



Multi-modal Emotion Recognition Based on Speech and Image

Yongqiang Li^{1,2}(✉), Qi He¹, Yongping Zhao¹, and Hongxun Yao²

¹ School of Electrical Engineering and Automation,
Harbin Institute of Technology, Harbin 150001, China
liyongqiang@hit.edu.cn

² School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China

Abstract. For the past two decades emotion recognition has gained great attention because of huge potential in many applications. Most works in this field try to recognize emotion from single modal such as image or speech. Recently, there are some studies investigating emotion recognition from multi-modal, i.e., speech and image. The information fusion strategy is a key point for multi-modal emotion recognition, which can be grouped into two main categories: feature level fusion and decision level fusion. This paper explores the emotion recognition from multi-modal, i.e., speech and image. We make a systemic and detailed comparison among several feature level fusion methods and decision level fusion methods such as PCA based feature fusion, LDA based feature fusion, product rule based decision fusion, mean rule based decision fusion and so on. We test all the compared methods on the Surrey Audio-Visual Expressed Emotion (SAVEE) Database. The experimental results demonstrate that emotion recognition based on fusion of speech and image achieved high recognition accuracy than emotion recognition from single modal, and also the decision level fusion methods show superior to feature level fusion methods in this work.

Keywords: Multi-modal emotion recognition · Feature level fusion
Decision level fusion

1 Introduction

Emotion is an intrinsic subjective experience, but always accompanied by certain external manifestations, such as changes in facial expressions, speech tones, rhythms, etc. Most traditional studies on emotion recognition focused on recognizing the current emotional state of human beings through single modal information, such as audio recognition. Cahn et al. [1] designed an emotional speech editor and analyzed the relationship between acoustic characteristics such as fundamental frequency, amplitude,

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-77380-3_81) contains supplementary material, which is available to authorized users.

formant and other emotional states; Tato et al. [2] found that sound quality not only expresses “Valence” information in the three-dimensional emotional space model, but also partly reflects its “control” information; Liu et al. [3] proposed a nonlinear method based on covariance descriptor and Riemannian manifold for audio emotion recognition. Some works [4–7] used speech spectral parameters LPC/LPCC/MFCC to study the short-term characteristics of speech signals. Besides, many scholars try to recognize facial expressions from images, such as Zeng et al. [8] used multi-scale and multi-directional Gabor wavelet to extract facial features for facial expression recognition; He et al. [9] used optical flow method with a priori knowledge of human face for facial expression recognition; Bouchra et al. [10] used the AAM model for the recognition and synthesis of facial expressions. Although emotion recognition based on single modal has made significantly progress, recognizing emotions from both image and speech will theoretically improve the recognition performance because of involving more information. Hence, more and more scholars turn to the field of multi-modal emotion recognition [11–16]. Some effects try to fuse the visual information obtained by the face image with the audio information obtained from the speech, to achieve a better recognition performance. In this paper, we explore the emotional recognition from multi-modal, i.e., speech and image. The target emotions include anger, disgust, fear, happiness, sadness, surprise and neutral. We extract 92-dimensional audio emotional features and 240-dimensional visual features, and make a systemic and detailed comparison among several feature level fusion methods and decision level feature fusion methods such as PCA based feature fusion, LDA based feature fusion, product rule based feature fusion, mean rule based feature and so on. All the compared methods are tested on the Surrey Audio-Visual Expressed Emotion (SAVEE) Database. A detailed comparison and analysis of the experimental results are provided.

2 Method

2.1 Database

The SAVEE emotional database, whose text material was selected from standard TIMIT database, was recorded from four native English male speakers, including postgraduate students and researchers at the University of Surrey aged from 27 to 31. There are seven emotional categories, namely anger, disgust, fear, happiness, neutral, sadness and surprise. Each emotion consists of 15 sentences: 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. There are also 30 sentences of neutral emotion. That is to say each recorder should speak 120 utterances, so there are 480 materials in total. The sampling rate was 44.1 kHz for audio and 60 fps for video.

To facilitate extraction of facial features, the actor’s face was painted with 60 markers. After data capture, markers were manually labeled for the first frame of a sequence and then tracked for the remaining frames using a marker tracker. This stage created a set of 2D marker coordinates for each frame of the visual data.

Each actor’s data were evaluated by 10 subjects, of which 5 were native English speakers and the rest of them had lived in UK for more than a year. Using the authoritative approach to evaluation, the final recognition results were audio 66.5%, visual 88.0%, Audio-visual 91.8%. The results showed that the database used for emotional recognition was effective.

2.2 Multi-modal Fusion Emotion Recognition

In this paper, SAVEE Database was selected to evaluate all the compared methods. We firstly extract the audio emotional features and visual emotional features from the data instances collected from SAVEE. All the compared methods are based on the same raw features. We test emotion recognition based on single modal, as well as based on multi-modal. The process of experiment is shown in Fig. 1.

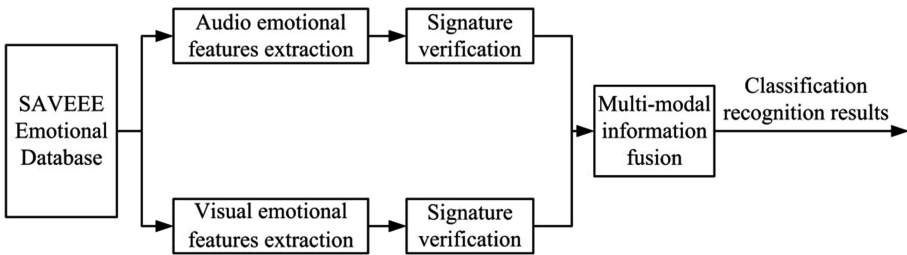


Fig. 1. Process of experiment

Specifically, 92-dimensional phonetic emotional features are extracted, including statistical parameters of short-term energy, duration of speech, pitch frequency, formants, and Mel Frequency Cepstral Coefficients (MFCCs). For visual features, we extract 240-dimensional features of the face image sequence by calculating the mean and standard deviation of the coordinates of 60 facial landmarks. The visual feature extraction flow is shown in Fig. 2. Then we use the “one-versus-one” strategy support vector machine (SVM) to perform emotional recognition experiments on the audio and visual features respectively, and we employ SVM as the classifier which is implemented using libsvm toolbox [17]. Finally, several the feature level fusion methods and the decision level fusion methods are implemented to fuse the audio modal information and the visual modal information, and then detailed comparison and analysis are provided.

PCA and LDA methods are used for feature level fusion and five rules are adopted for decision level fusion, including product rule, mean rule, summation rule, maximum rule and minimum rule. The general formula of decision level fusion is as follows:

$$\begin{cases} mid_p_j(n) = rule(origin_p_{ij}(n)) \\ new_p_j(n) = \frac{mid_p_j(n)}{\sum_j mid_p_j(n)} \end{cases}, i = 1, 2, \dots, k, j = 1, 2, \dots, c \quad (1)$$

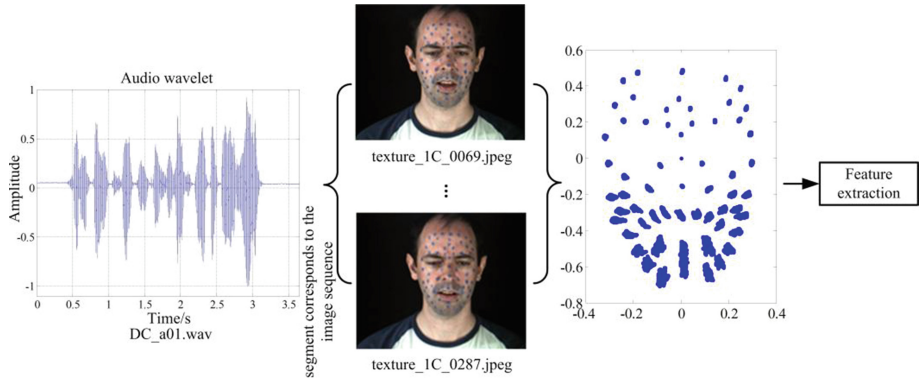


Fig. 2. Visual emotional feature extraction process

where $mid_p_j(n)$ is the recognition probability matrix after the regular operation, $rule$ denotes five rules as listed above, $origin_p_{ij}(n)$ is audio emotional recognition probability matrix and visual emotional recognition probability matrix, and $new_p_j(n)$ is the output recognition probability matrix, n is the current test sample, k denotes the number of emotional classifiers, c denotes the emotional category, and $k = 2$ and $c = 7$ in this paper. For data instance n , the final label is assigned to j where $new_p_j(n)$ has its maximum value.

3 Experiments and Analysis

We tested all the compared methods on SAVEE database, from which we collect 480 data instances and randomly select 320 as training data and test the methods on the rest. The results of the experiments are analyzed below. Here we only gives out a brief experimental result, and for the whole confusion matrix, the readers are referred to the supplemental material.

3.1 Audio Emotional Recognition

Table 1 shows the recognition results based on audio data.

Table 1. Audio emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	80.00%	70.00%	45.00%	35.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	80.00%	70.00%	65.00%	65.63%

The average recognition accuracy is significantly higher than that in the work [18]. From Table 1 we can see that, recognition rates of anger and neutral have the highest

accuracy. Anger is easily misidentified as disgust and neutral is easily misidentified as disgust and sadness. Disgust, sadness and surprise are all easily misidentified as sadness. Fear and happiness have the lowest recognition rate, where fear is easily misidentified as disgust, neutral or surprise, and happiness is easily misidentified as anger or fear.

3.2 Visual Emotional Recognition

The classification results based on visual data are shown in Table 2.

Table 2. Visual emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	70.00%	55.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	80.00%	87.50%

The average recognition accuracy based on visual data is higher than the speech emotional that of based on audio data. The recognition rates of anger, neutral and sadness are all 100%. Surprise is easily misidentified as fear, and disgust is easily misidentified as anger. Recognition rate of fear is higher than that on audio data, but is still the lowest, which is easily misidentified as surprise and sadness.

3.3 Feature Level Fusion Emotional Recognition

3.3.1 PCA Based Fusion

We feed all the 332 dimensional audio and visual features to PCA and 93 dimensional features were obtained. The classification results based on PCA based fusion are shown in Table 3.

Table 3. PCA fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	90.00%	85.00%	50.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	90.00%	85.00%	85.00%	83.75%

The average recognition accuracy is higher than that of audio emotional recognition but lower than that of visual emotional recognition, which is consistent with the conclusion of the paper [18]. Happiness, anger and neutral recognition accuracies are the highest. Disgust is easily misidentified as sadness, sadness is easily misidentified as fear, and surprise is easily misidentified as fear. Fear recognition accuracy is the lowest, which is easily misidentified as surprise and sadness.

3.3.2 LDA Based Fusion

We feed all the 332 dimensional audio and visual features to LDA and 6 dimensional features were obtained. The recognition results based on LDA based fusion are shown in Table 4.

Table 4. LDA fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	95.00%	90.00%	40.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	97.50%	100.00%	100.00%	89.38%

The average recognition accuracy is higher than that of audio emotional recognition and visual emotion recognition. The recognition accuracy of sadness and surprise recognition are the highest while that of fear is the lowest, which is easily misidentified as surprise and sadness.

3.4 Decision-Making Fusion Emotional Recognition

3.4.1 Product Rule

The recognition results of product rule based methods are shown in Table 5.

Table 5. Product rule based emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	95.00%	70.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	80.00%	92.50%

The average recognition accuracy of product rule based method is 92.5%, which is higher than that of feature level fusion methods, i.e., 83.75% for PCA based fusion and 89.38 for LDA based fusion. The recognition accuracies of surprise and fear were low, where surprise is easily misidentified as fear and fear is easily misidentified as surprise or disgust.

3.4.2 Mean Rule

The recognition results of mean rule based method are shown in Table 6. The average recognition accuracy of mean rule based method is 91.25%, which is higher than that of LDA based fusion but lower than the product rule based fusion. Anger, neutral and sadness have the highest recognition accuracies. Fear is also difficult to recognize since fear is easily misidentified as surprise, disgust or sadness.

Table 6. Mean rule fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	90.00%	65.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	80.00%	91.25%

3.4.3 Summation Rule

The recognition results of summation based method are shown in Table 7. The average recognition accuracy is the same as the mean rule.

Table 7. Summation rule fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	90.00%	65.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	80.00%	91.25%

3.4.4 Maximum Rule

The recognition results of maximum rule based methods are shown in Table 8.

Table 8. Maximum rule fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	85.00%	70.00%	90.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	75.00%	90.00%

The average recognition accuracy of maximum rule based method is 90.0%, which is lower than that of product rule based method. Surprise and fear are not well recognized, where surprise is easily misidentified as fear and fear is easily misidentified as surprise or sadness.

3.4.5 Minimum Rule

The recognition results of minimum rule based method are shown in Table 9.

Table 9. Minimum rule fusion emotional recognition results

Emotion	Anger	Disgust	Fear	Happiness
Recognition accuracy	100.00%	100.00%	70.00%	95.00%
Emotion	Neutral	Sadness	Surprise	Average
Recognition accuracy	100.00%	100.00%	85.00%	93.75%

The average recognition accuracy of minimum rule based method is 93.75%, which is higher than all other methods. Anger, disgust, neutral and sadness are all 100% accurately recognized. Fear is also difficult to recognize, where fear is easily misidentified as surprise or disgust.

3.5 Analysis of Experimental Results

The following is the detailed analysis of audio emotional recognition, visual emotional recognition and multi-modal fusion emotional recognition experimental results.

3.5.1 Comparison with State of the Art

The comparison of the recognition results of this work with that of work [18] is shown in Fig. 3. As shown in Fig. 3, blue bar indicates the evaluation results by ten evaluators, which indicates that the database is effective for emotional recognition. The green bar denotes the recognition results of paper [18] and brown bar is the recognition result of this paper. From Fig. 3 we can see that for audio recognition this paper achieved better performance, but for visual recognition and fused recognition the results of this paper is slightly lower than that of paper [18]. The reason is mainly because that the classifier used in this paper is different from that in paper [18], as well as some details for feature extraction, etc. In general, the recognition results obtained in this paper indicate that the fusion of multi-modal information can achieve better results than single-mode emotional recognition.

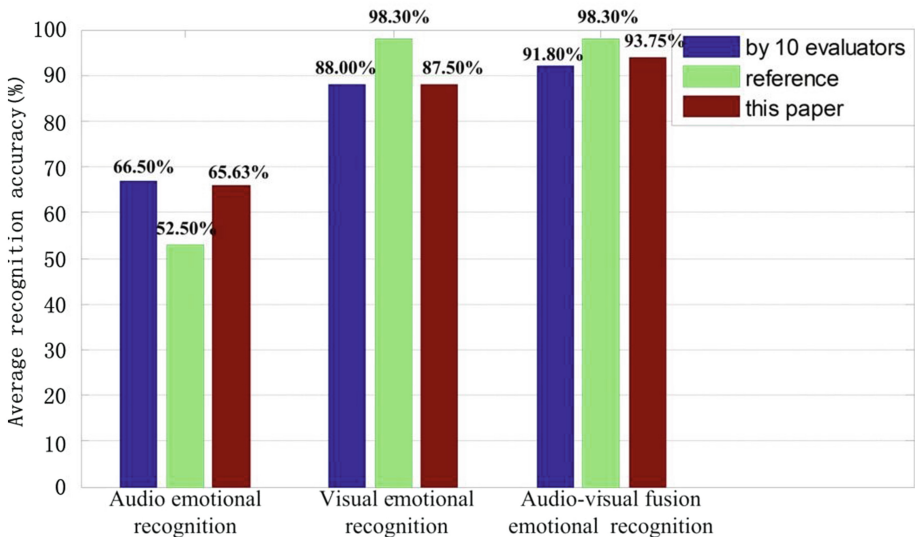


Fig. 3. Comparison of the recognition the results with work [18] (Color figure online)

3.5.2 Comparison of Feature Fusion Methods

In this paper, two different modal information fusion strategies are adopted, namely feature level fusion and decision level fusion. For feature level fusion, we employed PCA based feature fusion and LDA based feature fusion. For decision level fusion we adopt five different methods: product rule, mean rule, summation rule, maximum rule and minimum rule based feature fusion. The results of all these feature fusion methods are shown in Fig. 4.

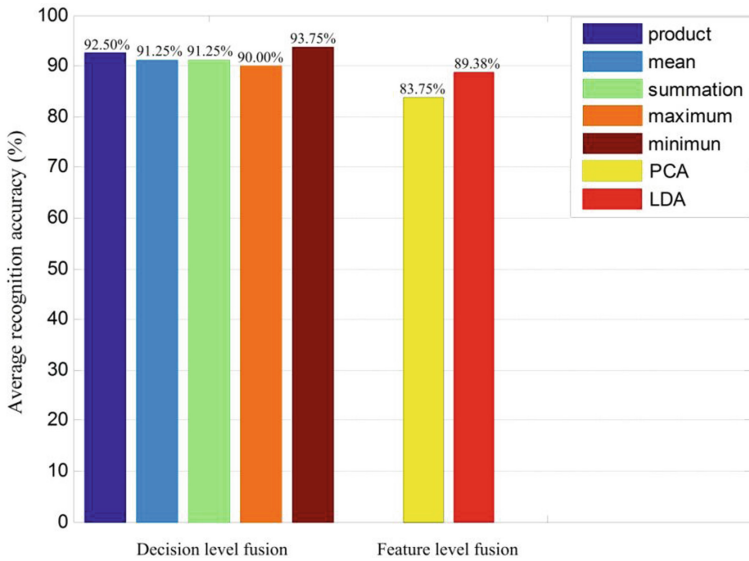


Fig. 4. Comparison of feature fusion methods.

From Fig. 4 we can see that decision level feature fusion methods outperform feature level feature fusion methods for multi-emotion recognition in this work. For instance, the minimum based feature fusion method achieved an average recognition accuracy of 93.75% while the PCA and LDA based feature fusion methods achieved an average recognition accuracy of 83.75% and 89.38% respectively. The LDA based feature fusion method outperformed PCA based feature fusion method, which is consistent with the work [18].

4 Conclusion

In this work, we implement several feature fusion methods for multi-modal emotion recognition, and evaluate all these methods on the SAVEE emotional database. The experimental results show that information fusion from multi-modal benefits the emotion recognition performance. And also, decision level feature fusion methods outperformed feature level feature fusion methods in this work.

Acknowledgements. This work is supported by National Natural Science Foundation of China Project (No. 61402219) and Postdoctoral Foundation Projects (Nos. LBH-Z14090, 2015M571417 and 2017T100243).

References

1. Cahn, J.: The generation of affect in synthesized speech. *J. Am. Voice I/O Soc.* **8**(1), 1–19 (1990)
2. Tato, R., Santos, R., Kompe, R., et al.: Emotional space improves emotion recognition. In: 7th International Conference on Spoken Language Processing, Denver, Colorado, pp. 2029–2032 (2002)
3. Liu, J., Chen, A., Ye, C.X., Bu, J.J.: Speech emotion recognition based on covariance descriptor and Riemannian manifold. *Pattern Recogn. Artif. Intell.* **05**, 673–677 (2009)
4. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. *Speech Commun.* **52**(7–8), 613–625 (2010)
5. Bozkurt, E., Erzin, E., Erdem, E.: Formant position based weighted spectral features for emotion recognition. *Speech Commun.* **53**(9), 1186–1197 (2011)
6. Schuller, B., Batliner, A., Steidl, S., et al.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun.* **53**(9), 1062–1087 (2011)
7. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech using source, system, and prosodic features. *Int. J. Speech Technol.* 1–25 (2012)
8. Zhang, Z., Lyons, M., Schuster, M., et al.: Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: Third IEEE International Conference on Face and Gesture Recognition, Nara, Japan, pp. 454–459 (1998)
9. He, K., Wang, G., Yang, Y.: Optical flow-based facial feature tracking using prior measurement. In: 7th IEEE International Conference on Cognitive Informatics, Stanford, CA, pp. 324–331 (2008)
10. Abboud, B., Davoine, F., Dang, M.: Facial expression recognition and synthesis based on an appearance model. *Sig. Process. Image Commun.* **19**(8), 723–740 (2004)
11. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: ACM Multimedia, pp. 47–56 (2014)
12. Lee, C.M., Narayanan, S.: Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* **13**(2), 293–303 (2005)
13. Zhao, S., Yao, H., Sun, X.: Video classification and recommendation based on affective analysis of viewers. *Neurocomputing* 101–110 (2013)
14. Zhao, S., Yao, H., Gao, Y., Ji, R., Ding, G.: Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE Trans. Multimedia* 1 (2016)
15. Kudiri, K.M., Said, A.M., Nayan, M.Y.: Emotion detection through speech and facial expressions. In: 2014 International Conference on Computer Assisted System in Health (CASH), pp. 26–31. IEEE (2014)
16. Zhao, S., Yao, H., Jiang, X.: Predicting continuous probability distribution of image emotions in valence-arousal space. In: ACM Multimedia, pp. 879–882 (2015)
17. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 27 (2011)
18. Haq, S., Jackson, P.J.B., Edge, J.: Audio-visual feature selection and reduction for emotion classification. In: Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP 2008), Tangelooma, Australia (2008)