47	0.48	0.48
LR (a1024_v2048)	0.49	0.45	0.46

LR: logistic regression

▼**Table 5. Recognition accuracy using the ensemble method with prior knowledge**

Accuracy	Audio	Video
LR (a60_v1024)	0.52 (0.086)	0.48 (0.032)
LR (a60_v2048)	0.51 (0.095)	0.49 (0.032)
LR (a_1024_v1024)	0.48 (0.113)	0.48 (0.023)
LR (a1024_v2048)	0.51 (0.111)	0.47 (0.011)

LR: logistic regression



▲**Figure 6. Confusion matrix of the best recognition accuracy using the ensemble method with prior knowledge (%).**

cates that our method performs well for "happy", "angry", "sad" and "disgust", but poorly for "fear" and "surprise" because of their relatively small number of examples resulting in inadequate training. Since "neutral" is the neutral state of a person and could be easily confused with other emotions, e.g. happy. The performance of "neutral" is also not good since it has an overlap with most of other categories.

The prior knowledge from the micro-F1 value could further improve the performance and robustness of multimodal emotion recognition. Actually, the micro-F1 value is calculated with the predicted results of different modalities. Therefore, the proposed method is based on data-driven without additional expert knowledge and can be applied to other multimodal fusion fields easily.

5 Conclusions

To improve the performance of multimodal emotion recognition with small amount of training data, we use transfer learning for the pre-trained deep neural network models on large-scale data for both audio and visual feature extraction. Compared with handcrafted audio features, the audio features extracted from deep speech recognition networks can improve the performance of audio emotion recognition significantly. Besides, the experiment results indicate visual emotion recognition obtains higher performance than audio emotion recognition. Finally, we utilize logistic regression and the ensemble method in decision level fusion, and then combine their advantages to improve the performance of multimodal emotion recognition. We consider the confidence of the emotion recognition results from each modality, and also introduce the prior knowledge obtained from F1-score to favor the modality with better performance in particular. This proposed method further improves the performance and robustness than the original ensemble method. The data-driven characteristic of this method makes it easier to be applied to other multimodal fusion fields without any additional expert knowledge. In the future, we will introduce text contexts into multimodal emotion recognition, and also utilize transfer learning to improve the performance of cross-language emotion recognition.

References

[1] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective Computing and Intelligent Interaction*, Beijing, China, Oct. 2005, pp. 981–995. doi: 10.1007/11573548_125.
[2] J. A. Russell, J. A. Bachorowski, and J. M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, Nov. 2003, pp. 329–349. doi: 10.1146/annurev.psych.54.101601.145102.
[3] C. Busso, Z. Deng, S. Yildirim, et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th International Conference on Multimodal Interfaces*, State College, USA, Oct. 2004, pp. 205–211. doi: 10.1145/1027933.1027968.
[4] Y. Li, L. Chao, Y. Liu, et al., "From simulated speech to natural speech, what are the robust features for emotion recognition?" in *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, Sept. 2015, pp. 368–373. doi: 10.1109/ACII.2015.7344597.
[5] F. Eyben, K. R. Scherer, B. W. Schuller, et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, Jan. 2015, pp. 190–202. doi: 10.1109/TAFFC.2015.2457417.
[6] B. Schuller, S. Steidl, A. Batliner, et al., "The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load," in *15th Annual Conference of the International Speech Communication Association*, Singapore, Singapore, Sept. 2014, pp. 427–431.
[7] F. Ringeval, B. Schuller, M. Valstar, et al., "Av+ ec 2015: the first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, Brisbane, Australia, Jan. 2015, pp. 3–8. doi: 10.1145/2808196.2811642.
[8] M. Valstar, J. Gratch, B. Schuller, et al., "AVEC 2016-depression, mood, and emotion recognition workshop and challenge," in *6th International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, Oct. 2016, pp. 1483–1484. doi: 10.1145/2988257.2988258.
[9] X. Xia, L. Guo, D. Jiang, et al., "Audio visual recognition of spontaneous emotions in-the-wild," in *Chinese Conference on Pattern Recognition*, Chengdu, China, Nov. 2016, pp. 692–706. doi: 10.1007/978-981-10-3005-5_57.
[10] J. Wu, Z. Lin, and H. Zha, "Multiple models fusion for emotion recognition in the wild," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 475–481. doi: 10.1145/2818346.2830582.
[11] B. Sun, L. Li, G. Zhou, et al, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 497–502. doi: 10.1145/2818346.2830586.
[12] K. Han, D. Yu, I. Tashev, et al., "Speech emotion recognition using deep neural network and extreme learning machine," in *14th Annual Conference of the International Speech Communication Association*, Malaysia, Singapore, Sept. 2014, pp. 223–227.
[13] Q. Mao, M. Dong, Z. Huang, et al., "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014. doi: 10.1109/TMM.2014.2360798.
[14] K. S. Ebrahimi, V. Michalski, K. Konda, et al., "Recurrent neural networks for emotion recognition in video," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 467–474. doi: 10.1145/2818346.2830596.
[15] B. K. Kim, H. Lee, J. Roh, et al., "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 427–434. doi: 10.1145/2818346.2830590.
[16] J. Lu, V. Behbood, P. Hao, et al., "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, May 2015, pp. 14–23. doi: 10.1016/j.knosys.2015.01.010.
[17] Y. Huang, M. Hu, X. Yu, et al., "Transfer Learning of Deep Neural Network for Speech Emotion Recognition," in *Chinese Conference on Pattern Recognition*,

Chengdu, China, Nov. 2016, pp. 721–729. doi: 10.1007/978-981-10-3005-5_59.
[18] H. W. Ng, V. D. Nguyen, V. Vonikakis, et al., "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 443–449. doi: 10.1145/2818346.2830593.
[19] H. Gunes, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, Jan.-Jun. 2010. doi: 10.4018/jse.2010101605.
[20] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. 2015 ACM on International Conference on Multimodal Interaction*, Seattle, USA, Nov. 2015, pp. 435–442. doi: 10.1145/2818346.2830595.
[21] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, Mar. 2002. doi: 10.1145/505282.505283.
[22] Y. Li, J. Tao, L. Chao, et al., "CHEAVD: a Chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 1–12, Nov. 2017. doi: 10.1007/s12652-016-0406-z.
[23] F. Grézl, E. Egorova, and M. Karafiát, "Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure," in Spoken Language Technology Workshop (SLT), South Lake Tahoe, USA, Dec. 2014, pp. 48–53. doi: 10.1109/SLT.2014.7078548.
[24] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, vol. 14, no. 4, Aug. 2012. doi:10.1109/TMM.2012.2186121.
[25] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Paris, France, Oct. 2014. doi:10.1109/ICIP.2014.7025068.
[26] X. Xiong and F. D. Torre, "Supervised descent method and its applications to face alignment," in Proc. IEEE Conference on Computer Vision And Pattern Recognition, Portland, OR, USA, Jun. 2013, pp. 532–539. doi: 10.1109/CVPR.2013.75.

Manuscript received: 2017-08-15

Biographies

HUANG Jian (jian.huang@nlpr.ia.ac.cn) is a Ph.D. candidate at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. His research interest covers affective computing, deep learning, and multimodal emotion recognition.

LI Ya (yli@nlpr.ia.ac.cn) received the B.E. degree from University of Science and Technology of China (USTC), China in 2007, and Ph.D. degree from the NLPR, CASIA in 2012. From November 2012 to December 2012, she was a visiting scholar in the University of Tokyo, Japan. From May 2014 to September 2014, she was a research fellow with Trinity College Dublin, Ireland. She is currently an associate professor with the NLPR, CASIA. She won several best student papers in INTERSPEECH, NCMMSC, etc. Her general interests include speech recognition and synthesis, affective computing, human computer interaction, and natural language processing.

TAO Jianhua (jhtao@nlpr.ia.ac.cn) received his Ph.D. from Tsinghua University, China in 2001 and MS. from Nanjing University, China in 1996. He is currently a professor with the NLPR, CASIA. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing, and pattern recognition. He has published more than eighty papers on major journals and proceedings including *IEEE Transactions on ASLP*, and got several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He serves as the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC, etc. He also serves as the steering committee member for *IEEE Transactions on Affective Computing*, an associate editor for *Journal on Multimodal User Interface and International Journal on Synthetic Emotions*, and Deputy Editor-in-chief for *Chinese Journal of Phonetics*.

YI Jiangyan (jiangyan.yi@nlpr.ia.ac.cn) is a Ph.D. candidate at the NLPR, CASIA, China. Her research interest covers deep learning, speech recognition, and transfer learning.