Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition

Linlin Chao
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
linlin.chao@nlpr.ia.ac.cn

Jianhua Tao*
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
jhtao@nlpr.ia.ac.cn

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

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

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

ABSTRACT
This paper presents our effort to the Audio/Visual+ Emotion Challenge (AV+EC2015), whose goal is to predict the continuous values of the emotion dimensions arousal and valence from audio, visual and physiology modalities. The state of art classifier for dimensional recognition, long short term memory recurrent neural network (LSTM-RNN) is utilized. Except regular LSTM-RNN prediction architecture, two techniques are investigated for dimensional emotion recognition problem. The first one is ε-insensitive loss is utilized as the loss function to optimize. Compared to squared loss function, which is the most widely used loss function for dimension emotion recognition, ε-insensitive loss is more robust for the label noises and it can ignore small errors to get stronger correlation between predictions and labels. The other one is temporal pooling. This technique enables temporal modeling in the input features and increases the diversity of the features fed into the forward prediction architecture. Experiments results show the efficiency of key points of the proposed method and competitive results are obtained.

Categories and Subject Descriptors
J [Computer Applications]: J.4 SOCIAL AND BEHAVIORAL SCIENCES; Subjects: Sociology

Keywords
Affective Computing; Emotion Recognition; Speech; Facial Expression; Multimodal; Challenge

*The author is further affiliated with Institute of Neuroscience, State Key Laboratory of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

AVEC'15, October 26 2015, Brisbane, Australia
©2015 ACM. ISBN 978-1-4503-3743-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2808196.2811634

1. INTRODUCTION
Emotion recognition plays an important role in human machine interaction, furthermore, has gained increasingly intensive attention [1]. Typically, the majority of the work in this field has focused on analysis of the acted or stereotypical emotions. The emotions are classified into some discrete basic emotions [2] (e.g., happiness, sadness, surprise, fear, anger and disgust). Although many promising recognition results have achieved recently, this work still cannot meet the needs of real life, because we exhibit non-basic, subtle and rather complex emotional states, which cannot be fully expressed by one category emotion label [3].

Meanwhile, a number of researchers argue that the dimensional approach to emotion modeling is more suitable to express our complex emotions [4, 5]. They try to learn emotion in a multi-dimensional emotion space rather than some basic discrete categories. For example, the arousal dimension refers to how excited or apathetic the emotion is. The valence dimension refers to how positive or negative the emotion is. The dominance dimension refers to the degree of power or sense of control over the emotion [6, 7]. In this space, the various emotional states locate in different positions and their similarities and differences can be expressed by their distances in this space. This research is not to detect several prototypic emotions but to recognize emotion states at each moment, which aims at working toward subtle, continuous, and context-specific interpretations of affective displays recorded in real world settings.

In this paper, we mainly focus on dimensional emotion recognition from audio, visual and physiology modalities. Based on the LSTM-RNN prediction architecture, which is the state of art classifier for dimensional emotion recognition, two techniques are introduced into dimensional emotion recognition. The two techniques are the followings:

1. ε-insensitive loss: ε-insensitive loss is utilized as the loss function to train the neural network. Label noise is an inevitable problem for dimensional emotion. Compared to squared loss function, which is the most widely used loss function for dimension emotion recognition, ε-insensitive loss is more robust for the label noises. Besides, ε-insensitive loss can ignore small errors to get stronger correlation between the predictions and labels;

2. Temporal pooling: temporal pooling among successive hidden layer representations is utilized before the LSTM layer. This technique enables temporal modeling in the input features and increases the diversity of the features fed into forward prediction architectures.

The rest of this paper is organized as follows. In section 2 and section 3, related work and the dataset we utilized are introduced separately. Section 4 gives the details of the regression model. Section 5 introduces the multimodal features. The experiments and results are presented in section 6. The conclusions are given in section 7.

2. RELEATED WORK

Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) is one of the state-of-the-art machine learning techniques in dimensional emotion recognition. It has the ability to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context. It has proven its success in Audio/ Visual Emotion Challenge 2011 (AVEC2011) [8]. In this challenge, regression problem for dimensional emotion is simplified to a positive and negative classification problem. Combining audio and visual modalities, [9] gets the state of art performance. For regression problem, which is more directly for dimensional emotion recognition, Ringeval et al. [10] investigates LSTM-RNN for audio, visual and physiological modalities based dimensional emotion recognition. In their work, data asynchronous between continuous ratings and data, analysis window size for emotion dimensions and multimodal fusion are analyzed. Gunes et al. [31] also utilizes LSTM-RNN to analyze the dimensional emotion recognition problem and competitive results are obtained.

Obtaining high inter-observer agreement is one of the main challenges in emotion data annotation, especially when for dimensional emotion [3]. Thus, the label noise is an inevitable problem. Michel et al. [11] tries to minimize the label noise by calculating the average ratings from all raters. [12] centers the ratings from different raters according to the mean value of all raters, then combines these ratings linearly weighted by their respective inter rater agreement. The agreement is measured by the concordance correlation coefficient (CCC). Except the above methods, [10] adds window to the label and features. By averaging the labels in the window, the label is smoothed and the noise is decreased to some extent. Researchers try to minimize the annotation noises in a variety of ways. However, when labels with noise are given, seldom considers this problem from the loss function.

Another work which relates to our work is temporal pooling. It is first introduced into dimensional emotion recognition by [13]. They utilize temporal pooling function in deep belief network. Temporal pooling functions enable temporal modeling in the forward network and competitive results are obtained. In this context, we utilize the temporal pooling function before LSTM layer. Two objectives are obtained. One is temporal modeling in short scale. The other one is increasing the diversity of the features fed into forward prediction layers.

3. DATASET

The challenge is evaluated on the RECOLA dataset [14]. In this dataset, subjects are recorded by audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA) modalities. Spontaneous and naturalistic interactions are collected when subject tries to solve a collaborative task. There is only one person in every recording.

This dataset is annotated in two emotion dimensions by 6 raters. These dimensions are arousal, valence. For each recording, the labels are the weighted linear combination of all the raters. The weight is determined by CCC. In each dimension, the range of the labels is scaled to [-1, 1], and a regression problem needs to be solved for each dimension. The average CCC of the two dimensions will be used to rank participants. Details of the dataset can be found in [14].

4. REGRESSION MODEL

For this challenge, we use the LSTM-RNN based neural network as prediction model. All the modalities (audio, visual, physiology) have the same prediction architecture, with one hidden layer, one temporal pooling layer, two LSTM-RNN layers and one linear prediction layer. Fig.1. depicts the architecture of the proposed regression model.

4.1 LSTM Layer

LSTM-RNN has the ability to learn long-term dynamic while avoiding the vanishing and exploding gradients problems. As research on LSTMs has progressed, hidden units with varying connections within memory unit have been proposed. We use the LSTM unit described in [15] (Fig.2.), which is a slight simplification of the one described in [16]. The LSTM updates for time step *t* given inputs m_t, h_{t-1}, and c_{t-1} are:

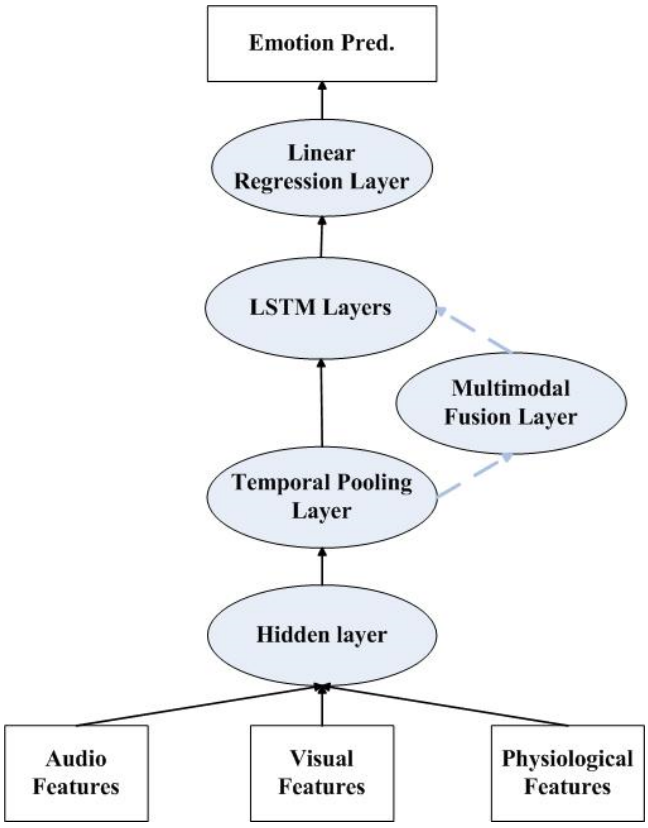

Figure 1 Overview of the proposed regression model. Feature level fusion method is also depicted. The dash line represents the feature level fusion path different with single modality prediction. For single modality prediction, only one type features are fed into the network.

$$i_t = sigmoid(W_{xi}m_t + W_{hi}h_{t-1} + b_i)$$
$$f_t = sigmoid(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$
$$o_t = \text{sigmoid}(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$
$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * g_t$$
$$h_t = o_t * \tanh(c_t)$$

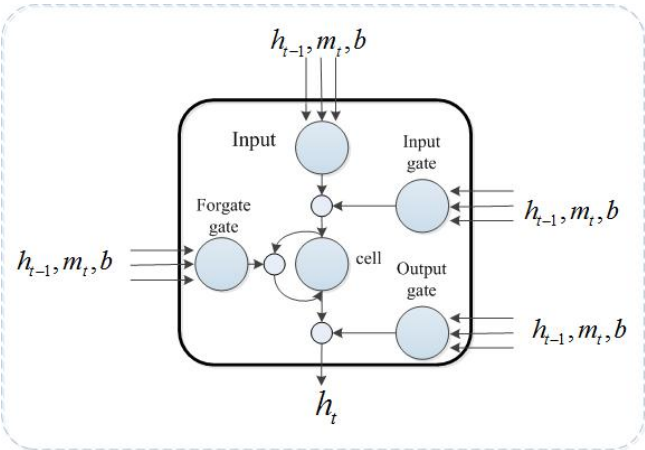

Figure 2 Architecture of the LSTM unit

4.2 Loss Function

For a regression problem, squared loss, absolute loss and ε -insensitive loss are the three widely used loss functions. Squared loss function is also the most widely used for neural network training. Suppose y is the target and $f(x,w)$ are the predicted value, the three loss functions can be shown in follow ways respectively:

$$L_{squared} = (y - f(x,w))^2$$
$$L_{absolute} = |y - f(x,w)|$$
$$L_{\varepsilon} = \begin{cases} 0 & if\ |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon & \text{otherwise} \end{cases}$$

where ε is a parameter determined by experiments. Fig.3. shows the plots of the above three types of loss functions.

Squared loss is one such function that is well-suited for the purpose of regression problems. However, it suffers from one critical flaw: outliers in the data (isolated points that are far from the desired target function) are punished very heavily by the squaring of the error. As a result, data must be filtered for outliers first, or else the fit from this loss function may not be desirable. Absolute loss is applicable to regression problems just like squared loss, and it avoids the problem of weighting outliers too strongly by scaling the loss only linearly instead of quadratically by the error amount. ε -insensitive loss is identical in behavior to the absolute loss function The main idea of ε -insensitive loss is to ignore 'small' errors during minimization, whereas 'large' errors are assigned absolute value loss. The conceptual difference between ε -insensitive loss and other types of loss functions (squared loss and absolute loss) is that the latter do not make qualitative distinction between small errors and large errors [17].

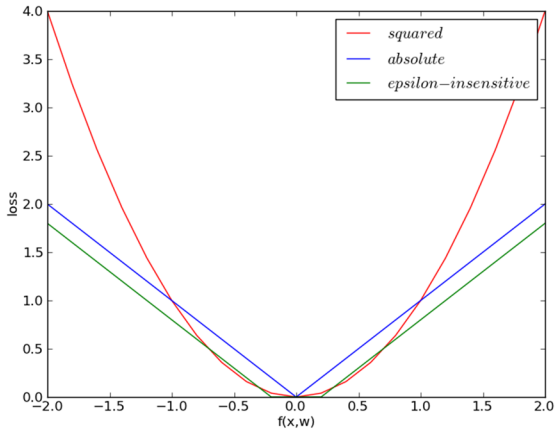

Figure 3 Plots of typical squared loss function, absolute loss function and ε -insensitive loss function

In the context of dimensional emotion recognition, label annotation is difficult. There is no coding scheme that is agreed upon and used by all researchers in the field [3]. Although the annotation noise can be decreased by averaging or linear combination of several raters, the noises can still exist. We believe that mainly of the large errors can be partly decreased and they can have little influence to the small errors. Because the squared loss is sensitive to label noises, especially for the "large errors", absolute loss function and ε -insensitive loss function are better than squared loss function for dimensional emotion recognition.

When judging the performance of dimensional emotion recognition, root mean square error (RMSE) is a judgment criterion. From this point of view, absolute loss function is better than ε -insensitive loss function. However, small RMSE is relatively easy to achieve. Getting a high correlation score between the label and the prediction is also what we want. That's because the correlation score reflects the overall trend between the label and the prediction. With the RMSE exists in a rational range, we prefer the prediction changing with the same trend of the label. The judgment criterias of AVEC2012-2015[23, 24, 25, 12] take the Person Correlation Coefficient (PCC) into account. PCC is an even harder objective to achieve than RMSE. We believe ignoring some "small errors" is helpful to get a higher PCC, or CCC, which is the judgment criterion of this Challenge. Thus ε -insensitive loss is more suitable than the other loss functions for dimensional emotion recognition.

4.3 Temporal Pooling Layer

Similar to pooling layer in convolutional neural network (CNN), temporal pooling layer pools the features from temporal input rather than spatial inputs. The differences are that the pooling window is one dimensional and pooling functions are added in temporal sequence inputs. The most widely used pooling ways are mean pooling and max-pooling. Mean pooling tends to get the context information. Max-pooling gets the maximum response.

For dimensional emotion regression problem, temporal modeling is one of the vital steps for getting good performance. The temporal pooling can get the statics of the successive frames, for example mean or maximum, which achieve short level temporal modeling.

Similar to temporal pooling, adding window to the input features and computing the average features is a widely used method. One

of the important reasons for computing the average features is that the emotion labels are also binned in the same size of window. Because label binning can smooth the label and decrease label noises to some extent. The proposed temporal pooling layer can also have the same function. When the hidden layer features are temporal pooled, the corresponding labels are also binned. However, as the temporal pooling operation is utilized after a hidden layer, the pooled features are changing dynamically with the weights changes in the backward hidden layer. The pooled features can be more expressive compared to the single average features from the raw features. Thus the temporal pooling operation can also increase the diversity of the features fed into the following layers.

In this challenge, mean pooling, max pooling and their combination are utilized. Experiments' result shows that mean pooling get the best performance. All the network architectures utilize the mean pooling in temporal pooling layer.

4.4 Multimodal Fusion Layer

Multimodal fusion layer is designed for feature level fusion specially. Suppose there are two input modalities. After temporal pooling layer, the outputs are a_t, s_t separately. In multimodal fusion layer, these two inputs are concatenated together. The equation is below:

$$m_t = tanh(W_m\,[a_t, s_t] + b_m)$$

where W_m and b_m is the weight and bias in this layer.

4.5 Multimodal Decision Level Fusion

The decision level fusion results are obtained by LSTM-RNN. For each dimension, the input are the prediction results from each feature sets on both dimensions. There are two fusion regression models are trained.

5. Multi-modal Features

There are three modalities in this challenge, including audio, visual and psychological. All the baseline features from different modalities are utilized. The audio features are the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [18] features set, with window analysis of 3 seconds, shifted forward at a constant 40ms rate. For visual modality, there are three features sets. The first one is appearance feature set. It is the LGBP-TOP features set [19]. The second one is geometric features, which are derived from facial landmarks. Details can be found in the baseline paper [12]. The last one is the pixel coordinates (image coordinate system) of 49 landmarks. We use the PCA whitened landmarks feature set and 30 dimensions are kept. For psychological modality，features are extracted from ECG and EDA signals. The ECG signals are filtered by 5th order Butterwoth bandpass. Then features like heart rate (HR) and its measure of variability (HRV), zero-crossing rate and other statistical data are extracted. For EDA, 62 features are extracted, including skin conductance response (SCR), skin conductance level (SCL), and relative statistics from EDA, SCR and SCL.

For visual modality, we also extract appearance features from CNN. We employ the Caffe [27] implementation. The combination of Celebrity Faces in the Wild (CFW) [28] and FaceScrub dataset [29] are utilized to train the CNN. These two datasets are designed for face recognition. In this context, they are utilized to train network for extracting face representations. Over 110,000 face images from 1032 people are used for training and the labels are their identities. The architecture is the same with [30]. The 9216 nodes' values of the last convolutional layer are used for face features. In order to prevent over-fitting, PCA dimension reduction is applied and only 41% energy of the total variance are kept as the final face representation.

There are total seven feature sets. We name these feature sets with audio, LGBP-TOP, geometric, landmarks, face-CNN, ECG and EDA respectively.

6. Experiments and Setup

We follow the challenge criterion of AVEC2015, with training set for training, development set for validation. For neural network, we use the implementation from Theano [20] [21]. For different feature sets, the architectures of the network keep same except the number of nodes in the input layer. 64 memory cells are utilized in the LSTM layers. Adadelta [22] optimization algorithm with batch size 5 is utilized. The maximum training epoch is 150 with dropout technique utilized in all layers except the LSTM layers. The drop rate is 0.5. Weight decay in the linear regression layer with the parameter 0.0005 is also applied to prevent over fitting. The best results are chosen by CCC in the development set.

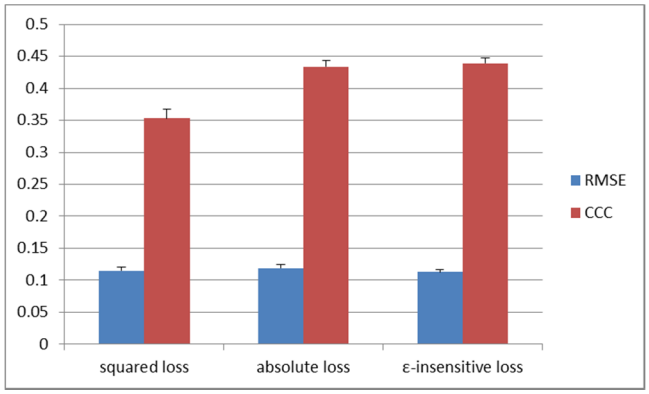

Figure 4 Performance comparisons among absolute loss, squared loss and ε -insensitive loss. The experiments are conducted on audio data and valence dimension performances are compared.

Table 1 Performance comparisons among different epsilon for ε -insensitive loss. All experiments are conducted on audio data and valence dimension performances are compared.

Epsilon	RMSE	CCC
0.01	0.1167	0.4374
0.001	0.1143	0.4340
0.0001	0.1127	**0.4393**
0.00001	**0.0920**	0.4364

6.1 Test for Loss Function

We test the performances of different loss functions with the proposed regression model. All the parameters keep the same, except the loss function. For different loss functions, five neural networks are trained. The input feature is audio features. The performances of valence dimension are compared. We present the averages and the standard deviations of RMSE and CCC in five trails in Fig.4. The experiment results show that absolute loss and ε -insensitive loss performs better than squared loss in CCC significantly. Compared to absolute loss, ε -insensitive loss also gets better performance in CCC. For RMSE, the three loss functions get similar results. Also, The performance of ε -

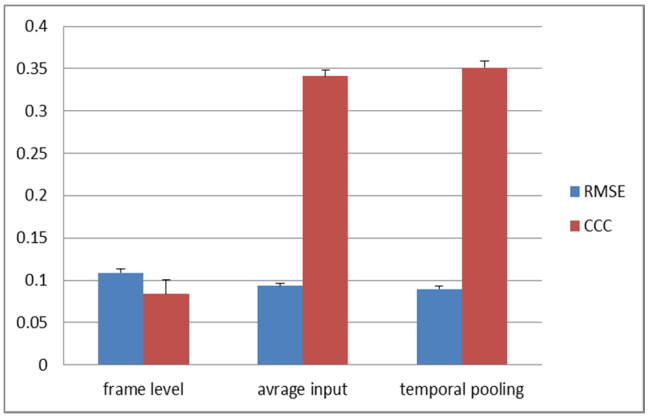

Figure 5 Performance comparisons among temporal pooling, average input and frame level input. The experiments are conducted on audio data and valence dimension performances are compared. Temporal pooling architecture is the proposed architecture. Average input architecture is the proposed network architecture without temporal pooling layer and the input feature and corresponding label are averaged among successive frames. Frame level architecture is the proposed prediction architecture without temporal pooling layer.

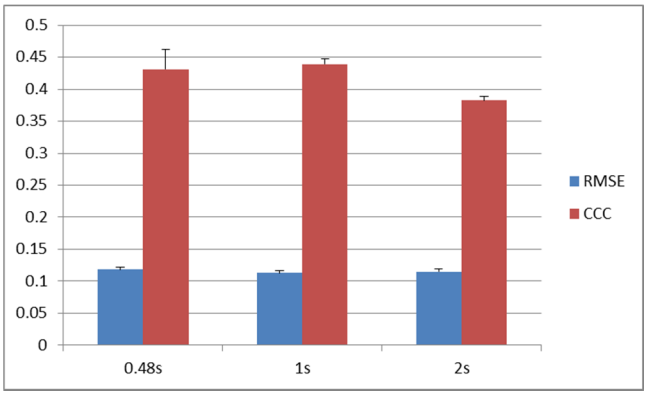

Figure 6 Performance comparisons for different window size for temporal pooling. The experiments are conducted on audio data and valence dimension performances are compared. The average and standard derivate of RMSE and CCC are shown.

insensitive loss for RMSE does not be influenced. These results prove that ε-insensitive loss is the most suitable loss function for dimensional emotion recognition among these three loss functions.

We also compare different epsilon parameters for ε-insensitive loss function. For different epsilon, five neural networks are trained, with audio features input and valence dimension predicted. The average RMSE and CCC are listed in table 1. From the table, we can see when the epsilon equals to 0.00001, the best RMSE is get. The RMSE is change monotonous with the epsilon parameter. And the best CCC is obtained when epsilon equals to 0.0001. As the CCC combines the PCC and RMSE together, the best performance is chosen by CCC. We set the epsilon parameter to be 0.0001.

6.2 Test for Temporal Pooling

This part shows the experiments results about the proposed temporal pooling architecture compared with average input architecture and frame level input architecture. Different with temporal pooling architecture, the average input architecture does not have temporal pooling layer and the input features and labels are averaged in a given window. Frame level architecture is the raw features are fed into the neural network directly and temporal pooling layer is also not applied. Five neural networks for different architectures are trained. These experiments are also conducted on audio modality and judged by valence dimension. Fig.5. shows the experiment results. Frame level architecture is significantly worse than the other two, especially for CCC. One of the reasons is that without label binning or smoothing, the label contains too much noise. When the label is binned or smoothed, the results of CCC improve by a large margin. The average input shows the results after label binned or smooth. Compared to the average input, the proposed temporal pooling architecture improves the CCC further.

The window size for temporal pooling influences the final performances. Fig. 6 shows the performances comparison among different window sizes. Five neural networks are trained and the input feature is audio features. Valence dimension is predicted. From Fig.6, we can see when the window size equals to 1 second, the best results are obtained. The window size for the temporal pooling is set to 1 second in all the experiments.

6.3 Experiment Results

Experiments results for each single feature set are shown in Table 2. Seven feature sets are compared. There are one feature set for audio modality, four feature sets (LGBP-TOP, face-CNN,

Table 2.Performance comparisons with the proposed regression model and different feature sets for the AVEC 2015 training set and development set

	Training						Development						
	Arousal			Valence			Arousal			Valence			Avg.
	RMSE	PCC	CCC	RMSE	PCC	CCC	RMSE	PCC	CCC	RMSE	PCC	CCC	CCC
Audio	0.101	0.896	**0.861**	0.097	0.712	0.707	0.123	0.798	**0.798**	0.115	0.494	0.483	**0.640**
LGBP-TOP	0.134	0.752	0.740	0.098	0.689	0.672	0.188	0.550	0.535	0.121	0.488	0.463	0.499
Face-CNN	0.135	0.864	0.738	0.054	0.923	**0.916**	0.203	0.348	0.336	0.116	0.561	**0.538**	0.437
Geometric	0.148	0.672	0.641	0.102	0.658	0.598	0.189	0.427	0.411	0.111	0.514	0.488	0.449
Landmarks	0.162	0.456	0.371	0.103	0.644	0.621	0.212	0.182	0.137	0.118	0.493	0.483	0.310
ECG	0.165	0.426	0.349	0.122	0.403	0.270	0.202	0.341	0.222	0.122	0.272	0.182	0.202
EDA	0.177	0.235	0.108	0.130	0.282	0.119	0.213	0.177	0.062	0.119	0.317	0.153	0.108

Table 3.Performance comparisons with the proposed regression model and with different feature set fused in the feature level for the AVEC 2015 training set and development set. The Lgb, Geo, Aud, Lan, Fac are short for LGBP-TOP, geometric, audio, landmarks, face-CNN respectively.

	Training				Development				
	Arousal		Valence		Arousal		Valence		Avg.
	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC	CCC
Lgb+Geo	0.146	0.584	0.103	0.573	0.166	0.470	0.107	0.492	0.481
Aud+ Lgb +Geo	0.114	0.808	0.102	0.593	0.122	0.790	0.103	0.541	0.665
Aud+ Lgb +Geo+Lan	0.109	0.821	0.104	0.585	0.123	0.786	0.105	0.540	0.663
Aud+ Lgb +Geo+Lan+ECG	0.112	0.816	0.103	0.579	0.124	0.787	0.102	0.550	0.669
Aud+Lgb+Geo+Lan+ECG+EDA	0.122	0.798	0.100	0.651	0.126	**0.791**	0.105	0.582	**0.687**
Aud+Lgb+Geo+Lan+ECG+EDA+Fac	0.115	0.813	0.078	0.802	0.126	0.777	0.102	**0.590**	0.684

Table 4 Decision level fusion results from audio, LGBP-TOP, geometric, landmarks, ECG and EDA feature sets for AVEC 2015 training set, development set and testing set.

	Arousal			*Valence*			*Avg.*
	RMSE	PCC	CCC	RMSE	PCC	CCC	CCC
Training	0.101	0.907	0.848	0.078	0.811	0.803	0.826
Development	0.113	0.813	0.801	0.097	0.657	0.635	0.718
Testing	0.126	0.711	0.693	0.101	0.612	0.579	0.636

Table 5 Performance comparison between the proposed approach and baseline in AVEC 2015 testing set.

	Arousal			*Valence*			*Avg.*
	RMSE	PCC	CCC	RMSE	PCC	CCC	CCC
Our approach	0.137	0.718	0.716	0.103	0.627	0.618	0.667
Baseline [12]	0.164	0.354	0.444	0.113	0.490	0.382	0.413

geometric and landmarks) for visual modality and two feature sets (ECG, EDA) for physiological modality. Among the three modalities, audio modality gets the best CCC, which is much higher than the others. Physiological modality shows the worst results in CCC. The best result for arousal dimension is achieved by audio modality and the best result for valence dimension is achieved by visual modality, which comes from face-CNN feature set. These results are in agreement with previous studies [3] [26]. In visual modality, the four feature sets can be classified to two categories, appearance based and shape based. The LGBP-TOP and face-CNN feature sets are the appearance based. The shape based feature sets are geometric feature and landmarks feature. The appearance based feature set shows better performance than the shape based feature set in both arousal dimension and valence dimension. The LGBP-TOP feature set gets the best CCC for arousal dimension except audio feature, while the face-CNN feature set gets the best CCC for valence dimension among all the feature sets. Both ECG feature set and EDA feature set in physiological model have significant gaps from the other two modalities, which are in agreement with previous study [10]. For RMSE, all the feature sets get similar results on valence dimension. While the results on arousal dimension from all the seven feature sets show no significant regularity

Table 3 shows the result from feature level fusion. We list some combination of the feature sets. After feature level fusion, the performances improve significantly, especially for valence dimension. With more feature sets added, the CCC keeps increasing. The best performance of valence dimension is got when all the feature sets are fused. On the other hand, performance on arousal dimension shows quite different results. When audio feature sets fused with other feature sets, the CCC decreased slightly. Without audio feature, both the results of arousal dimension and valence dimension decrease significantly. This proves the importance of audio modality for both arousal dimension and valence dimension.

Table 4 shows the result of decision level fusion for results from audio feature, LGBP-TOP feature, geometric feature, landmarks feature, ECG feature and EDA feature. The results show that better performance is achieved on development set compared with feature level fusion. Meanwhile, over fitting also exist in both dimensions.

Table 5 shows the final submitted result on testing set. The result comes from feature level fusion from audio feature, LGBP-TOP feature, geometric feature, landmarks feature, ECG feature and EDA feature. Compared with the baseline [12], the proposed

approach gets better performance in both arousal dimension and valence dimension. However, over-fitting exists in arousal dimension. As we also test the performance of audio feature only on test set and the over-fitting also exists.

7. CONCLUSION

This article presents our approach that models dimensional emotion recognition using LSTM-RNN based neural network. Better loss function for dimensional emotion recognition is analyzed. On one hand, ε -insensitive loss function is more robust for the label noises, which is inevitable for dimensional emotion recognition. On the other hand, ε -insensitive loss can ignore small errors to get stronger correlation between the predictions and labels. Experiment results shows ε -insensitive loss function is the more suitable loss function for dimensional emotion recognition compared with squared loss function and absolute loss function.

Temporal pooling technique is utilized to the LSTM-RNN based neural network. This technique enables short level temporal modeling for the input features. As it pools the features after a hidden layer rather than the raw features, it also increases the diversity of the features fed into forward layers compared to when the averaged raw feature is fed into the neural network. Experiment results also show the effectiveness of the proposed technique.

Although ε -insensitive loss function show its advantage in dimensional emotion recognition, better loss function or more techniques should be put forward to deal with the label noises and improve the correlation relationship between the predictions and the labels. Thus, we will put more effort on the two points for dimensional emotion recognition.

8. 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, No.61332017, No.61375027, No.61305003, No.61203258) and the Major Program for the National Social Science Fund of China (13&ZD189).

The authors would like to thank the developers of Theano and the organizers of AV+EC2015.

9. REFERENCES

[1] J. Tao and T. Tan, “Affective Computing: A Review,” Proc. First Int’l Conf. Affective Computing and Intelligent Interaction, J. Tao, T. Tan, and R.W. Picard, eds., pp. 981-995, 2005.

[2] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58. doi:10.1109/TPAMI.2008.52.

[3] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” Int. J. Synthetic Emotions, vol. 1, no. 1, pp. 68-99, 2010.

[4] J. R. Fontaine, K. R. Scherer, E. B. Roesch, P. C. Ellsworth, “The world of emotions is not two-dimensional,” Psychological science, 18(12), 1050-1057,2007.

[5] C. Breazeal, “Emotion and sociable humanoid robots” [J]. International Journal of Human-Computer Studies, 2003, 59(1): 119-155.

[6] A. Mehrabian, and J. Russell, An approach to environmental psychology. Cambridge, MA: MIT Press.

[7] J. Davitz, Auditory correlates of vocal expression of emotional feeling. In J. Davitz (Ed.), The communication of emotional meaning (pp. 101-112).New York: McGraw-Hill, 1964.

[8] B. Schuller, M. Valstar, F. Eyben, G. Mckeown, R. Cowie and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge." Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg, 2011. 415-424.

[9] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework, Image and Vision Computing, 2012.

[10] Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J. P., Ebrahimi, T.... & Schuller, B. (2014). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters.

[11] Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J. & Pantic, M. (2014, November). Avec 2014: 3d dimensional affect and depression recognition challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 3-10). ACM.

[12] Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D. & Pantic, M. (2015). The AV+EC 2015 Multimodal Affect Recognition Challenge: Bridging Across Audio, Video, and Physiological Data.

[13] Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2014, November). Multi-scale Temporal Modeling for Dimensional Emotion Recognition in Video. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 11-18). ACM.

[14] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on (pp. 1-8). IEEE.

[15] Zaremba W, Sutskever I. Learning to execute [J]. arXiv preprint arXiv:1410.4615, 2014.

[16] Graves A., Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014: 1764-1772.

[17] Cherkassky, V., & Ma, Y. (2002). Selecting of the loss function for robust linear regression. Neural computation.

[18] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C.Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, 2015. to appear.

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

[20] Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian, Bergeron, Arnaud,

Bouchard, Nicolas, and Bengio, Yoshua. Theano: new features and speed improvements. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2012.

[21] Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In Proceedings of the Python for Scientific Computing Conference (SciPy), June 2010.

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

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

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

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

[26] Gunes, H., Schuller, B., 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing: Affect Analysis in Continuous Input 31,120–136.

[27] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv: 1408. 5093, 2015.

[28] X. Zhang, L. Zhang, X.-J. Wang and H.-Y. Shum. Finding celebrities in billions of web images. Multimedia, IEEE Transactions on, 14 (4):995–1007, 2012.

[29] H.-W. Ng, S. Winkler. A data-driven approach to cleaning large face datasets. Proc. IEEE International Conference on Image Processing (ICIP), Paris, France, Oct. 27-30, 2014.

[30] Krizhevsky, A., Sutskever, I., Hinton. G., ImageNet Classification with Deep. Convolutional Neural Networks, NIPS 2012.

[31] H. Gunes, M. Piccardi, M. Pantic, From the Lab to the Real World: Affect Recognition Usng, Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition. I-Tech Education and Publishing, Vienna, Austria, pp. 185 - 218, 2008.