The SYSU System for CCPR 2016 Multimodal Emotion Recognition Challenge

Conference Paper · November 2016

DOI: 10.1007/978-981-10-3005-5_58

4 authors, including:

Ming Li
Duke Kunshan University
93 PUBLICATIONS **977** CITATIONS

SEE PROFILE

The SYSU System for CCPR 2016 Multimodal Emotion Recognition Challenge

Gaoyuan He[1], Jinkun Chen[1], Xuebo Liu[1], and Ming Li[1,2(✉)]

[1] SYSU-CMU Shunde International Joint Research Institute, Foshan, China
[2] School of Electronics and Information Technology,
SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University,
Guangzhou, China
liming46@mail.sysu.edu.cn

Abstract. In this paper, we propose a multimodal emotion recognition system that combines the information from the facial, text and speech data. First, we propose a residual network architecture within the convolutional neural networks (CNN) framework to improve the facial expression recognition performance. We also perform video frames selection to fine tune our pre-trained model. Second, while the text emotion recognition conventionally deal with the clean perfect texts, here we adopt an automatic speech recognition (ASR) engine to transcribe the speech into text and then apply Support Vector Machine (SVM) on top of bag-of-words (BoW) features to predict the emotion labels. Third, we extract the openSMILE based utterance level feature and MFCC GMM based zero-order statistics feature for the subsequent SVM modeling in the speech based subsystem. Finally, score level fusion was used to combine the multimodal information. Experimental results were carried on the CCPR 2016 Multimodal Emotion Recognition Challenge database, our proposed multimodal system achieved 36 % macro average precision on the test set which outperforms the baseline by 6 % absolutely.

Keywords: Multimodal emotion recognition · Residual network · Speech recognition · Text emotion recognition

1 Introduction

A computer with powerful emotion recognition intelligence has a wide range of applications, such as human-computer interaction, psychological research, video recommendation services, etc. Although there are many existing works on this topic, understanding human emotion precisely is still a challenging task for researchers. First, signals from multiple modalities provide complementary information about emotion states. How to make full use of this information to make an accurate decision has been a difficult point. Second, there are many variances in the emotion datasets, such as age, gender, identity, background, etc. Third, human emotion can be characterized in the continue Valence-Arousal

T. Tan et al. (Eds.): CCPR 2016, Part II, CCIS 663, pp. 707–720, 2016.
DOI: 10.1007/978-981-10-3005-5_58

space [1]. Labelling the data with categorical classes can result in ambiguous and inconsistent labels among multiple evaluators.

Although there remain a number of difficulties in solving the emotion recognition problem, many existing works have been proposed to improve the performance. Many of those early works focus on single modality or "lab-controlled" environments [2,3]. Recent works on emotion recognition mainly focus on recognizing the emotional states in more real and spontaneous environments with multimodal signals [4]. In this paper, we first utilize three emotion recognition subsystems that analyzing video, text, and speech signals, respectively. Then we fuse these three subsystems on the score level to further enhance the performance.

For the video based subsystem, traditional approaches are based on hand-engineered features, such as Local Binary Pattern (LBP) [5] and Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [6] features. They are proved to work well on certain datasets, but lack the generality on other datasets [7]. Recently, deep learning methods have achieved state-of-the-art results in many computer vision tasks [8,9]. Therefore, researchers also start to apply the deep learning methods on the emotion recognition task [10]. In this paper, we propose a residual network [11] architecture within the convolutional neural networks (CNN) framework to improve the facial expression recognition performance. In addition, we also propose a method to select certain video frames to fine tune our pre-trained model.

For the text based subsystem, some machine learning based methods have been proposed [12]. Conventionally, in multimodal emotion recognition, the text emotion recognition mainly deals with perfect texts or manually transcribed subtitles while the speech processing focuses on the acoustic features rather than the semantic meanings of speech utterances. And in recent years, the automatic speech recognition (ASR) has boomed and been potentially practical with higher accuracy, which means that the ASR error rate is not the principal detrimental factor any more. All these factors motivated us to integrate the emotion recognition on ASR generated text into our system, which contributes to better practicality and robustness. We use millions of short film subtitles and short conversation sentences to build a N-gram language model for rescoring the ASR lattices. And more importantly, we proposed a hierarchical classifier constructed with two Support Vector Machines to avoid the over fitting problem on small datasets.

The remainder of the paper is organized as follows. The related works are explained in Sect. 2. Our proposed methods are described in Sect. 3. Experimental results are presented in Sect. 4 while conclusions and future works are provided in Sect. 5.

2 Related Works

2.1 The Video Based Subsystem

Convolutional neural networks (CNN) has already generated state-of-the-art performance in a wide range of computer vision fields. Novel modified CNN models

have also been proposed in recent years, for example, the bottleneck structure in Residual-Net [11] has been proposed to reduce the number of parameters and reduce the computational complexity.

Recently, deep learning method is also heavy used in emotion recognition task. In particular, Yu [10] proposed an innovative voting framework to combine multiple CNN models to get the final decision. In addition, Yu used the FER [13] dataset to train a model firstly, then fine tune the model on SFEW [14]. Kahou [15] proposed a hybrid CNN-RNN architecture for facial expression recognition in video. A CNN model was pre-trained on additional face expression datasets to extract facial features as the input of a simple recurrent neural network (RNN) model. He also extracted audio features and activity features from videos as multiple modalities. Two different fusion strategies were proposed to take advantage of these different modalities, operating on the feature and decision level, respectively.

2.2 The Text Based Subsystem

Some machine learning based methods for text emotion recognition have been proposed, such as naive Bayes (NB), maximum entropy (ME) classification and support vector machine (SVM) [12], which are generally based on bag-of-words (BoW) model [16] in word-level studies. The emotions or sentiments associated with topics is based on topic model [17], which introduces an intermediate topical layer into latent Dirichlet allocation (LDA) [18,19]. In the task of text emotion recognition, the methods based on BoW model work well on short texts or small datasets while the methods based on topic model have better performances on long documents. In this challenge, since the speech utterances are short, we employ SVM based on BoW feature to handle the multi-class text based emotion recognition task. Moreover, we adopt a Mandarin ASR system to automatically decode the speech utterances into text as the input of our text based emotion recognition system. Some works related to emotion recognition on ASR generated text have been reported [20]. However, for mandarin, there are very few prior works on the idea of using automatically transcribed text for emotion recognition and participating the visual-audio multimodal fusion.

2.3 The Speech Based Subsystem

It has been shown in [21,22] that speech emotion recognition can be modeled at various levels, such as acoustic, prosodic, phonetic and linguistic levels. Due to the different aspects of modeling, combining different classification methods with different features can significantly improve the overall performance. The openSMILE toolkit [23] has widely been used to perform extraction of utterance level acoustic and prosodic features followed by the Support Vector Machine (SVM) for the subsequent classification on the spontaneous short utterances [24].

Despite the openSMILE features, several Gaussian Mixture Model (GMM) based supervectors have also been proposed as features for paralinguistic speaker states recognition [25,26]. These supervectors originally were proposed for

speaker verification and language identification tasks but also performed well in the paralinguistic challenges. However, when the duration of speech utterance is very short (e.g. less than 2 s), the performance of those supervectors replyiproceeding13ng on the first-order Baum-Welch statistics (mean supervector, i-vector, Maximum Likelihood Linear Regression (MLLR) supervector, etc. [25]) drops as there are not enough feature frames to calculate the sufficient statistics. The zero-order statistics based posterior probability feature achieves better performance in these short duration scenarios with limited training data [25].

It is worth noting that deep learning has been applied on the speech emotion recognition recently [27]. However, due to the lack of large scale mandarin emotional speech database for model training, in this work we only extract the conventional openSMILE based utterance level feature and MFCC GMM based zero-order statistics feature, then use SVM to perform the classification.

3 Methods

The Multimodal Emotion Recognition Challenge (MEC) is part of the 2016 Chinese Conference on Pattern Recognition (CCPR). The task of this competition is to classify the given audio and video into 8 emotion categories, i.e., "happy", "sad", "angry", "surprise", "disgust", "worried", "anxious", and "neutral". The dataset is the Chinese Natural Audio-Visual Emotion Database (CHEAVD) [28], which contains 140 min spontaneous emotional segments extracted from Chinese movies and TV programs. The 2852 samples are divided into three sets: training set, validation set and testing set, containing 1981, 243 and 628 clips, respectively. More details about the database and the challenge can be found in [29]. The emotion recognition we built consisted of three subsystems, which are video, text, and speech based emotion recognition. In this section, we will respectively explain the methods used in each subsystem in details.

3.1 The Video Based Subsystem

In addition to the video data, the organizers also provide face images extracted by IntraFace toolkit [30], where the OpenCV's Viola and Jones face detector is applied for face detection and initialization of the Intraface tracking library. We refer to this dataset as CHEAVD-faces. Each video clip contains multiple frames, but a particular emotion state may only occur in some certain frames. Thus labeling all frames with the same emotion label will introduce noise. Therefore, two additional emotion datasets of static images, Facial Expression Recognition dataset (FER2013) [13] and Static Facial Expression in the wild dataset (SFEW) [14], were used to pre-train our CNN model. In addition, we do not use all frames of each competition video clip to fine-tune our model, but select a portion of frames.

3.1.1 Pre-processing

Since CHEAVD-faces, FER2013 and SFEW use different face detection and alignment techniques. Thus we re-aligned all datasets to FER2013 using the method proposed by Kahou [15]. First, we detected five facial keypoints(left eye center, right eye center, nose tip, left mouse corner and right mouse corner) for all images on the FER2013, SFEW, and CHEAVD-faces training set using the method in [9]. Second, for each dataset we computed the mean shape by averaging the coordinates of keypoints. Third, we use a similarity transformation between mean shapes to map all datasets to FER2013. Finally, all three dataset images are preprocessed with standard histogram equalization, followed by normalization with mean and standard deviation. CHEAVD-faces validation and test sets were mapped using the same transformation learned from the training set.

3.1.2 The Residual Network Architecture

We pre-train our residual network on the FER+SFEW datasets. Figure 1 shows an overview of the networks architecture. In our CNN model, the stride of every convolutional layers is 1. The size of first convolutional filter is 5×5, and the number of first convolutional layer's channels are 64. After the first layer, the size of each convolutional filer in main branch is 3×3, the size of each convolutional filter in another branch is 1×1. The stride length of each pooling layer is 2, and the window size for each pooling layer is 3×3. We use average pooling layer in our CNN model, because after standard histogram equalization, the global features of images are more distinct than local features. There are three fully connected layers at the end. To avoid over-fitting, we add dropout after each fully connected layer.

3.1.3 Networks Pre-training on FER+SFEW

We use a deep learning library named Keras [31] to pre-train our CNN model on the FER+SFEW datasets. The stochastic gradient descent (SGD) method is taken to optimize our loss function. The batch size is set to 128. The initial learning rate is set to 0.01 while the minimum learning rate is set to 0.0005.

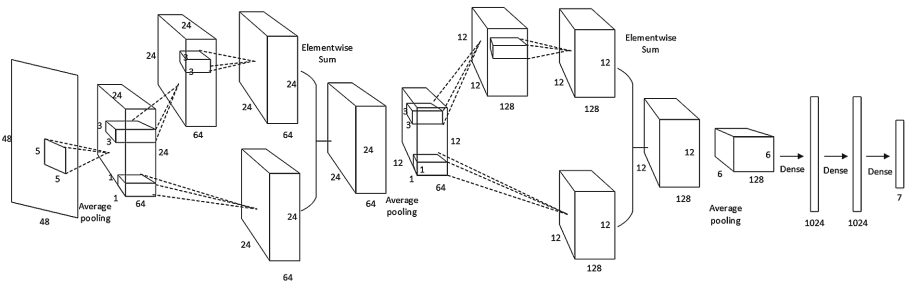

Fig. 1. The residual network for emotion recognition

The loss and trained network parameters in each epoch are recorded. If the training loss keeps increasing in more than five consecutive epochs, the learning rate is reduced by half. We then select the model with the best accuracy on held-out development data as our pre-train model.

3.1.4 Networks Fine-Tuning on CHEAVD-Faces

The pre-trained model achieved 65.65 % accuracy on the FER test dataset. As previously stated, as labeling all frames in one video clip with the same emotion label will introduce noise, we do not use all image frames to fine-tune our model. Figures 2 and 3 show the method that we use to select training CHEAVD-faces dataset for model fine-tuning. For each video in CHEAVD training dataset, we use our pre-trained model to predict all frames of the video. Our pre-trained model will produce an estimated label of seven emotions categories for every face image frame that is fed to the model, we select the most common one as the final label of the video. If the label is different from the ground truth, we use all the image frames of this video to fine-tune our model. Otherwise, we selected the frames with estimated labels matching with the ground truth, discard the frames with estimated labels are different from the ground truth. What needs illustrating is that since FER and SFEW both have seven categories, but CHEAVD-faces has eight categories. In this step, we merge the worried category and anxious category into fear category.

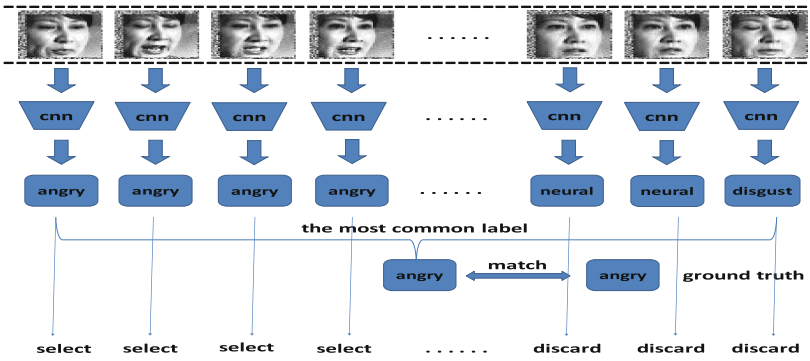

Fig. 2. The estimated label is as same as ground truth

Next we fine-tune our CNN model on the selected training CHEAVD-faces dataset. As previously stated, these three datasets have different categories. Therefore, we change the size of last fully connected layer from seven to eight and only retrain the weights of fully connected layers in the pre-trained model. Because the macro average precision (MAP) is the final criterion of this competition, the validation accuracy is based on MAP. We not only record MAP but also confusion matrix at each epoch. We finally choose the fine-tuning model with relatively high MAP and relatively balanced confusion matrix. On the test

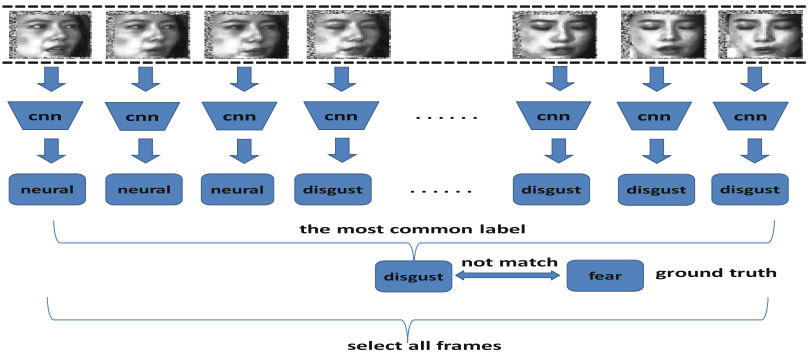

Fig. 3. The estimated label is different with ground truth

dataset, our fine-tuned model will produce an estimated label of eight emotions categories for each face image frame that is fed to the model, we select the most common one as the final label of the test video clip.

3.2 The Text Based Subsystem

Although some films are dubbed and have subtitles, we recognize the utterances from speech signals instead of using the existing subtitles. This is for better practicality and applicability in emotion recognition on any speech signals without word describing data provided. The following paragraphs explain the ASR procedure on speech signals and emotion recognition on film subtitles.

3.2.1 Speech Recognition on CHEAVD-Audio

We build our Mandarin automatic speech recognition system based on the KALDI toolbox [34]. To decrease the word error rate in ASR, we crawl about fifteen million sentences of short film subtitles from the Internet, then a special N-Gram language model is built with these subtitles and is applied in the ASR rescoring. To clarify, the subtitles of the movies or TVs in CHEAVD dataset are not contained in the training set for ASR language model. In order to measure the performance of our ASR engine, first, we manually transcribe 200 audio waves provided by CHEAVD dataset into text to form the speech evaluation data and ground truth. Second, we call the speech recognition service of iFLYTEk[1] and Microsoft[2], as well as our ASR engine to generate three versions of ASR recognized text. Third, we calculate the word error rate[3] as the metric of speech recognition.

As the experimental results of the validation dataset shown in Table 1, Sect. 4, the iFLYTEK speech recognition service achieves the best performance among

[1] http://www.xfyun.cn/services/voicedictation.
[2] https://www.microsoft.com/cognitive-services/en-us/speech-api.
[3] https://en.wikipedia.org/wiki/Word_error_rate.

the three ASR engines and our engine ranks the second. However, the differences of emotion recognition results on ASR decoded texts, corresponding to ASR engine of iFLYTEK and our laboratory respectively, are nearly negligible. Thus, we choose our in-house speech recognition engine in this challenge.

3.2.2 Emotion Recognition on Short Text

Considering the film subtitles are very colloquial and may carry emotional meanings, we extract 267043 short conversations with emotional labels from thousands of Chinese novels to fit and train the classifier. First, we manually create a dictionary containing a set of adjectives which have emotional tendencies and classify these adjectives into 36 emotion categories. The set of 8 emotional categories in the challenge, i.e., "happy", "sad", "angry", "worried", "anxious", "surprise", "disgust" and "neutral", is a subset of the 36 categories. Second, we find out all the conversions which have the structures fulfilling the pattern of "someone [says] {ADVERB} something", where ADVERB can be any adverb, to qualify the verbs such as "say", "talk", "speak", etc. Third, we select the conversation if the adverb has a strong emotional tendency, and choose the corresponding adjective as the emotional label of the conversation. Finally, we obtain 267043 conversation sentences with 360 different labels approximately and classify all conversation sentences into 36 emotional categories according to the dictionary we built.

We propose a hierarchical classifier constructed with two SVMs to recognize emotions on the ASR generated text. With the aforementioned conversations dataset, we calculate the term frequencies as the text feature and train the first stage SVM classifier, which acts as a tandem feature extractor. For each utterance u_i from the CHEAVD-audio data, either from train dataset, validation dataset or test dataset, is sent to the first stage SVM classifier to generate a 36-dimension feature vector $P_i = (p_0, p_2, ..., p_{35})$, where p_j $(0 \leq j \leq 35)$ is the probability of the c_j class in the 36 categories. Furthermore, another backend SVM classifier is trained with the 36 dimensional tandem features to perform 8-class classification. The flow of emotion recognition on ASR generated text is shown as Algorithm 1.

In the beginning, we chose the classifiers of NB and SVM respectively and trained classifiers with BoW feature of the training dataset directly. And then we predicted the labels of the samples in test dataset. Unfortunately, this yielded over fitted results. Most of the samples were predicted as "neutral" for the reason that forty percent of samples in training set are "neutral". The small sample size of dataset can also result in over fitting. With the two stage hierarchical SVM method on ASR generated text, we avoid the over fitting problem and improve both the accuracy and average macro precision of text emotion recognition.

3.3 The Speech Based Subsystem

Due to the lack of large scale mandarin emotional speech database for model training, in this work, we only extract the conventional openSMILE based

Algorithm 1. Flow of emotion recognition on film subtitles

Require:
The datasets of short conversations and its labels, Cs, Lb_cs;
The utterances of ASR generated text of train dataset and its labels, U_trn, Lb_trn;
The utterances of ASR generated text of test dataset, U_tst;

Ensure:
The labels of the ASR generated text of test set, Lb_tst;

1: Train the text SVM classifier $text\_CLF$ with Cs, Lb_cs;
2: Input U_trn and U_tst to $text\_CLF$, get feature matrix $mat\_trn_{|U\_trn|\times 36}$ and $mat\_tst_{|U\_tst|\times 36}$ respectively;
3: Train another SVM classifier vec_CLF with feature matrix $mat\_trn_{|U\_trn|\times 36}$ and labels Lb_trn;
4: Find the optimal parameters (C, γ) for RBF kernel function with 5-fold cross validation and grid search;
5: Classify the vectors in test feature matrix $mat\_tst_{|U\_tst|\times 36}$ via vec_CLF, get the labels of test set Lb;
6: **return** Lb;

utterance level feature and MFCC GMM based zero-order statistics feature, then adopt SVM to perform the classification.

The utterance level openSMILE features provided by the challenge organizers are extracted by openSMILE with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [29].

For each utterance in the training and validation sets, zero-order statistics feature extraction is performed using the Universal Background Model (UBM) trained by the training dataset. Given a frame-based MFCC feature $\boldsymbol{x_t}$ and the GMM-UBM λ with M Gaussian components (each component is defined as λ_i),

$$\lambda_i = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, i = 1, \cdots, M, \tag{1}$$

the occupancy posterior probability is calculated as follows,

$$P(\lambda_i|\boldsymbol{x_t}) = \frac{w_i p_i(\boldsymbol{x_t}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\Sigma_{j=1}^{M} w_j p_j(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})}. \tag{2}$$

This posterior probability can also be considered as the fraction of this feature $\boldsymbol{x_t}$ coming from the i^{th} Gaussian component which is also denoted as partial counts. The larger the posterior probability, the better this Gaussian component can be used to represent this feature vector. The zero-order statistics supervector is defined as follows,

$$\boldsymbol{b} = [b_1, b_2, \cdots, b_M], b_i = \frac{y_i}{T} = \frac{1}{T}\Sigma_{t=1}^{T} P(\lambda_i|\boldsymbol{x_t}), \tag{3}$$

$$MGPP_{feature} = \sqrt{\boldsymbol{b}}. \tag{4}$$

Equation 3 is for calculating the zero-order Baum-Welch statistics and is exactly the same as the weight updating equation in the expectation-maximization (EM)

algorithm in GMM training. In order to apply Bhattacharyya probability product (BPP) kernel [36], we adopt $\sqrt{\boldsymbol{b}}$ as our zero-order statistics features [25].

Based on the above mentioned features, we employed the LibLinear toolbox [35] for the backend linear kernel SVM modeling. To reduce the data imbalance problem, we set the SVM weight parameter of the i^{th} emotion class as the inverse ratio of its samples vs the neutral class $(\#neutral)/(\#i^{th}class)$. Finally, the two SVM prediction scores with OpenSMILE features and zero-order statistics features are linear fused as the outputs of our speech based subsystem.

3.4 Score Level Fusion

Due to the limited amount of training data, we simply employed the weighted summation fusion approach with parameters tuned by cross validation. When the evaluation was performed on the testing set of the CHEAVD database, both the training and validation sets were used for modeling and the weight vector was exactly the same as the one tuned on the validation set. It is worth noting that other advanced score fusion approaches, like the logistic regression method in the popular FoCal toolkit [37], can also be adopted here to increase the performance which is a topic for our future work.

4 Experiment Results

As classes are unbalanced, this challenge choose macro average precision (MAP) as the primary measure to rank the results, and secondly the accuracy. The formula of MAP and accuracy are given in Eqs. 5 and 8, respectively.

$$macro\ average\ precision = \frac{1}{s} \times \sum_{i=1}^{s} P_i \tag{5}$$

$$P_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$R_i = \frac{TP_i}{TP_i + FP_i} \tag{7}$$

$$accuracy = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s}(TP_i + FN_i)} \tag{8}$$

In these four formulas, s represent the number of the emotion labels. FP_i, FN_i TP_i represent the number of false positive , the number of false negative and the number of true positive in the i^{th} emotion class, respectively. P_i is the precision of the i^{th} emotion class and R_i is the recall of the i^{th} emotion class [29].

The Table 1 presents the word error rate and emotion recognition result on validation dataset. The ASR engine of iFLYTEK has the lowest word error rate and our engine ranks the second. As for the accuracy and MAP, we get the similar emotion recognition result on the text generated by ASR of iFLYTEK and ours respectively.

Table 1. Word error rate and emotion recognition results of validation set

	Our tools	iFLYTEK API	Microsoft API
Word error rate	32.26 %	24.78 %	47.94 %
Emotion accuracy	40.1 %	40.2 %	38.8 %
Average macro precision	46.4 %	45.8 %	42.9 %

There are 1981, 243 and 628 samples in the train, validation and test dataset respectively in this challenge. We use 10-folder cross-validation method to select our model. Since the validation dataset is very unbalanced, we find that the performance on validation dataset does not have representativeness. But the test dataset is relatively balanced, we only compare our system with the baseline system on the test dataset.

The Table 2 lists the accuracy and MAP of our system and the baseline system. In each subsystem, our methods are better than the baseline. And the fusion system gives 29.93 % classification accuracy and 36.42 % MAP, which has the best recognition rate. It shows that our fusion method can truly improve the performance.

Table 2. Comparison of baselines and our result

	Accuracy(%)			MAP(%)		
	Baseline-1	Baseline-2	Our result	Baseline-1	Baseline-2	Our result
Video	19.59	21.02	**27.38**	34.28	30.41	**36.56**
Text			**27.10**			**33.84**
Audio	24.36	21.91	**26.11**	24.02	20.48	**25.98**
Fusion	24.52	21.18	**29.93**	24.53	30.63	**36.42**

The Figs. 4 and 5 show the confusion matrices of our system and the baseline system respectively. In the baseline system, the video and audio subsystem confusion matrices are very unbalanced. Such as, for audio confusion matrix of the baseline system, the precision and recall of disgust and worried categories are both 0 %. The precision of angry and anxious categories are both 100 %, but the recall of angry and anxious are 9.80 % and 0.97 % respectively. But for our audio confusion matrix, none of the precision or recall of these categories is 0 % or 100 %. Therefore, our system is much more robust than the baseline system.

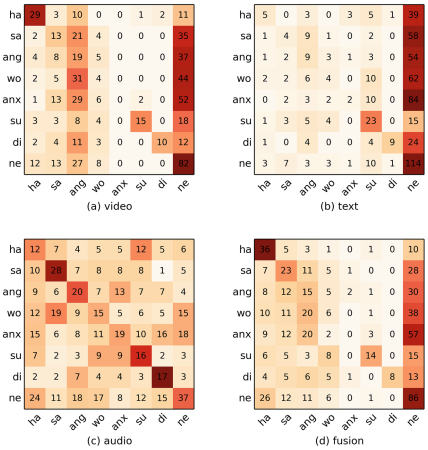

Fig. 4. Our system

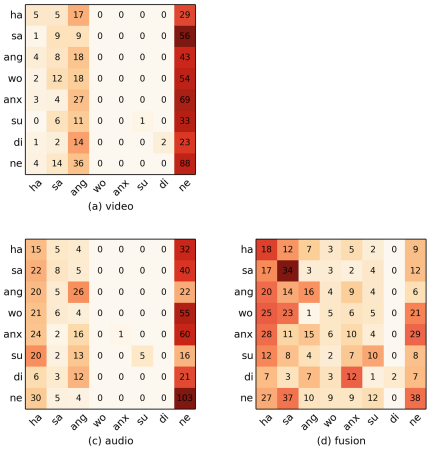

Fig. 5. Baseline system

5 Conclusion

In this paper, we propose a multimodal emotion recognition system that combines the information from the facial, text and speech data. The residual network architecture within the CNN framework and video frames selection for fine tuning can significantly improve the facial expression recognition performance. We extract the openSMILE based utterance level feature and MFCC GMM based zero-order statistics feature for the subsequent SVM modeling in the speech based subsystem and these two features are complimentary. For the ASR decoded text subsystem, the hierarchical classifier constructed with SVMs performs well and provide a new idea to avoid over fitting on short texts and small dataset. In this challenge, our multimodal system achieves better macro average precision in comparison to baseline, which proves its effectiveness.

Acknowledgement. This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), the Fundamental Research Funds for the Central Universities (15lgjc10) and National Key Research and Development Program (2016YFC0103905).

References

1. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Sci. Inf. **44**(4), 695–729 (2005)
2. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Third IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings, pp. 200–205 (1998)

3. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vision Comput. **28**(5), 807–813 (2010)
4. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 461–466 (2014)
5. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vision Comput. **27**(6), 803–816 (2009)
6. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
7. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia **19**(3), 34–41 (2012)
8. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1106–1114 (2012)
9. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483 (2013)
10. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442 (2015)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
13. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., et al.: Challenges in representation learning: a report on three machine learning contests. In: International Conference on Neural Information Processing, pp. 117–124 (2013)
14. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2106–2112 (2011)
15. Kahou, S.E., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 467–474 (2015)
16. Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cybern. **1**(1–4), 43–52 (2010)
17. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine learning, pp. 977–984. ACM (2006)
18. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
19. Ramage, D., Hall, D., Nallapati, R., et al.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, vol. 1, pp. 248–256 (2009)
20. Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., Steidl, S.: Emotion recognition using imperfect speech recognition. ISCA (2010)

21. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. **43**(2), 155–177 (2015)
22. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
23. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)
24. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: INTERSPEECH, pp. 312–315 (2009)
25. Li, M., Han, K.J., Narayanan, S.: Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Comput. Speech Lang. **27**(1), 151–167 (2013)
26. Li, M., Metallinou, A., Bone, D., Narayanan, S.: Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1937–1940 (2012)
27. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204 (2016)
28. Bao, W., Li, Y., Gu, M., Yang, M., Li, H., Chao, L., Tao, J.: Building a Chinese natural emotional audio-visual database. In: 2014 12th International Conference on Signal Processing (ICSP), pp. 583–587 (2014)
29. Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., Jia, J.: MEC 2016: the multimodal emotion recognition challenge of CCPR 2016, submitted to CCPR 2016
30. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
31. Chollet, F.: Keras. Github (2015). https://github.com/fchollet/keras
32. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. Comput. Speech Lang. **16**(1), 69–88 (2002)
33. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2003)
34. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011)
35. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**(Aug), 1871–1874 (2008)
36. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. J. Mach. Learn. Res. **5**, 819–844 (2004)
37. Brümmer, N.: FoCal multi-class: toolkit for evaluation, fusion and calibration of multi-class recognition scores–tutorial and user manual– (2007). http://sites.google.com/site/nikobrummer/focalmulticlass