Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network

Jian Huang
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
jian.huang@nlpr.ia.ac.cn

Ya Li
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
yli@nlpr.ia.ac.cn

Jianhua Tao
National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation Chinese Academy of Sciences, Beijing, China
jhtao@nlpr.ia.ac.cn

Zheng Lian
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
lianzheng2016@ia.ac.cn

Zhengqi Wen
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
zqwen@nlpr.ia.ac.cn

Minghao Yang
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
mhyang@nlpr.ia.ac.cn

Jiangyan Yi
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
jiangyan.yi@nlpr.ia.ac.cn

ABSTRACT

The continuous dimensional emotion can depict subtlety and complexity of emotional change, which is an inherently challenging problem with growing attention. This paper presents our automatic prediction of dimensional emotional state for Audio-Visual Emotion Challenge (AVEC 2017), which uses multi-features and fusion across all available modalities. Besides the baseline features provided by the organizers, we also extract other acoustic audio feature sets, appearance features and deep visual features as complementary features. Each type of feature is trained using Long Short-Term Memory Recurrent Neutral Network (LSTM-RNN) for every dimensional emotion prediction separately considering annotation delay and temporal pooling. To overcome overfitting problem, robust models are chosen carefully for individual model. Finally, multimodal emotion fusion is achieved by utilizing Support Vector Regression (SVR) with the estimates from different feature sets in decision level fusion. The experimental results indicate that our extracted features are beneficial to performance improvement and our system design achieves very promising results with Concordant Correlation Coefficient (CCC), which outperform the baseline system on the testing set for arousal of 0.599 vs 0.375 (baseline) and for valence of 0.721 vs 0.466 and for liking 0.295 vs 0.246.

AVEC'17, October 23, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5502-5/17/10…$15.00
https://doi.org/10.1145/3133944.3133946

Keywords

Dimensional Emotion Recognition; LSTM-RNN; Delay; Temporal pooling; Overfitting; Multimodal fusion

1 INTRODUCTION

Due to the essential role of affective computing in artificial intelligence, emotion recognition has gained increasingly intensive attention [1], especially for the recognition of non-acted spontaneous emotions in the continuous dimensional space. The past emotion recognition focused on laboratory settings, which is difficult to be applied in real world situations effectively [2]. Moreover, there is a shift from discrete emotion model to dimensional emotion model which describes emotion in a continuous multi-dimensional emotion space [3][4] and thus can model subtle and complicated emotional behaviors.

The Audio-Visual Emotion Challenge (AVEC) [5-11], an annual challenge since 2011, aims at promoting the development of multimedia processing and machine learning methods for automatic continuous emotion recognition in the wild. It provides a framework for non-acted spontaneous emotion recognitions and a fair benchmark to evaluate various emotion recognition methods. The first AVEC [5] simplifies the continuous dimensional emotion prediction as a classification

task, while the following challenges regard it as a regression task to be corresponding to natural emotional state. The depression recognition sub-challenge is added since AVEC 2013 [7], which is related to emotion recognition well. The physiological modality is added in AVEC 2015 [9] and AVEC 2016 [10], but doesn't appear in AVEC 2017 [11]. AVEC 2017 introduces text modality in the first time. For the dimensional modeling of emotion, arousal and valence dimension are widely used in most of AVEC challenges. Besides these two dimensions, AVEC 2017 introduces prediction of likability indicating the participants' tastes for the video/audio, which are explained in Fig. 1.

Audio modality plays a key role in emotion recognition especially in arousal dimension. For the audio signals, AVEC organizers provide the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [13] as baseline features. The acoustic feature sets are main audio features for emotion recognition in most of researches. Bottle-neck (BN) features are firstly proposed to improve the performance of speech recognition [14] and then found to be effective in language identification [15] and speaker identification [16]. BN features are generated from a narrow hidden layer of Deep Neural Network (DNN), which are adopted for emotion recognition in AVEC 2015 [17] and AVEC 2016 [18]. Filip et al. [17][18] adopt two Neural Networks (NNs) structure, the output of the first network is stacked in time, defining context-dependent input features for the second NN. Two stacked BN features are used as acoustic feature set trained as French and several languages, which achieve promising performance for emotion recognition.

For video signals, AVEC organizers provide geometric features related to the position and expression of the subjects' face. Empirically, facial expressions have an important influence on emotion recognition especially in valence dimension. Bo et al. [19] explore various visual features for continuous emotion recognition, including Local Binary Gabor Patterns (LGBP) [20], geometric features, multi-scale dense SIFT and deep visual features. They utilize Convolutional neutral networks (CNNs), AlexNet and ResNet pre-trained on other dataset and fin-tuned on the AVEC dataset to extract deep visual features, which have positive effect on emotion recognition.

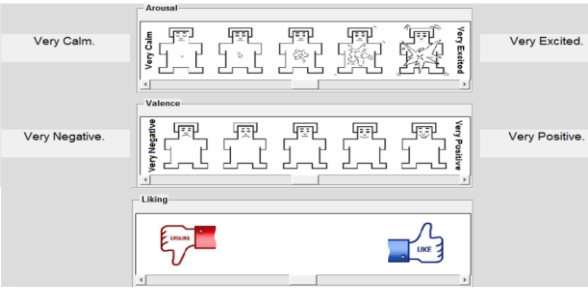


Figure 1: The descriptions of arousal, valence and liking dimensions [12]

People can perceive emotion straightway from hearing and vision. In addition, the semantic of the words can provide helpful and important information in emotion recognition. For example, some special words, laughter and sob can reflect the current emotion state of the person indeed. Two main methods are applied to take advantage of text modality, lexicon-based approach and word embedding. The lexicon-based approach obtains semantic information from the lexicon of emotional words to estimate emotion. Word embedding maps the words to real number vectors in a lower dimensional space. Due to the lack of text modality data of AVEC 2016, Filip et al. [18] use an automatic speech recognition system to get text transcriptions. They investigate the effectiveness of text based features using lexicon-based approach and word embedding, and the results indicate that word embedding can improve the performance of arousal dimension. In AVEC 2017, organizers provide a bag-of-words text feature representation based on the transcription of the speech as text features. The baseline experimental results [11] show that text features achieve best and robust performance in liking dimension.

With respect to the continuous emotion recognition, various regression models have been used. Two regression models are used frequently; one is static regression models represented by Kernel based SVR and the other introduces recurrent networks such as Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs). SVR can achieve satisfactory performance and is chosen as the baseline method in AVEC challenges [9][10]. However, dimensional emotion is continuous temporal process related closely to contextual information. Therefore, successful application of temporal information has a critical impact on performance improvement of emotion recognition. Wöllmer et al. [21] utilize LSTM-RNN to perform regression analysis on arousal and valence dimension, which improves recognition performance significantly. Later, Wöllmer and Nicolaou et al. [22] further improve the LSTM architectures to bidirectional LSTM (BLSTM). Compared to SVR, LSTM can capture the temporal information of the emotional dimension better, which is also adopted by the first place winner [23] and second place winner [24] of AVEC 2015.

Benefited from the complementarity of different modalities, multimodal emotion fusion can achieve significant recognition improvements and provide the robustness when feature extraction fails. Feature level fusion and decision level fusion strategies are widely utilized [3]. Feature level fusion extracts features from every modality separately and then concatenates them into feature vector for final emotion recognition. Feature level fusion suffers from the curse of dimensionality and demands a strict time synchrony between the modalities. Decision level fusion assumes each modality is independent which eliminates some disadvantages of feature level fusion. It builds separate emotion recognition models and combines the predictions of different modalities to train a second level model. The baseline systems [10][11] and Bo et al. [19] utilize multi-modal decision level fusion to improve system performance. Chao et al. [24] adopts LSTM for decision level fusion to achieve better performance than feature level fusion. Therefore, decision level fusion is considered in this challenge.

In the following, Section 2 briefly introduces the database. Section 3 presents feature sets adopted in this challenge. Section 4 describes the regression model. Section 5 show details of the entire experiment and results. Section 6 concludes this paper.

2 DATABASE

AVEC 2017 is based on Sentiment Analysis in the Wild (SEWA). This database collects spontaneous and naturalistic interactions consisting of audio, video and text modalities. All recordings were recorded 'in the wild' through human-human interactions, but only the behaviors of one person are in every recording. Thus, the recording of audio can record the sound of another interlocutor, which would influence the effectiveness of features. Speaker turn timings are derived to know which subject is specking when. We add the turn timings information to feature sets to decrease this influence. The duration time of the recording ranges from 40 seconds to 3 minutes, which is different from former challenges having similar duration time for all the recordings. This dataset is annotated in three dimensions, namely arousal, valence, liking. The results are evaluated using CCC, which combines the Pearson correlation coefficient of two times series with mean square error.

3 MULTI-MODAL EMOTION FEATURES

3.1 Audio Features

3.1.1 Baseline Audio Features. This challenge adopts eGeMAPS as the baseline audio features. Both segment-level acoustic feature types are computed over segments of 4/6 seconds. Overall, the acoustic baseline feature sets contain 88 dimensional features. The extraction of the LLDs and the computation of the functionals are done using the openSMILE toolkit [25].

3.1.2 Other Audio Feature Sets. Due to the importance of audio modality in emotion recognition, we extract other acoustic feature sets to complement the baseline audio features. Specifically, we extract the baseline audio feature set of INTERSPEECH 2010 Paralinguistic Challenge (IS10) [26] which can better reflect a border coverage of paralinguistic information assessment. It includes 38 LLDs computed by 21 functionals resulting in 1582 acoustic features. Finally, we apply a Principal Component Analysis (PCA) to retain 418 dimensional features.

In addition, MFCC features have wide application and excellent performance in the field of speech recognition. The winner of AVEC 2016 [27] extracts MFCC features as one type of audio features which achieve better performance than other audio features in arousal and valence dimension. Therefore, we compute MFCC features using mel-scale filterbank to extract 39 dimensional features including 13 dimensional MFCC along with their delta and acceleration coefficients.

3.1.3 BN Features. Filip et al. [17][18] have verified the effectiveness of BN features. We extract high level BN features from bottleneck DNN network, which is different from Filip's two Neural Networks (NNs) structure. The bottleneck DNN network, designed for speech recognition, has six hidden layers and each layer has 1024 nodes except 60 nodes of last layers. We utilize 300 hours spontaneous English speech recognition corpus to train this network. The trained deep speech recognition network acts as DNN based feature extractor.

3.2 Video Features

3.2.1 Baseline Video features. AVEC organizers provide geometric features related to the position and expression of the subjects' face including face orientation, eye points and facial landmarks. We handle origin features to obtain robust and efficient features [9]. The 49 facial landmarks are aligned with a mean shape from stable points (located on the eye corners and on the nose region). Then, we compute the difference between the coordinates of the aligned landmarks and those from the mean shape, and also between the aligned landmark locations in the previous and the current frame. The same operations are applied for face orientation and eye points. Then, the facial landmarks are split into three regions: i) the left eye and left eyebrow, ii) the right eye and right eyebrow and iii) the mouth. For each of these groups, the Euclidean distances (L2-norm) and the angles (in radians) between the points are computed. We also computed the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. The geometric feature sets are interpolated by a piecewise cubic Hermite polynomial to cope with dropped frames. In a results, geometric features have 372 dimensional features.

3.2.2 Appearance Features. Facial expressions are usually quantified by two types of facial descriptors: appearance and geometric features. We extract Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) and Histogram of Oriented Gradients (HOG) as the appearance features. LGBP-TOP features are extracted based on blocks of frames to capture dynamic information effectively. Firstly, we detect face pictures with frame step 20ms from video stream. The failed frames are replaced with neighboring successful frame. Then, we follow the operating steps [9] to extract LGBP-TOP features. HOG [28] features describe the distribution information of intensity gradients or edge directions. The descriptor decomposes a local region into small squared cells, computes the histogram of different bins of oriented gradients in each cell, and normalizes the results using block-wise pattern (each block contains several cells). Finally, a feature reduction is performed by applying a PCA for LGBP-TOP keeping 500 dimensional features and HOG features keeping 40 dimensional features.

3.2.3 Deep Visual Features. As previously mentioned, Bo et al. [19] verify the effectiveness of deep visual features based on AlexNet and ResNet for continuous emotion recognition. Therefore, we extract deep visual features based on AlexNet. The effective Deep CNN model is trained using 110,000 face images from 1032 people in all, inspired by Liu's work [29]. The trained CNN model acts as CNN based feature extractor. The detected face images are fed into trained CNN network to extract the 9216-dimensional features from the 5th pooling layer. Finally, 500 dimensional features are kept based on the importance of the features.

3.3 Text Features

AVEC 2017 firstly introduces text modality features, which are bag-of-words feature representation. The dictionary for these textual features is learnt from the training partition taking only

the terms with at least two occurrences into account. The generated dictionary consists of 521 words, where only unigrams are considered. Finally, bag-of-text-words (BoTW) features are extracted by openXBOW [30]. In this challenge, text features are used to improve the recognition performance.

4 MULTI-MODAL EMOTION REGRESSION MODEL

For this challenge, we adopt the LSTM-RNN based neutral network as basic emotion regression model. The overview of system framework is shown in Figure 2(a). Different types of features from audio, video and text modalities described in detail in section 3 are extracted respectively. Then, each feature set is trained individually to get its best model. The concatenated estimates from different modalities are utilized to obtain final emotion predictions using SVR. Figure 3(b) presents specific procedure of model training. The turn timings information is added to feature sets as mentioned in section 2. Then, the features are fed into LSTM to train emotion regression model by taking the factor of annotation delay and temporal pooling into consideration.

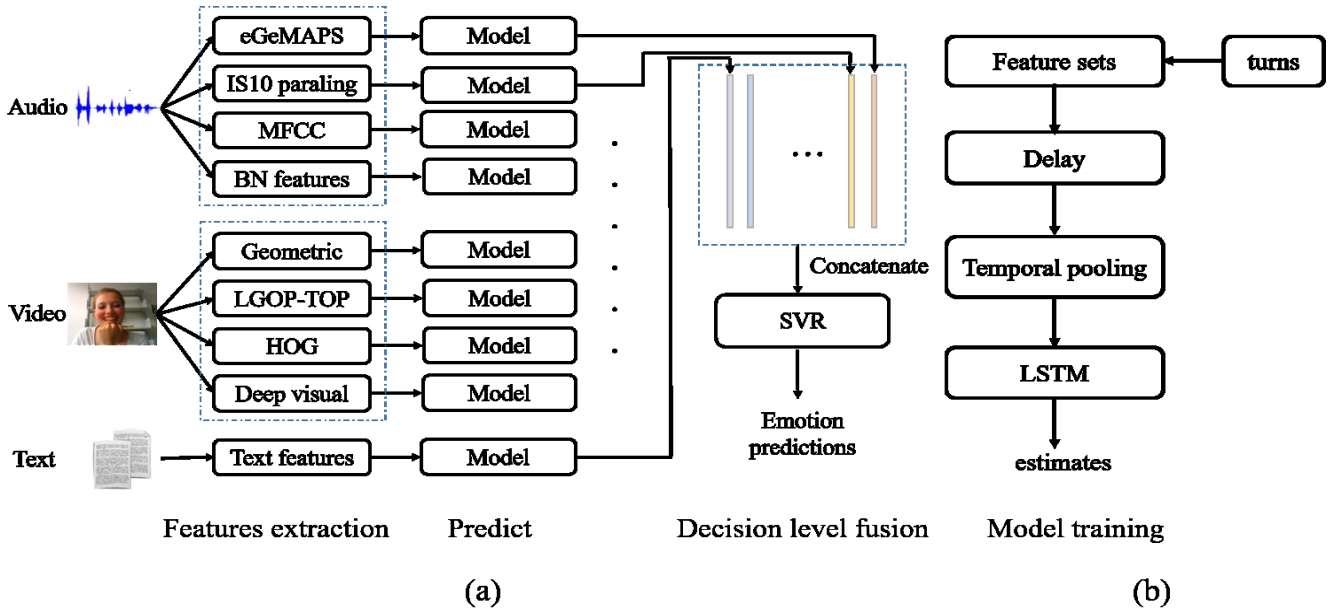


Figure 2: (a) Overview of the proposed multimodal emotion regression method. The features from different modalities are extracted in feature extraction. Each feature set is trained individually based on LSTM model. The estimates from different modalities are concatenated to obtain final emotion predictions using SVR in decision level fusion. (b) Modeling training. The feature sets combined with turns time information are fed into LSTM to train emotion regression model considering the factor of delay and temporal pooling.

4.1 Annotation Delay

Continuous dimensional emotion accords with nature emotional dynamic, while time-continuous annotations bring problems about the reliability of the labels. When the annotators label the recordings, a serious problem is their reaction lag caused by observing, appraising and responding to the expressive behaviors [31]. Soroosh et al. [31] propose to compensate for this reaction lag by finding the time-shifting that maximizes the mutual information between the expressive behaviors and the time-continuous annotations. The factor of annotation delay is processed in many emotion recognition systems of AVEC [11][23][32] to improve the prediction performance. Therefore, we consider the influence of annotation delay in our emotion regression model. This is achieved by shifting all annotations forward in time before training a model. The duration time of delay is regarded as a parameter to be optimized when training LSTM model. For each feature set, we find its proper duration time of delay adapted to LSTM training. The parameter delay ranges from 0 s to 2 s with a step of 0.2 s.

4.2 Temporal Pooling

There exists inevitable label noise owing to continuous long time annotation. Michel et al. [8] calculate the average ratings from all raters to minimize the label noise. In AVEC 2017 [11], Hermitian resampling and EWE approach are performed to create one unique gold standard from the annotations and decrease the noisy of the labels. In SEWA, every frame (100ms) is labeled one dimensional value and the maximum frames is 1756. Therefore, there also exists redundant information among adjacent frames. Ringeval et al. [32] add average window to the labels and features to smooth the labels and decrease the noise. Chao et al. utilize temporal pooling function in the forward network for deep belief network [33] and LSTM [24] to reduce

the label noise and redundant information. The temporal pooling operation add the window to average both the features and labels, which can get the statics of the successive frames to achieve short level temporal modeling. On the other, temporal pooling of the features is also regarded as sub-sampling. The labels are also averaged in the same size of window to decrease the noisy to some extent. Thus, we utilize the temporal pooling, specifically mean pooling same as Chao's work [24]. The window length of temporal pooling is also regarded as a parameter to be optimized when training LSTM model.

4.3 LSTM-RNN Model

LSTM-RNN can learn long-term dynamic information since the output of LSTM layer is influenced by the outputs of hidden layer involving previous information and the current input. Emotion is a temporally expression event which can be better inferred by LSTM network structure. Each feature set is inputted to LSTM to train individual emotion regression model. The turn timings information is added to the feature sets to decrease the noise of the recordings. We quantize the turn timings information by setting the time range when the recorder is talking to 1 and the time range when another interlocutor is talking to 0. We use ε-insensitive loss function to ignore small errors and assign absolute value loss to large errors, which is more suitable than the other loss functions [24]. For different feature sets, the architectures of the network including one LSTM layer keeps same except the number of nodes in the input layer. Adadelta [34] optimization algorithm is utilized. Weight decay in the linear regression layer is also applied to prevent over-fitting. The hyper-parameters, the number of hidden layer nodes along with the delay time and the window length of temporal pooling, are chosen based on the CCC performance by random combination in the development set. The maximum training epochs are 70. We also use dropout after LSTM with the rate 0.5.

4.4 Decision Level Fusion

Having finished model training of every feature set, the estimates from different features are concatenated to train further regression model. Three conventional regression algorithms, logistic regression, random forest and SVR are utilized in this study. In random forest, the number of trees varies from range [50, 500] and the maximum depth of the tree varies from range [5, 10]. In SVM, the parameter C varies from range [1, 100] and gamma varies from range [0.01, 0.5]. The experimental results indicate that SVR with RBF kernel can achieve best performance. Therefore, we utilize SVR with RBF kernel for decision level fusion. The hyper-parameters of SVR are chosen based on the CCC performance in the development set. Given space limitations, we only present the experimental results of SVR in the following.

5 EXPERIMENTS AND ANALYSIS

Limited to the size of SEWA database, the LSTM model is hard to be trained adequately, easy to fall into overfitting. The experimental results reveal that the CCC performance of training can achieve near 0.9, but it degrades seriously on development and testing set, which will be shown in Fig. 3. To solve this problem, we split the development set into two subsets: the first 9 subjects are regarded as *dev1* set to adjust the parameters and the left 5 subjects are regarded as *dev2* set to test the model. We choose the model that achieves better and close performance on both *dev1* and *dev2* set. The predicted performance is reported in terms of CCC with the root mean squared error (RMSE) and Pearson's correlation coefficient (PCC).

5.1 Unimodal Emotion Prediction

For every dimension, different feature sets are trained separately to get best performance through the settings described in section 4.3. There are nine feature sets in all, four feature sets (eGeMAPS, IS10, MFCC and BN) for audio modality, four feature set (Geometric, LGBP-TOP, HOG and Deep visual) for video modality and one feature set (BoTW) for text modality. The experimental results on *dev2* set are shown in Table 1, and the corresponding results of *dev1* set are in brackets. The symbol "-" represents that the feature sets are abandoned because of worse performance.

In the following, we assess the feature sets based on the CCC performance of *dev2* set. Deep visual features achieve best perform 0.682 in arousal dimension, while geometric features achieve best performance 0.740 in valence dimension. Thus, video modality achieves better performance than audio modality both in arousal and valence dimension, which is different from previous statements that audio modality can achieve better performance in arousal dimension and video modality can achieve better performance in valence dimension. However, only BoTW features can perform well on both *dev1* (0.444) and *dev2* set (0.473) for liking dimension. LGBP-TOP features achieve promising performance on *dev1* set (0.581) but generalize badly on *dev2* set (0.236). As a whole, the performance of arousal and valence dimension are much higher than that of liking dimension. The performance of valence dimension is better than that of arousal dimension, which is also contrary to previous AVEC challenges.

The delay of different feature sets is recorded in Table 2. We can observe that arousal and valence dimension have shorter delay less than 1 second and liking dimension has longer delay more than 1 second except the IS10. It indicates that the annotators can react quickly to arousal and valence dimension when perceiving others' emotional state, but slowly to liking dimension when expressing their preferences for the recordings. Besides, the duration of delay time is shorter than the experimental results of baseline system. An explain could be that LSTM network capturing the temporal information has alleviated some influence of annotation delay.

Table 1: Performance comparisons with the proposed emotion regression model and different feature sets for the AVEC 2017 *dev2* set (*dev1* set).

	Arousal			Valence			Liking		
	RMSE	PCC	CCC	RMSE	PCC	CCC	RMSE	PCC	CCC
eGeMAPS(88)	0.123 (0.144)	0.522 (0.509)	0.506 (0.485)	0.165 (0.167)	0.500 (0.484)	0.455 (0.478)	0.133 (0.133)	0.203 (0.468)	0.193 (0.426)
IS10(418)	0.119 (0.125)	0.502 (0.655)	0.465 (0.631)	0.131 (0.139)	0.442 (0.540)	0.440 (0.536)	0.103 (0.126)	0.229 (0.427)	0.227 (0.373)
BN(60)	0.107 (0.128)	0.543 (0.627)	0.533 (0.604)	0.128 (0.122)	0.485 (0.647)	0.466 (0.643)	-	-	-
MFCC(39)	0.139 (0.165)	0.356 (0.406)	0.341 (0.400)	0.127 (0.142)	0.425 (0.468)	0.421 (0.453)	-	-	-
Geometric(372)	0.092 (0.117)	0.645 (0.691)	0.639 (0.662)	0.087 (0.109)	0.742 (0.718)	**0.740 (0.713)**	0.110 (0.144)	0.169 (0.327)	0.166 (0.310)
LGBP-TOP(500)	0.108 (0.112)	0.647 (0.724)	0.604 (0.708)	0.100 (0.114)	0.696 (0.691)	0.695 (0.686)	0.139 (0.117)	0.284 (0.588)	0.236 (0.581)
HOG(40)	0.100 (0.130)	0.649 (0.615)	0.602 (0.591)	0.109 (0.141)	0.597 (0.544)	0.590 (0.526)	0.134 (0.135)	0.165 (0.354)	0.153 (0.329)
Deep visual(500)	0.093 (0.117)	0.703 (0.699)	**0.682 (0.673)**	0.090 (0.103)	0.724 (0.738)	0.720 (0.720)	0.096 (0.147)	0.314 (0.374)	0.302 (0.351)
BoTW(521)	0.112 (0.135)	0.463 (0.558)	0.451 (0.503)	0.125 (0.135)	0.518 (0.539)	0.518 (0.526)	0.090 (0.124)	0.478 (0.472)	**0.473 (0.444)**

5.2 Multi-modal Emotion Prediction

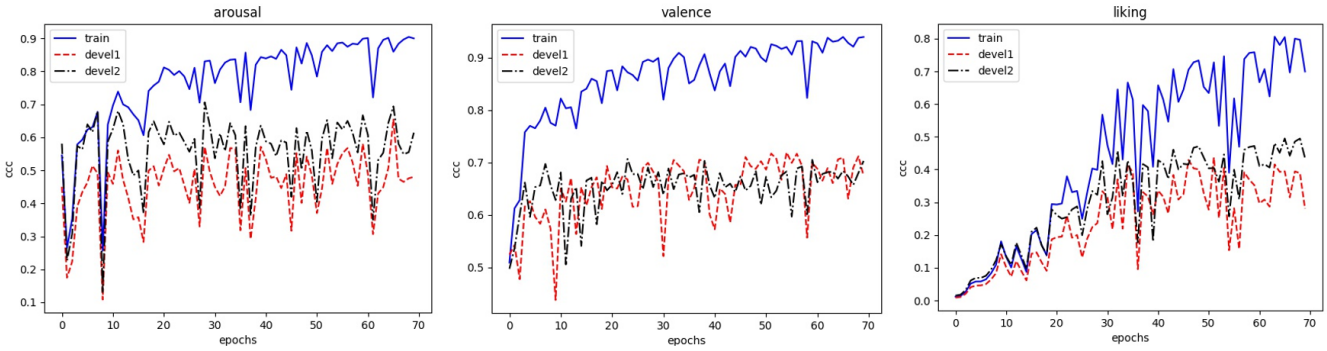
For each feature set, we need to obtain robust model respectively. To be specific, the chosen model should perform well on both *dev1* and *dev2* set, meanwhile avoid overfitting. In reality, there is high chance that overfitting would exist if the number of the training epochs is set too large. We can observe from Fig. 3(b) in valence dimension that the performance of *dev1* and *dev2* set have reached a stable state in near 10 epochs, but the performance of training set is still increasing after 10 epochs.

Therefore, the final model near 70 epochs is overfitting. The same situation also exists in Fig. 3(a) for arousal dimension and in Fig. 3(c) for liking dimension. To overcome this problem, the models are chosen once the performance of *dev1* and *dev2* set reach a stable state. For example, we choose the model near 10 epochs in Fig. 3(b) rather than the model near 70 epochs. We also notice that the training of valence dimension is stable, but that of arousal and liking dimension have serious fluctuation, which influences their robustness.

Table 2: The delay of different feature sets in three

Delay(s)	Arousal	Valence	Liking
eGeMAPS	0.4	0.6	1.8
IS10	1.0	1.6	0.4
BN	0.0	0.4	-
MFCC	0.8	0.0	-
Geometric	0.2	0.2	1.2
LGBP-TOP	0.8	0.8	1.6
HOG	0.2	0.2	1.2
Deep visual	0.0	0.2	1.0
BoTW	0.8	1.0	1.4

After obtaining appropriate model of every feature set, their estimates are combined by frame-wise concatenation. Then, SVR is utilized for decision level fusion. The model of SVR are trained on training set, optimized on dev1 set and tested on dev2 set. In order to select an optimal combination of feature sets, a greedy feature selection strategy is utilized. Firstly, we rank the feature sets according to their CCC performance of dev2 set. Then, the feature sets are added sequentially to the combination in order, and the feature set is retained if the performance of dev2 set increases, otherwise abandoned. The combination with the highest value is selected. Finally, the post-processing median-filtering [9] is applied to smooth the prediction with the window size optimized on *dev1* set. Prediction results on *dev1* and *dev2* set, with different combinations are shown in Table 3 for arousal dimension, Table 4 for valence dimension, Table 5 for liking dimension. We can observe that the feature selection strategy is effective because the selected feature combinations is better than the combinations including all available feature sets for three dimensions. The optimal combination of arousal dimension is deep visual features, geometric features, LGBP-TOP, HOG and eGeMAPS. The optimal combination of valence dimension is geometric features, deep visual features, HOG, IS10, eGeMAPS and BoTW. The optimal combination of liking dimension is BoTW, deep visual features, LGBP-TOP, IS10 and geometric features.



(a) (b) (c)

Figure 3: (a) The CCC performance over training epochs for arousal dimension. (b) The CCC performance over training epochs for valence dimension. (c) The CCC performance over training epochs for liking dimension.

Multimodal fusion improves the performance obviously compared to unimodal prediction. For valence dimension, the performance of *dev1* and *dev2* set are both increasing when adding the new features. It's worth noting that the performance of *dev1* set might decrease, contrary to *dev2* set when adding the new features for arousal and liking dimension. It indicates that multimodal fusion is stable and effective for valence dimension, but a little fluctuant for arousal and liking dimension. The final results for AVEC 2017 development and testing set are shown in Table 6 including baseline results. Our proposed method achieves promising results. The system results with Concordant Correlation Coefficient (CCC), outperform the baseline system on the testing set for arousal of 0.599 vs 0.375 (baseline) and for valence of 0.721 vs 0.466 and for liking 0.295 vs 0.246.

Table 3: Multi-modal prediction results of arousal dimension on *dev2* set (*dev1* set).

	RMSE	PCC	CCC
A: Deep visual+ Geometric	0.082 (0.110)	0.725 (0.742)	0.708 (0.697)
B: A+ LGBP-TOP	0.082 (0.110)	0.732 (0.736)	0.709 (0.690)
C: B+ HOG	0.072 (0.117)	0.763 (0.722)	0.750 (0.635)
D: C+ eGeMAPS	0.071 (0.112)	0.773 (0.756)	0.762 (0.649)
All	0.072 (0.114)	0.766 (0.735)	0.754 (0.633)

6 CONCLUSIONS

This paper presents our approach for AVEC 2017. Besides the baseline features, we extract extra feature sets as additional features, IS10, MFCC features and BN features for audio modality; LGBP-TOP, HOG and deep visual features for video modality. LSTM-RNN is adopted to train dimensional emotion regression model for every feature set considering the factor of annotation delays and temporal pooling. The turn timings information is added to feature sets to decrease the noise of the recordings. Visual features achieve better performance both in arousal and valence dimension, especially deep visual features and geometric features. Audio features can achieve a good complementary role for emotion recognition. Text features are beneficial to robustness of system especially liking dimension. To overcome overfitting problem, better and robust models are chosen carefully for each feature set, which improve the generalization performance of emotion regression system.

Table 4: Multi-modal prediction results of valence dimension on *dev2* set (*dev1* set).

	RMSE	PCC	CCC
A: Geometric + Deep visual+ LGBP-TOP	0.088 (0.116)	0.744 (0.718)	0.718 (0.630)
B: A+ HOG	0.083 (0.106)	0.760 (0.753)	0.754 (0.697)
C: B+ IS10+ eGeMAPS	0.079 (0.105)	0.784 (0.771)	0.774 (0.699)
D: C+ BoTW	0.078 (0.103)	0.786 (0.777)	0.776 (0.706)
All	0.079 (0.104)	0.785 (0.773)	0.772 (0.692)

Table 5: Multi-modal prediction results of liking dimension on *dev2* set (*dev1* set).

	RMSE	PCC	CCC
A: BoTW+ Deep visual	0.081 (0.113)	0.517 (0.566)	0.509 (0.529)
B: A+ LGBP-TOP	0.081 (0.114)	0.519 (0.565)	0.513 (0.530)
C: B+ IS10	0.077 (0.120)	0.542 (0.521)	0.530 (0.445)
D: C+ Geometric	0.077 (0.121)	0.559 (0.531)	0.553 (0.458)
All	0.076 (0.118)	0.550 (0.525)	0.534 (0.400)

Table 6: Decision level fusion results for AVEC 2017 development set and testing set, including the proposed method and

	Arousal			Valence			Liking		
	RMSE	PCC	CCC(Baseline)	RMSE	PCC	CCC(Baseline)	RMSE	PCC	CCC(Baseline)
Development	0.102	0.726	0.721(0.525)	0.094	0.771	0.728(0.507)	0.106	0.524	0.481(0.314)
Testing	0.093	0.609	**0.599**(0.375)	0.085	0.725	**0.721**(0.466)	0.150	0.338	**0.295**(0.246)

Multimodal emotion fusion is achieved by utilizing Support Vector Regression (SVR) with the estimates from different feature sets in decision level fusion. The experimental results reveal that our proposed system achieves very promising results on both the development and testing set of AVEC 2017.

ACKNOWLEDGMENTS

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017), the National Key Research & Development Plan of China (No. 2016YFB1001404) and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES

[1] Tao J, Tan T. Affective computing: A review. International Conference on Affective computing and intelligent interaction. Springer, Berlin, Heidelberg, 2005: 981-995..

[2] Kächele M, Schels M, Meudt S, et al. Revisiting the EmotiW challenge: how wild is it really? J. Multimodal User Interfaces, 2016, 10(2): 151-162.

[3] Gunes H, Pantic M, Ashour A S. Automatic, Dimensional and Continuous Emotion Recognition. International Journal of Synthetic Emotions, 2010, 1(1):68-99.

[4] Gunes H, Schuller B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing, 2013, 31(2): 120-136.

[5] Schuller B, Valstar M, Eyben F, et al. Avec 2011–the first international audio/visual emotion challenge. Affective Computing and Intelligent Interaction, 2011: 415-424.

[6] Schuller B, Valster M, Eyben F, et al. AVEC 2012: the continuous audio/visual emotion challenge. Proceedings of the 14th ACM international conference on Multimodal interaction. ACM, 2012: 449-456.

[7] Valstar M, Schuller B, Smith K, et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, 2013: 3-10.

[8] Valstar M, Schuller B, Smith K, et al. Avec 2014: 3d dimensional affect and depression recognition challenge. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, 2014: 3-10.

[9] Ringeval F, Schuller B, Valstar M, et al. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 1335-1336.

[10] Valstar M, Gratch J, Schuller B, et al. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016: 3-10.

[11] Ringeval F, Schuller B, Valstar M, et al. AVEC 2017: Real-life depression and affect recognition workshop and challenge.

[12] Miranda-Correa J A, Abadi M K, Sebe N, et al. AMIGOS: A dataset for Mood, personality and affect research on Individuals and GrOupS. arXiv preprint arXiv:1702.02510, 2017.

[13] Eyben F, Scherer K R, Schuller B W, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[14] Grézl F, Egorova E, Karafiát M. Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure. Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014: 48-53.

[15] Fér R, Matějka P, Grézl F, et al. Multilingual bottleneck features for language recognition. Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[16] Garcia-Romero D, McCree A. Insights into deep neural networks for speaker recognition. Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[17] Popková A, Povolný F, Matějka P, et al. Investigation of Bottle-Neck Features for Emotion Recognition. International Conference on Text, Speech, and Dialogue. Springer International Publishing, 2016: 426-434.

[18] Povolny F, Matejka P, Hradis M, et al. Multimodal Emotion Recognition for AVEC 2016 Challenge. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016: 75-82.

[19] Sun B, Cao S, Li L, et al. Exploring multimodal visual features for continuous affect recognition. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016: 83-88.

[20] Almaev T R, Valstar M F. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013: 356-361.

[21] Wöllmer M, Kaiser M, Eyben F, et al. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 2013, 31(2): 153-163.

[22] Wöllmer M, Kaiser M, Eyben F, et al. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 2013, 31(2): 153-163.

[23] He L, Jiang D, Yang L, et al. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 2015: 73-80.

[24] Chao L, Tao J, Yang M, et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 2015: 65-72.

[25] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010: 1459-1462.

[26] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2010 paralinguistic challenge. Eleventh Annual Conference of the International Speech Communication Association. 2010.

[27] Brady K, Gwon Y, Khorrami P, et al. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016: 97-104.

[28] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[29] Liu M, Wang R, Li S, et al. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 2014: 494-501.

[30] Schmitt M, Schuller B W. Openxbow-introducing the passau open-source crossmodal bag-of-words toolkit. arXiv preprint arXiv:1605.06778, 2016.

[31] Mariooryad S, Busso C. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. IEEE Transactions on Affective Computing, 2015, 6(2): 97-108.

[32] Ringeval F, Eyben F, Kroupi E, et al. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters, 2015, 66: 22-30.

[33] Chao L, Tao J, Yang M, et al. Multi-scale temporal modeling for dimensional emotion recognition in video. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, 2014: 11-18.

[34] Zeiler M D. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.