Combining Emotional History Through Multimodal Fusion Methods

Linlin Chao, Jianhua Tao and Minghao Yang
National Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{linlin.chao, jhtao, mhyang}@nlpr.ia.ac.cn

***Abstract*—Continuity, one of the important characteristics of emotion, implies us that the emotional history may provide useful information for emotion recognition. Meanwhile, classification-based multimodal fusion methods show effective results in multimedia analysis tasks. In this study, two-stage classification method is proposed and emotional history is combined by the support vector machine fusion method. Evaluations on the Audio Sub-Challenge of the 2011 Audio/Visual Emotion Challenge dataset show that: (i) combining emotional history to recognition improves the accuracy significantly, (ii) classification-based multimodal fusion methods can effectively combine emotional history.**

I. INTRODUCTION

Emotion recognition plays a important role in human-computer interaction, furthermore, having increasingly intensive attention [1]. Typically, the majority of the work in this field has focused on analysis of the acted or stereotypical emotions, and the emotions are classified into some discrete basic emotions [2, 3] (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 emotional state label.

Meanwhile, a number of researchers argue that the dimensional approach to emotion modeling is more suitable to expressing our complex emotions [4, 5]. They try to learn emotion in a multi-dimensional emotion space rather than some basic discrete categories. 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.

In cognitive sciences, researchers argue that the dynamics of human behavior are crucial for its interpretation [6]. Moreover, a number of recent studies [7-10] in affective computing demonstrate this point of view. These imply us combining the emotional history into emotion recognition would improve the accuracy.

On the other hand, classification-based multimodal fusion methods [11] (e.g., support vector machine, Bayesian inference, Dempster-Shafer theory, neutral network) have been widely used to classify the multimodal observation into one of the pre-defined classes. These methods have already proved their effectiveness in multimodal fusion field.

In this article, we propose to use two-stage classification approach for dimensional emotion classification. First, support vector machine is used to classify the emotional state for every single period. Second, one of the classification-based fusion methods combines the past periods' classification results from the first stage to make the final decision. Previously, we use Bayesian inference to combine emotional history [12]. In this work, support vector machine is utilized, which shows more competitive results.

The rest of this paper is organized as follows. In Section II, several related works are introduced. Section III introduces the dataset and audio features. The proposed two-stage classification approach is introduced in detail in Section IV. The experiment results and discussions are presented in Section V. The conclusions are given in Section VI.

II. RELATED WORK

This work is based on the 2011 Audio/Visual Emotion Challenge [13]. In this challenge, most of the winners take temporal context information into account in their classifiers.

The first place winners [7] of the Audio Sub-Challenge[1] proposes a three-stage classification methods. Thirteen K-nearest neighbours (KNN) classifiers are used as the first stage classifiers, and their outputs are directly put into thirteen Hidden Markov Modals (HMM), which are the second stage classifiers, corresponding to each of the first stage classifier. At last the outputs of the thirteen HMMs are put into another HMM to make the final prediction. Glodek [8], the second place winner uses five different HMMs, and their outputs are summed to make the final decision. In these two methods the temporal context information is mainly included by HMM.

Long Short-Term Memory Recurrent Neutral Networks (LSTM-RNN), one of the state-of-the-art machine learning techniques is employed in this dataset [10]. In the Audio/Visual Sub-Challenge, it gets the best performance. LSTM-RNN can incorporate long rang emotional history to the final recognition, and this is one of the main reasons for its success. Besides, the authors think that long range temporal context modeling tends to increase the emotion recognition performance, which is also supported by their other work [14].

[1] http://sspnet.eu/avec2011/

III. DATASET AND FEATURES

The AVEC 2011 dataset is constructed from the SEMAINE database [15], which consists of emotionally colored conversations. Users are recorded while holding conversations with an operator who adopts in sequence four roles designed to evoke emotional reactions. The dataset utilized in this experiment is the Solid-SAL part. There are 24 recordings, with approximately 4 character conversation sessions per recording. It is split into three partitions for the AVEC challenge: training, developing, and testing partition, each consisting of 8 recordings. Because the testing partition labels are not public available, we use the training partition for training and the developing partition for testing.

This dataset is annotated in four emotion dimensions by 2 to 8 raters. These dimensions are arousal, expectancy, power and valence. These labels are given in binary form, where the database's creators compute the average value of each dimension over all interactions in the dataset, then compare each dimension j at every frame t with the average ratings value to give a final binary label: '0', '1'. '0' denotes below the average ratings value and '1' means above the average value. For the challenge the originally continuous dimensions are refined as binary ones, this reduced a regression problem into a classification one.

All the dataset in this experiment comes from the audio part. In the audio part, every uttered word is extracted and a label is given for each dimension. This dataset also provides the baseline features for every uttered word. The baseline feature set consists of 1941 features, composing of low-level descriptors (LLD) related to energy, spectral, voicing, etc, and their statics. For that the set of LLD covers a standard range of commonly used features in audio signal analysis, the whole feature set is utilized in this experiments.

IV. CLASSIFIER

The proposed classification architecture is shown in Fig 1. Two stages are composed. In the first stage, a classifier is used to make a decision about the state of every single period. Next, these isolated decision values are combined by another classifier.

A. First stage classifier

A discriminative classifier or any generative classifiers that returns posterior probabilities can be adopted in this stage. The posterior probabilities, enabling post-processing, represent the confidence of each class membership assignment. Classifiers like KNN, linear discriminant analysis (LDA), SVM can be utilized.

In this experiment, SVM is utilized as the first stage classifier. It shows a competitive performance on problems where the data are sparse, namely too many features and relatively few data, as is the case with the dataset of the audio part. It is a maximum margin classifier whose decision hyper-plane is chosen to maximize the separability of the two classes. In the feature space, data points that are closer to the margin are more easily confused with the opposing class than data points further from the margin. Thus, the distances from the margin provide a measure of the classifier confidence. The distances in SVM can be transformed into posterior probability distribution, namely the soft label.

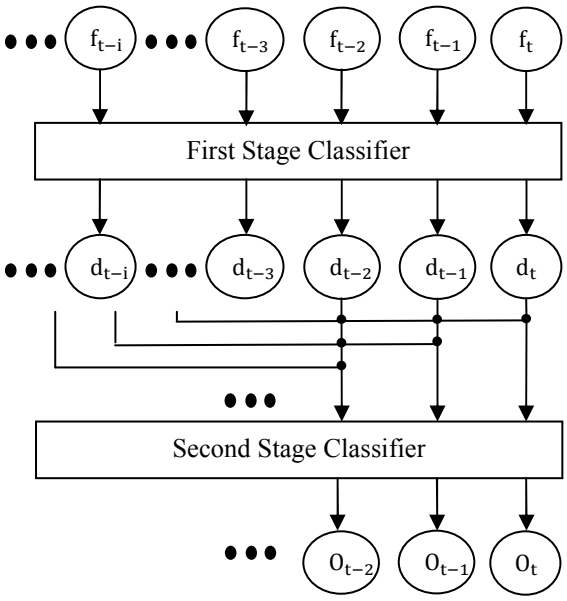

Fig.1. Overview of the proposed classification architecture. f_t represents features of signals at period t and d_t denotes the output decision values from the first stage classifier. O_t is the final decision value.

B. Second stage classifer

The second stage classifier is aimed at combing the results from the stage classifier. Classifiers in this category include the support vector machine, Bayesian inference, Dempster-Shafer theory and neutral networks, etc. We test the Bayesian inference and SVM, and their performances are convincing.

a) *Bayesian inference*

The Bayesian inference is often referred to as the 'classical' sensor fusion method because it has been widely used and many other methods are based on it [16]. From the perspective of multimodal fusion, Bayesian inference combines different modalities by the joint probability of an observation or a decision.

For the audio emotion classification, the temporal sequence is modeled as a Markov Chain [17], and Bayesian inference combines the past periods' emotional state by sum rule [18]. Details about this fusion process can be found in our previous work [12].

b) *Support vector machine*

From the perspective of multimodal fusion, SVM is used to solve a pattern classification problem, where the input to classifier is the scores given by the individual classifier. The basic SVM method is extended to create a non-linear classifier by using the kernel concept, where every dot product in the basic SVM formalism is replaced using a non-linear kernel function.

In order to combine emotional history, the inputs to the SVM from several individual classifiers are replaced by the same classifier' outputs, which come from different periods.

V. RESULTS AND DISCUSSION

The SVM used in the experiment is implemented by the LibSVM tool [19] and linear kernel is used in the first stage, RBF kernel for the second stage. All this two classifier are trained on the training set. In order to study the influence of different ranges of emotional history to the final classification accuracy, we change the ranges for comparison. In this experiment, the number of words represents the range of the emotional history and the window size represents the number of words it includes. We change the window size from one to thirty-one for comparison. So for every dimension, thirty classifiers in the second stage are trained.

From Fig.2, we can see with the increase of window size, the results in all four dimensions are increasing first and then remaining stable (expectancy and valence is not strictly monotone in its increasing region). Table I shows the increase rate when combines the emotional history through Bayesian inference and SVM. The result of Bayesian inference fusion comes from our previous work [12] and in SVM fusion, we choose the window size equals to thirty-one, namely the previous thirty words classification results are taken into account. We can see both the Bayesian inference and SVM improve the weighted accuracy in all dimensions significantly. Especially in arousal dimension, it increases 12.8% and 20.4% separately, which are much higher than the other three dimensions. One of the reasons may be that for audio modality variability of ratings of arousal appears to be smaller than other dimensions [3]. As a whole, SVM fusion is slightly better than Bayesian inference. Bayesian inference is a generative classifier, while SVM is a discriminative classifier, which is more suitable to classification tasks. Except this, the influence of emotional history' range to Bayesian inference and SVM is different. For Bayesian inference, with the increase of window size, the results in all four dimensions are increasing first and then decreasing gradually, and their best results peak when the window size equals to eight, which is very different from SVM.

TABLE I THE INCREASE RATE OF BAYESIAN INFERENCE AND SVM FUSION METHOD

	The increase rate (%)				
	Arousal	Expectancy	Power	Valence	Avg.
Bayesian[12]	12.8	1.2	6.6	6.8	6.9
SVM	20.4	2.4	4.0	5.4	8.1

TABLE II COMPARISON OF RECOGNITION RATES ON DEVELOPMENT DATASET BETWEEN PROPOSED METHOD AND SEVERAL LEADING METHODS

Classifier	Weighted accuracy(%)				
	Arousal	Expectancy	Power	Valence	Avg.
MLP-HMM[8]	66.9	62.9	63.2	65.7	64.68
GMM[20]	65.1	61.3	64.0	54.3	61.18
BLSTM[10]	71.3	66.2	66.0	**65.9**	67.38
SVM[13]	63.7	63.2	65.6	58.1	62.65
SVM*	60.8	65.9	64.5	61.2	63.10
SVM-Bayesian[12]	68.6	66.7	**68.7**	65.3	67.33
SVM-SVM	**73.2**	**67.5**	67.1	64.5	**68.08**

* This is the experiment result of SVM with linear kernel and probability output.

Table II shows weighted accuracies (WA) obtained by training on the training set of audio part of the AVEC 2011 dataset and testing on the developing set. BLSTM networks [10], the methods of the top two winners [8, 20], the SVM producing the challenge baseline in [13] and the proposed

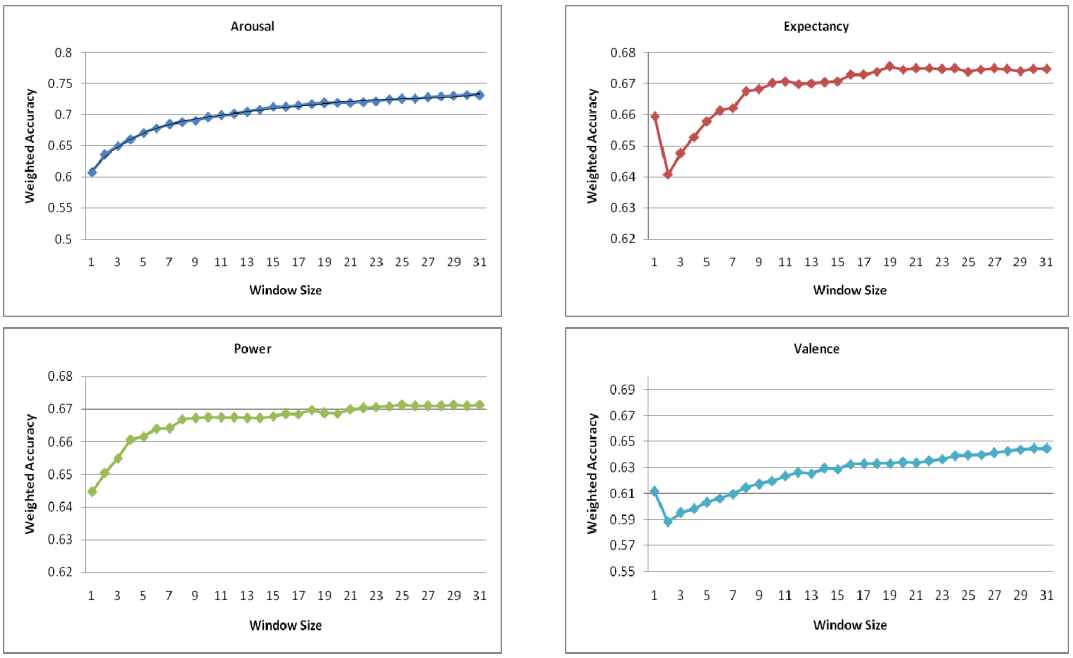

Fig.2. Weighted accuracy comparison among different ranges of emotional hitory for four affective dimensions. The range of emotional history is represented by the window size..

method are compared. The table shows that, under the same condition, the proposed method achieves the best accuracy in arousal, expectancy and power dimensions. In average, the proposed method also obtains fairly good results, ranking first compared with several leading methods in the weighted accuracy. These results show the importance of emotional history and the effectiveness of the proposed methods.

VI. CONCLUSIONS

This article discusses the importance of emotional history to emotion recognition and the practicability of classification-based multimodal fusion methods to combing emotional history.

The analysis provided in this article indicates that emotional history provides important information in the dimensional emotion recognition. One of the reasons is that in dimensional modeling of emotion, the continuity of emotion is expressed by the dimensional value sufficiently. Thus, modeling the emotional history effectively to emotion recognition will improve the performance.

Classification-based multimodal fusion methods are often used to combine different classifiers' output. In this article we use these methods to combine the outputs from the same classifier but from different periods. Evaluations on a public contest show the method we put forward is effective. Furthermore, this method can be helpful for other sequential pattern recognition problems.

Emotional information is conveyed by multimodal cues, including speech and facial expression. Thus, we will put our effort in combining multimodal information and emotional history information in the future.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC)(No.61273288, No.61233009, No.61203258, No. 61011140075, No. 90820303), and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

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] K. R. Scherer, "Appraisal Theory," Handbook of Cognition and Emotion, T. Dalgleish and M.J. Power, eds., pp. 637-663, Wiley, 1999.

[7] H. Meng, N. Bianchi-Berthouze, "Naturalistic Affective Expression Classification by a Multi-stage Approach Based on Hidden Markov Models," In Affective Computing and Intelligent Interaction (pp. 378-387). Springer Berlin Heidelberg, 2011.

[8] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, F. Schwenker, "Multiple Classifier Systems for the Classification of Audio-Visual Emotion States," In Affective Computing and Intelligent Interaction (pp. 378-387). Springer Berlin Heidelberg, 2011.

[9] G. A. Ramirez, T. Baltrušaitis, L. P. Morency, "Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals," In Affective Computing and Intelligent Interaction, (pp. 396-406). Springer Berlin Heidelberg, 2011.

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

[11] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhall, "Multimodal fusion for multimedia analysis : a survey," Multimedia Syst., vol. 16, no. 3, pp. 1432–1882, 2010.

[12] L. Chao, J. Tao, M. Yang and Y. Li, "Bayesian Inference based Temporal Modeling for Naturalistic Affective Expression Classification," In Affective Computing and Intelligent Interaction, 2013, in press.

[13] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, "AVEC 2011 – The First International Audio/Visual Emotion Challenge," in Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, (Memphis, Tennessee, USA), pp. 415–424, 2011.

[14] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, "Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependences," In Proc. Interspeech, pp. 597-600, 2008.

[15] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE Corpus of Emotionally Coloured Character Interactions," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 1079-1084, July 2010.

[16] D. Hall, J. Llinas, "An introduction to multisensor data fusion," Proceedings of the IEEE, 85(1), 6-23, (1997).

[17] J. L. Doob, J. L. Doob, "Stochastic processes," (Vol. 7, No. 2). New York: Wiley, 1953.

[18] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 226–239, Mar. 1998.

[19] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[20] J. C. Kim, H. Rao, M. A. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech. InAffective Computing and Intelligent Interaction," (pp. 369-377),Springer Berlin Heidelberg, 2011.