Video Emotion Recognition in the Wild Based on Fusion of Multimodal Features

Shizhe Chen
School of Information
Renmin University of China
cszhe1@ruc.edu.cn

Xinrui Li
School of Information
Renmin University of China
lialialia@ruc.edu.cn

Qin Jin*
School of Information
Renmin University of China
qjin@ruc.edu.cn

Shilei Zhang
IBM Research Lab
Beijing, China
slzhang@cn.ibm.com

Yong Qin
IBM Research Lab
Beijing, China
qinyong@cn.ibm.com

ABSTRACT

In this paper, we present our methods to the Audio-Video Based Emotion Recognition subtask in the 2016 Emotion Recognition in the Wild (EmotiW) Challenge. The task is to predict one of the seven basic emotions for the characters in the video clips extracted from movies or TV shows. In our approach, we explore various multimodal features from audio, facial image and video motion modalities. The audio features contain statistical acoustic features, MFCC Bag-of-Audio-Words and MFCC Fisher Vectors. For image related features, we extract hand-crafted features (LBP-TOP and SPM Dense SIFT) and learned features (CNN features). The improved Dense Trajectory is used as the motion related features. We train SVM, Random Forest and Logistic Regression classifiers for each kind of feature. Among them, MFCC fisher vector is the best acoustic features and the facial CNN feature is the most discriminative feature for emotion recognition. We utilize late fusion to combine different modality features and achieve a 50.76% accuracy on the testing set, which significantly outperforms the baseline test accuracy of 40.47%.

CCS Concepts

•**Computing methodologies** → *Artificial intelligence; Computer vision; Machine learning;*

Keywords

Video Emotion Recognition, Multimodal Features, CNN, Late Fusion

*The corresponding author.

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2997629

1. INTRODUCTION

Automatic video emotion recognition is a challenging task for machine learning, and it has a wide range of applications such as human computer interaction, e-learning and mental health care for the depressed people.

In order to provide benchmark for the automatic emotion recognition in the spontaneous environment, the Emotion Recognition in the Wild challenges [1, 2] have been held for four years using the Acted Facial Expression in Wild (AFEW) database. The AFEW database consists of short emotional video clips extracted from movies or TV shows, which mimics the emotional behaviours in real life. The task is to assign the video clip with a single emotion label from the seven basic emotions (Anger, Disgust, Fear, Happiness, Neutral, Sad and Surprise). The emotion category distributions of the data on training and validation sets are presented in Table 1. The difficulty of the task results from the wide variety in the videos such as context scenes, subjects, poses, illuminations, occlusions and so on.

Previous research works [3–5] on the challenge show the great benefits from the fusion of multimodal features. Facial expression features play the most important role for emotion recognition in the challenge. The winner of the EmotiW2015 [3] automatically learns the AU-aware features for each pair of different emotion and encodes the latent relations of the learned AU patches as the facial feature. The CNN based features [6] have shown the state-of-the-art performance for expression recognition. Hand-crafted visual features such as HOG, LBP-TOP, LPQ-TOP are also proved to be effective in previous works [5]. The performance of the audio features can not beat that of the facial features, but the combination of them can further boost the emotion recognition performance due to the complementary information of audio and visual modalities. Text features extracted from the speech content in the videos, however, are less useful as shown in the work [7] because of the inaccurate transcription from the speech recognizer and insufficient training data. The Bag-of-Words framework is typically used to generate the fixed dimensional video-level features. To capture temporal information, Samira et al. [8] propose to use the recurrent neural networks to encode the sequential information. The SVM is often used as the classifier for each kind of features.

The fusion strategies for different modalities can be divided into 3 categories, namely feature-level (early) fusion,

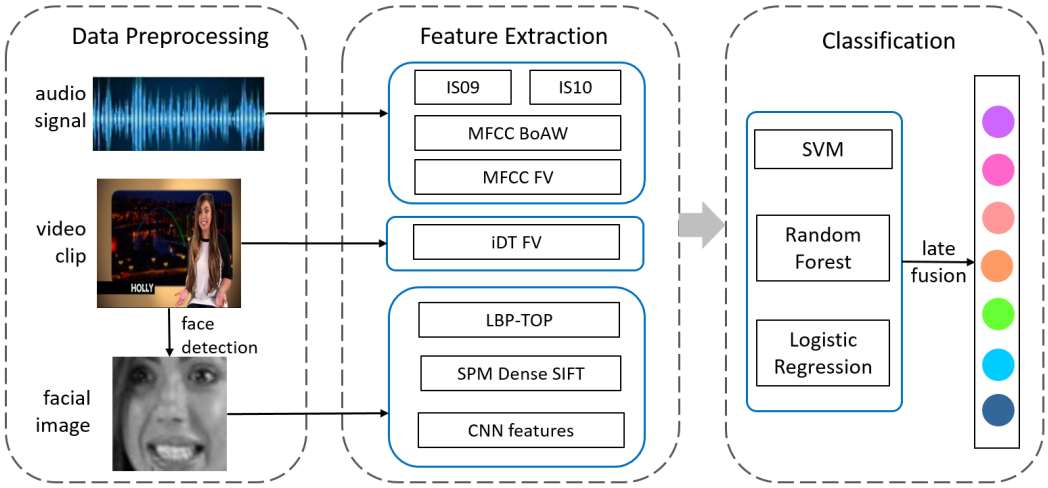


Figure 1: The framework of multimodal feature fusion system

Table 1: Emotion Category Distributions on Training and Validation Set

	ang	dis	fea	hap	neu	sad	sur	total
train	133	74	81	150	144	117	74	773
val	64	40	46	63	63	61	46	383

decision-level (late) fusion and model-level fusion [9]. Early fusion concatenates different features as the input feature for classifiers. It has been widely used in the literature to successfully improve performance [10]. However, it suffers from the curse of dimensionality. Late fusion trains a second level model like RVM [11] using the predictions of different modality features and thus eliminates the high dimensionality of feature concatenation. But it ignores interactions and correlations between the features. Model-level fusion is a compromise between the two extremes. Some typical model-level fusion methods are kernel fusion [12] or concatenating hidden layers of different neural networks.

In this task, we extract various multimodal features from audio, facial image and video motion. Acoustic features include statistical acoustic features, MFCC Bag-of-Audio-Words and MFCC Fisher Vector. The facial expression features contain hand-crafted features such as LBP-TOP and dense SIFT and learned deep convolution neural network (CNN) features. The improved Dense Trajectory is used as the motion feature. We utilize multiple classifiers including SVM, Random Forest and Logistic Regression for each unimodal feature to explore the effectiveness of different features. Late fusions are applied over different combinations of these features. The framework of our emotion recognition system is shown in Figure 1.

The paper is organized as follows. Section 2 describes all the multimodal features we extract in the challenge in details. Section 3 provides our experimental results. Section 4 concludes the paper and presents our future work.

2. MULTIMODAL FEATURES

Emotions are conveyed through multi-modalities. In this section, we present all the features from audio, face image and video motion modalities to represent the emotion expressed in the video from different aspects.

2.1 Audio Features

2.1.1 *Statistical Acoustic Features*

Statistical acoustic features are proved to be effective in speech emotion recognition. We use the open-source toolkit OpenSMILE [13] to extract two kinds of statistical acoustic features `IS09` and `IS10`, which use the configuration in INTERSPEECH 2009 [14] and 2010 [15] Paralinguistic challenge respectively. Low-level acoustic features such as energy, pitch, jitter and shimmer are first extracted over a short-time window. And then statistical functions like mean, max are applied over the set of low-level features to generate sentence-level features. The difference between `IS09` and `IS10` is that `IS10` uses more low-level features and statistical functions. The dimensionality of the `IS09` and `IS10` features are 384 and 1582, respectively.

2.1.2 *MFCC based Features*

The Mel-Frequency Cepstral Coefficients (MFCCs) [16] are the most widely used low-level features and have been successfully applied in many tasks such as speech recognition and audio event detection. Therefore, we use MFCCs as our fundamental frame-level feature which are extracted with window of 25ms and shift of 10ms. We use two encoding strategies, Bag-of-Audio-Words (BoAW) [17] and Fisher Vector Encoding (FV) [18], to transform the variant length of MFCCs to the sentence-level features. Both of the BoAW and FV are based on an intermediate representation, the audio vocabulary built in the low-level feature space.

Bag-of-Audio-Words: We first generate an acoustic codebook by K-means clustering algorithm with K=1024 using MFCCs features in the training set. Then we assign the MFCCs in one sentence to the discrete set of codewords in the codebook, thus providing a histogram of codewords count. L1-norm is used to get the probability distributions on the codebook for each sentence.

Fisher Vector: The Fisher encoding uses Gaussian Mixture Models (GMM) to construct an audio word dictionary.

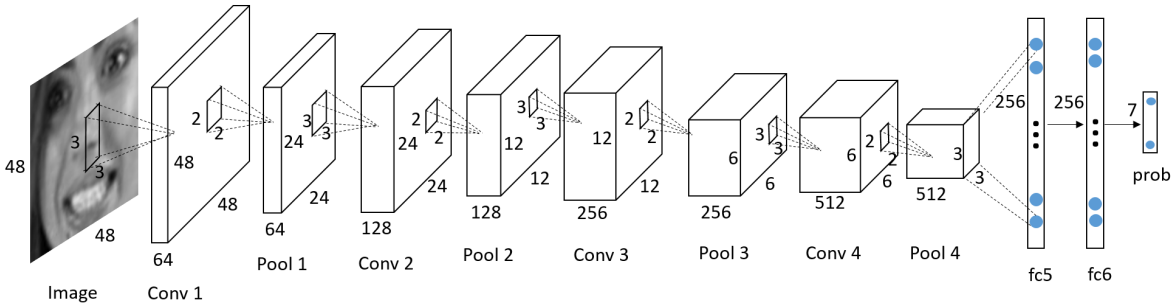


Figure 2: The structure of the convolution neural network

We compute the gradient of the log likelihood with respect to the parameters of the GMM to make the model best fit the MFCC features in one audio segment. The Fisher Vector is the concatenation of these partial derivatives. In our experiment, we use a GMM with 8 mixtures to generate FV representation and apply L2-norm on the FV features.

2.2 Facial Image Features

2.2.1 Face Pre-processing

We use Deformable Parts Model proposed in [19] to detect faces in each video frame and then align faces with facial landmarks extracted by methods in [20, 21]. Faces are scaled into the size of 128×128.

2.2.2 LBP-TOP

The Local Binary Patterns of Three Orthogonal Planes (LBP-TOP) [22] has been successfully applied to facial expression recognition which models the dynamic image textures. We first divide every facial image in a video into 4×4 patches. Then we extract the statistical histogram of LBP features from the XY, XT and YT planes and concatenate them into a fixed-dimensional representation.

2.2.3 SPM Dense SIFT

Scale Invariant Feature Transform (SIFT) [23] is wildly used as the image feature, which is invariant to image scale and rotation. In this task, we first extract dense SIFT descriptors in every 7×7 patch with one shift pixel. We encode the local features using the Bag-of-Feature (BoF) [24] framework. A codebook with 1024 codewords is generated by K-means clustering and then each local SIFT feature can be represented by the distributions on its top-10 nearest codewords. Then we use the Spatial Pyramid Matching (SPM) [4] method to encode the set of SIFT features in an image. To be specific, a image is partitioned into blocks with different levels (1×1, 2×2, 4×4). In each image block, the transformed SIFT BoF local features are applied with mean pooling to generate the fixed dimensional block features. The block features are concatenated as the image feature. The video features are obtained using max pooling over all the facial images features, the dimensionality of which is 21504. We call this feature as `DSIFT`.

2.2.4 CNN Features

The convolution neural networks (CNN) [25] have shown the state-of-the-art performance in many visual tasks such as object detection and scene recognition. It benefits from its hierarchical network architecture and parameters sharing in different local patches. In this task, we train our CNN models with the Facial Expression Recognition dataset (FER2013) [26] which contains 35,887 images with seven basic expressions: angry, disgust, fear, happy, sad, surprise and neutral. The images in FER2013 are gray-scale with size 48×48. We use a subset of 25,887 images from FER2013 as the training set and the remained 1000 images as the validation set.

Since the facial emotion dataset is quite small, we use two different strategies to train our CNN models. The first is to finetune a sophisticated CNN network pre-trained on a large facial dataset. We use VGG-Face [27] as our basic CNN model, which is based on the VGG-Very-Deep-16 CNN architecture [27] and trained for the face recognition task. We only finetune the last three layers in the task. And the activations from the antepenultimate, penultimate and the last softmax layers are extracted as our VGG-Face CNN features, called `VGG_fc6`, `VGG_fc7` and `VGG_prob` respectively.

As the another strategy, we build a small CNN model from scratch. The CNN architecture is shown in Figure 2. Data augmentation and dropout are used to prevent overfitting. We extract the outputs of `pool4`, `fc5`, `fc6` and the last `prob` layer as the image feature respectively. In order to better transfer the CNN model on the AFEW dataset, the network is further finetuned through active learning. In each epoch, we select correctly classified facial images in the videos using the CNN to finetune the CNN model. We finetune the model for 10 epochs with decreased learning rate. The features from the finetuned CNN is called `ft_pool4`, `ft_fc5`, `ft_fc6`, `ft_prob` respectively.

For all image-level CNN features except `*prob`, mean pooling is applied over the videos to generate video-level features. For `*prob` features, we calculate the mean and standard deviation to better represent the temporal dynamics of the facial expression in the video.

2.3 Motion Related Features

To exploit the temporal information in the video, we extract the improved Dense Trajectory feature (iDT) [28], which is proved to be effective in many video classification tasks. The main purpose of `iDT` features is to represent the dynamic movements of the human poses and facial muscles changes. HOG, HOF and MBH features are extracted from the video using the dense sampling strategy and are encoded with Fisher Vector Encoding. The GMM codebook consists of 32 mixtures which leads the dimensionality of the `iDT` fea-

Table 2: The accuracy of acoustic features on validation set (%)

Feature	Dimension	SVM	RF	LR
IS09	384	33.94	35.77	30.03
IS10	1582	31.59	**36.03**	26.63
mfccBoAW	1024	33.16	34.20	28.20
mfccFV	624	**35.51**	35.51	**34.99**

Table 3: The accuracy of facial image features on validation set (%)

Feature	Dimension	SVM	RF	LR
LBP-TOP	2832	39.69	34.20	40.99
DSIFT	21504	37.86	32.64	40.73
VGG_fc6	4096	36.55	36.03	34.73
VGG_fc7	4096	34.46	35.77	36.55
VGG_prob	14	36.29	35.77	33.94
pool4	4608	36.03	38.90	37.60
fc5	256	39.95	40.47	40.73
fc6	256	**40.21**	38.38	38.64
prob	14	38.90	38.38	38.64
ft_pool4	4608	38.90	39.16	40.47
ft_fc5	256	39.43	**42.30**	41.51
ft_fc6	256	38.69	40.21	41.25
ft_prob	14	38.12	39.95	**42.30**

tures to be 25,344. We use the entire areas of the video and only the facial parts of the video respectively to extract iDT features, called the video_iDT and face_iDT respectively.

3. EXPERIMENTS

3.1 Experimental Setup

Following standard data split, we randomly select 20% of the training data for local validation, which is to say that we use 616 videos of the original training data as the local training set and the remained 157 videos as the local validation set. We report experiment results on the original validation set of 383 videos in the following experiments. SVM [29], random forest (RF) [30] and logistic regression (LR) are used as our unimodal classifiers. Hyper parameters of the models are selected according to the accuracy on our local validation set using grid search. For SVM, RBF kernel is applied and the cost is searched from 2^{-3} to 2^{15}. And for random forest, the number of trees is selected from 100 to 1000 with step 100 and the depth of the tree is searched from 2 to 20. For logistic regression, we use the one-vs-rest scheme to train multiple binary classifiers without any hyper parameters.

3.2 Unimodal Results

The performance of the different acoustic features is presented in Table 2. As shown in Table 2, the Fisher Vector Encoding is superior to Bag-of-Audio-Words, which makes the mfccFV feature to be the most robust and discriminative features among all acoustic features.

Comparing the second and third block in Table 3, we can see that the features extracted from the small CNN model are more effective than those of finetuned VGG-Face CNN model. The reason for the relative low performance of finetuned VGG-Face model might be the variant size and channels between VGG-Face dataset and FER2013 facial expression dataset.

Also as shown in Table 3, the features extracted from the small CNN model significantly surpass the hand-crafted features. Furthermore, after finetuning the small CNN model using the facial images from AFEW videos through active learning, the performance of the CNN features is improved and both ft_fc5 and ft_prob achieve 42.30% accuracy on the validation set. The fc5 layer feature achieves the best performance among the CNN features in average.

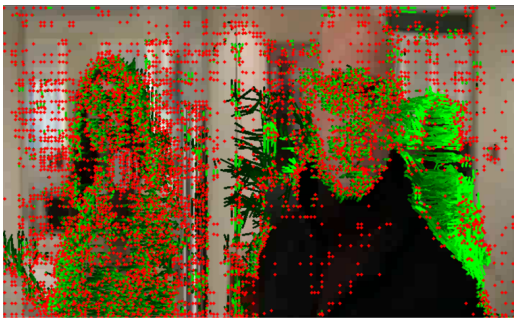
The motion features of the entire video video_iDT achieved a poor performance with validation accuracy of 27.68% using logistic regression as shown in Table 4. We visualize the dense trajectories in Figure 3. Besides motions of the human body and facial muscles, the video_iDT also capture a lot for the background movements, which add much noise for classification. When we only extract the motion feature on the facial area, the performance is improved to 30.03% but the motion modality still is worse than acoustic and facial image modality. In general, visual modality performs the best and the acoustic modality performs the second best.

Confusion matrixes on the validation set of audio feature mfccFV, facial image feature ft_fc5 and motion feature face_idt are presented in Figure 4 (a)(b)(c) respectively. The acoustic feature can distinguish the angry emotion best

Table 4: The accuracy of motion features on validation set (%)

Feature	SVM	RF	LR
video_iDT	23.24	25.59	27.68
face_iDT	27.94	25.33	30.03

(a) video clip



(b) improved dense trajectory

Figure 3: The visualization of improved Dense Trajectory. The red dots is the dense samplers and the green line is the trajectories.

and the CNN feature can recognize the happy emotion best. The `face_idt` is slightly more distinguishable for neutral emotion since there are less facial movements in neutral expressions. The different classification effects show that different kinds of modality features are complimentary.

3.3 Multimodal Fusion Results

Since different features are complementary for emotion recognition, we explore late fusion strategies on different combinations to fully make use of all the multimodal features since the early fusion can suffer from the high dimensionality of the features. The logistic regression is used as the second level classifier in order to avoid overfitting. The probabilities of the unimodal classifiers are concatenated as the late fusion features. Table 5 presents the performances of late fusion of different modalities on both the validation set and the testing set.

We explored training the late fusion model on our local validation set only or on the combination of local validation and original validation sets. The performance on the testing set shows that only using local validation can achieve better performance. This might because that the data distribution of the testing set and the validation set is different. By late fusing different modality features, we achieve the testing accuracy of 50.76% which is 10.29% higher than the baseline accuracy of 40.47%. The confusion matrix of our best submission on the validation and testing set is shown in Figure 5. The multimodal fusion system has a fairly high classification performance for angry and happy emotion classes, which can be classified best by audio and facial expression respectively.

4. CONCLUSIONS

In this paper, we investigate various multimodal features from audio, face image and video motion modalities and late fusion strategies for the EmotiW 2016 audio-visual emotion recognition challenge. Among all the extracted unimodal features, the facial CNN feature is the most discriminative feature for emotion recognition. And the MFCC fisher encoding is the most effective acoustic feature. The late fusion of different modality features significantly boost the recognition performance. In the future, we will continue to extract more discriminative features including learned feature for audio modality and 3-D CNN features for motion modality and utilize some personalized fusion methods. Also, we will explore modelling the temporal information for video emotion recognition.

5. ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202.

REFERENCES

[1] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[2] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey,

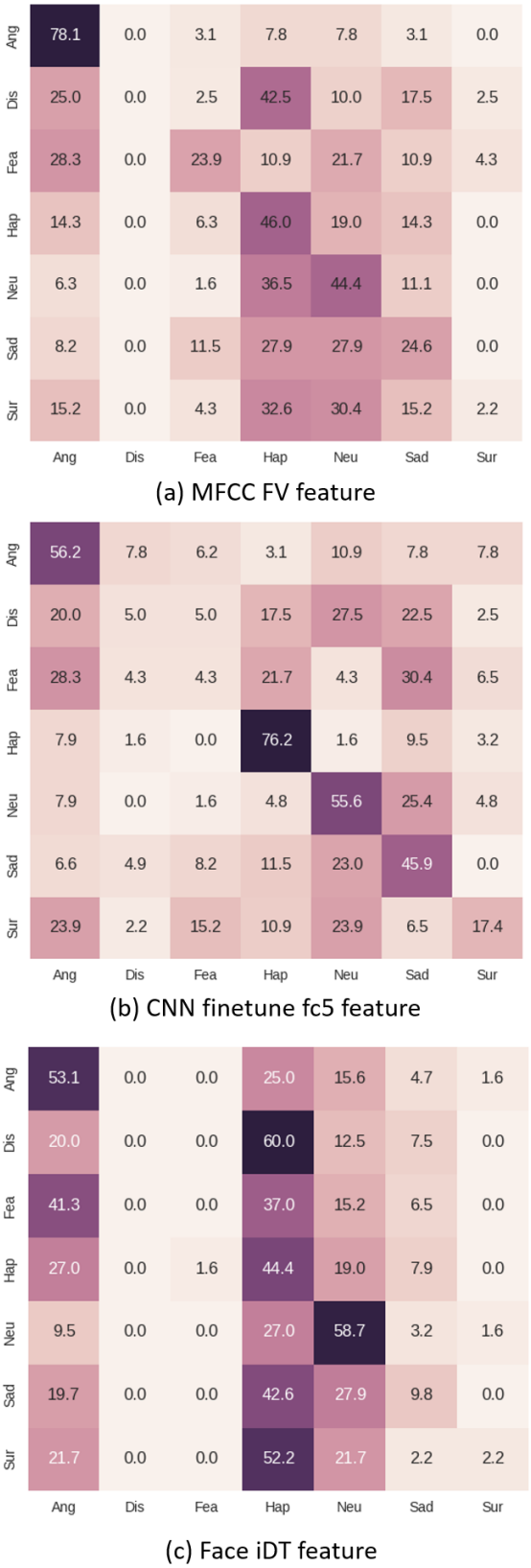


Figure 4: The confusion matrix of unimodal features

Table 5: Late fusion accuracies on the validation and testing sets

feature	train set	submission	Val	Test
IS09+mfccFV+ft_fc6+ft_pool4+fc5+LBP-TOP+DSIFT	val local	1	50.65	**50.76**
	val local + val	2	55.61	48.74
IS09+mfccFV+mfccBoAW+ft_prob+ft_fc6+ft_pool4+fc5+DSIFT	val local	3	51.44	50.59
	val local + val	4	55.87	49.58

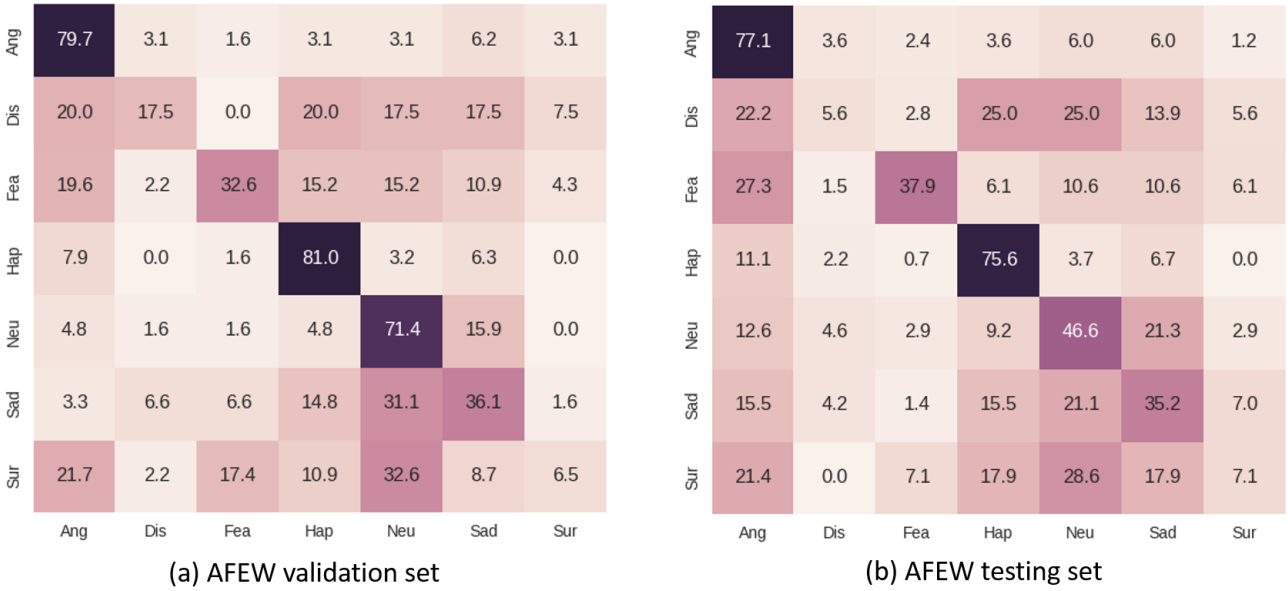


Figure 5: Confusion matrix of the best submission on the validation and testing sets

and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. ACM ICMI, 2016.

[3] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 451–458. ACM, 2015.

[4] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Multiple models fusion for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 475–481. ACM, 2015.

[5] Bo Sun, Liandong Li, Guoyan Zhou, Xuewen Wu, Jun He, Lejun Yu, Dongxue Li, and Qinglan Wei. Combining multimodal features within a fusion network for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 2015*, pages 497–502, 2015.

[6] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 2015*, pages 435–442, 2015.

[7] Albert C. Cruz. Quantification of cinematography semiotics for video-based facial emotion recognition in the emotiw 2015 grand challenge. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 2015*, pages 511–518, 2015.

[8] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Reddy Konda, Roland Memisevic, and Christopher Joseph Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 2015*, pages 467–474, 2015.

[9] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3:e12, 2014.

[10] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. Emotion recognition using acoustic and lexical features. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 366–369, 2012.

[11] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *The International Workshop on Audio/visual Emotion Challenge*, 2015.

[12] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turke, 2014*, pages 508–513, 2014.

[13] Florian Eyben, Martin Llmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia, Mm*, pages 1459–1462, 2010.

[14] Björn W. Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 312–315, 2009.

[15] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011.

[16] Steven B. Davis. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in Speech Recognition*, 28(4):65–74, 1990.

[17] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1370–1374, 2014.

[18] Jorge SÃąnchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[19] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

[20] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[21] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[22] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using volume local binary patterns. In *Dynamical Vision, ICCV 2005 and ECCV 2006 Workshops, WDV 2005 and WDV 2006, Beijing, China, October 21, 2005, Graz, Austria, May 13, 2006. Revised Papers*, pages 165–177, 2006.

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[24] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[25] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[26] Yichuan Tang. Deep learning using support vector machines. *CoRR, abs/1306.0239*, 2, 2013.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[28] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[29] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[30] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.