Facial Emotion Recognition

Ma Xiaoxi, Lin Weisi
Nanyang Technological University
Singapore
e-mail: {XMA007, wslin}@e.ntu.edu.sg

Huang Dongyan, Dong Minghui, Haizhou Li
Institute for Infocomm Research
Singapore
e-mail: {huang, mhdong, hli}@i2r.astar.edu.sg

***Abstract*—The interest on emotional computing has been increasing as many applications were in demand by multiple markets. This paper mainly focuses on different learning methods, and has implemented several methods: Support Vector Machine (SVM) and Deep Boltzmann Machine (DBM) for facial emotion recognition. The training and testing data sets of facial emotion prediction are from FERA 2015, and geometric features and appearance features are combined together. Different prediction systems are developed and the prediction results are compared. This paper aims to design an suitable system for facial emotion recognition.**

Keywords-facial emotion recognition; support vector machine; deep Boltzmann machine

I. INTRODUCTION

Emotional Computing [1] aims to enable machines to recognize and synthesize human emotions. As we all know, a change of user's emotion is one of the foundation of communication. Emotional states can motivate human's actions, and can also supplement the meaning of communication. However, the traditional HCI (Human-Computer Interaction) ignores the emotional states of users, and misses out a great part of information during the interaction process. Comparatively, emotion-sensitive HCI systems are more powerful and desired by users. Nowadays the interest on emotional computing has been increasing with the increasing requirements of applications related to amusement, commerce, physical and psychological health, and education. Because of this, many products of emotion-sensitive HCI systems have been produced in last several years, even though the ultimate solution has not been proposed to this research field.

Audiovisual signals are most significant parts of Emotional Computing. Our expressive face and voice contain the more than half of our emotion state, beyond what our words contain. Thus, numerous research focus on audiovisual features of emotional computing. Besides audiovisual features, some research works also focus on 3D expressions [2], Body micro-expressions [3] and combination of all these features, which is still a complex and open problem.

Nevertheless, existing algorithms and systems are mainly based on intentionally displayed and overstated expressions of original emotions, and ignore the difference between intentional behaviors and naturally occurring behaviors in all aspects: visualization, audio, and time. To solve this problem, recently many algorithms for processing spontaneously generated emotional behaviors have been proposed. Moreover, increasing research towards fusion methods is proposed for analyzing human emotion, such as feature-level fusion, decision-level fusion and multimodal fusion. Besides, the fusion methods for merging the information about facial expressions, head movements and body gestures [4] are also proposed.

This paper mainly focuses on different learning methods. As we know, different learning methods are suitable to different prediction problems. In this paper, we have implemented several learning methods which are ones of the best methods in general cases. Besides, these prediction systems are based on the characteristics of feature set and simple fusion is also implemented. The aim is to design and construct the prediction system which is most suitable to the challenges.

II. BACKGROUND

A. Feature Extraction

Current emotional facial recognition is mainly based on 2D facial features of spatial and temporal dimension. In general, extraction of facial features can be divided into two categories: geometric features and appearance features.

Usually, the geometric facial features represent the locations of facial salient points (corners of the eyes, mouth, etc.) and the shapes of the facial components (eyes, mouth, etc.). And the locations and shapes are usually defined by facial landmarks [5]. Besides, the appearance features represent the facial texture, including crinkles, wrinkles, and bulges. The appearance feature descriptors include Gabor wavelets [6]-[10], Haar features [11], and the using of spatial and temporal templates [12], [13] are also been proposed. From the opinion of several research (e.g. [14]), merging the information of both geometric and appearance features can obtain the best performance emotional facial recognition.

B. Machine Learning

For the learning stage, there are numerous machine learning methods that can be chosen. In general, machine learning methods can be divided into two categories: context-dependent methods and the methods with the static predictor. Typically, the context-dependent frameworks based methods include Hidden Markov Models [15] and

978-1-5386-0969-9/17/$31.00 ©2017 IEEE

Long Short-Term Memory Neural Networks (LSTMNN) [16], which obtain the advantage of dynamic learning. And for methods with the static predictor, the well-known Support Vector Machine [17][18] is one of the typical examples. Several typical machine learning methods for different types are described and implemented in this paper.

For emotion recognition, SVM is the widely used learning method, where SVC is for classification problems and SVR is for regression problems. The key point to adjust to different problems is to select an appropriate kernel function [19], and the parameters of kernel function are also needed to be determined.

Random Forest is a relatively new machine learning method. One of the classical machine learning methods is Neural Network method, which has a history of more than half a century. Neural Network method has good accuracy in prediction, but its computation complexity is very high. Breiman et al. presented the algorithm of Classification and Regression Tree (CART) in the 1980s, which recursively divides residual data into two parts for classification or regression and significantly reduces the computation complexity. While in 2001, Leo Breiman [20] presented an improved method of Random Forest, which combines Breiman's idea of bagging and the selection of features randomly, and summarize predicting results of every CART. Random Forest method highly improves the accuracy in prediction while the computation complexity is not significantly increased. Compared with decision trees, Random forests overcome the disadvantages of overfitting. And it is also robust to missing data and unbalanced data. Besides, this method can also predict precisely in high dimension data (e.g. thousands of feature attributes) without feature selection approach.

Besides, deep learning method is proposed nowadays. For many training tasks, their features obtain a natural hierarchical structure. Take image recognition as an example, the original input of the images is pixels, and the adjacent pixels constitute lines. Multiple lines compose textures, further forming the graphic pattern. Then the graphic patterns constitute local objects, until constituting the appearance of the entire object.

We can see that it is easy to find the relationship between the original input and shallow features, and through the middle features, the relationship between high-level features can be achieved step by step. However, directly spanning from original inputs to the high-level features is undoubtedly difficult. From Geoffery E. et al. [21], there are two main opinions: The artificial neural network of multiple hidden layers possesses excellent ability of the feature learning, and the features that have been learned are better at capturing the nature of the data, thereby facilitating visualization or classification. The difficulty of training the deep neural network can be effectively overcame by Layer-wise Pre-training, and the unsupervised Layer-wise Pre-training is given in their paper.

It remains unclear whether some learning methods are better than others, as only the recognition results given from specific same environment can be compared. For example, Marian has found that the combination of Adaboost and SVM's can obtain the best performance on Cohn and Kanade's DFAT-504 dataset [22], comparing with other learning methods. However, comparative results will differs from different database, we cannot qualify and compare these methods only from the results of precision rates. Similarly, for the categories of supervised and unsupervised learning methods, supervised learning methods can usually obtain better experimental results. However, some unsupervised methods, such as deep learning, outperform some of the supervised methods sometimes.

C. Fusion Methods

In general, fusion strategies can be divided into three categories: feature-level fusion, decision-level fusion and model-level fusion. Feature-level fusion (Early fusion) is to accumulate different feature sets into one structure and input it to the classifier entirety. For decision-level fusion (Late fusion), the whole feature set is divided into several groups, and each group is as an input of sub-systems and the results of all the sub-systems are merged together into the final prediction results. However decision-level fusion ignores the correlation between different feature sets, for example, facial feature set and audio feature set. As for this defect of decision-level fusion, a lot of research of model-level fusion [23]-[28] is proposed to solve this problem, for example, HMM-based fusion [26]-[28], NN-based fusion [24], [29], and BN-based fusion [30].

Comparing these three different fusion strategies, simple accumulating different feature sets as feature-level fusion ignores the different information in time and space for different features. At the same time, finding an appropriate joint feature vectors is still an unsolved problem. For model-level fusion, achieving multi-time-scale labeling multimodal fusion is still an unsolved problem.

In conclusion, despite the limitation of decision-level fusion in correlation between different feature sets, decision-level fusion is chosen by most majority of researchers currently.

III. LEARNING METHODS

A. Support Vector Machine

As we know, Support vector machines (SVM) [31] is recognized as one of the best learning methods for the databases in general cases in these years. Sequential Minimal Optimization – SMO is used to train SVM. The memory requirement for SMO is linear in the size of training data set. The computation time of SMO depends on SVM evaluation, thus SMO is fastest for linear SVM and data sets with large sparsity.

B. Deep Boltzmann Machine

Shallow model need to rely on artificial experience to select features of samples, the inputs of the model are these features that have been selected, the model is only responsible for classification and prediction. For the shallow model, the pros and cons of the model are not usually the most important, but the pros and cons of selecting features are. Therefore, most researchers are devoted into the feature

extraction and feature selection, it not only need us to have a deep understanding on the domain of the thesis, but also spend a lot of time to explore repeated experiments, which also limits the performance of the shallow model.

C. Fusion Method

In order to improve the performance of system, a simple fusion method is applied. The details of simple fusion can be found in [37]. Weights are assigned to a few sub-systems in training process, and the weighted combination of sub-systems results is the final result. For classification, the probability of a sample in each class is combined from different sub-systems. And the formula of simple fusion is described as follows:

$$s(t) = \sum_{i=1}^{l} \alpha_i * s_i(t) \quad (1)$$

where l is the number of subsystems, $s_i(t)$ is the prediction results of ith subsystem, a is the weight to be optimized, and is the final prediction results.

IV. Prediction System on Emotional Facial Feature Data

The emotional facial feature data is from FERA 2015 (Second Facial Expression Recognition and Analysis Challenge). And the data onto implementation is from SEMAINE [32] database, which is used for the classification of Occurrence Detection Sub-Challenge to detect the occurrence of Action Unit (AU). The Matthews Correlation Coefficient (MCC) [33] is used to represent inter-coder reliability. Only AUs where MCC is larger than 0.6 can be selected, as a result, AU2, AU12, AU17, AU25, AU28 and AU45 are selected for implementation from SEMAINE database.

For feature extraction, both geometric features and appearance features are used. The appearance features are extracted by LGBP, and it is generated by applying LBP to Gabor feature.

Classifier using different machine learning is implemented for the classification problem Occurrence Detection, which is described in Fig. 1.

Firstly, the features of training data are used to train the prediction model of SVC. Besides, test data is inputted to the prediction model. Finally, the prediction results of testing data can be obtained.

The detection of occurrence status of the five AUs (AU2, AU12, AU17, AU25, AU28 and AU45) is a binary classification problem. Specifically, label 0 indicates this AU didn't occur in this frame sample, and label 1 indicates this AU occurred in this frame sample. And this classification system is implemented several times for different AUs.

Besides, the kernel function of SVC is set as polynomial kernel. For parameters of kernel function, the polynomial kernel is set as 0.001 and the cost of c-SVC is set as 100 for implementation.

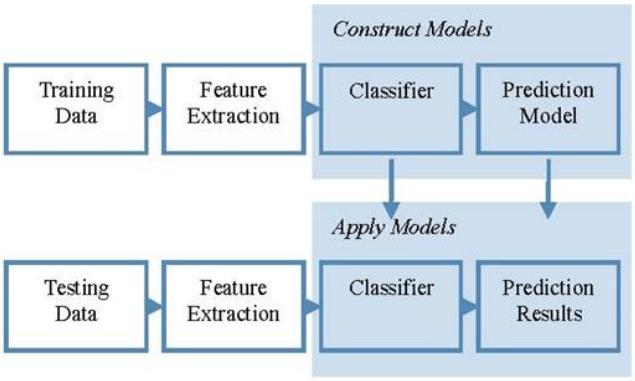

Figure 1. Emotional facial classification system

In SEMAINE database, 3000 frames are manually selected from the whole video, with more balanced classification results. The video of rec1, rec14, rec19, rec23, rec25, rec37, rec39, rec43, rec48, rec54 is selected for training process, and the video of rec13, rec15, rec20, rec24, rec38, rec42, rec46, rec49, rec51, rec53, rec55 is selected for testing process. As a result, 30000 frames are used for training, and 33000 frames are for testing. Considering the time of implementation, both the training data set and testing data set are randomly selected 10% of frame samples to be the input of the prediction system.

Similarly, the method of Deep Boltzmann Machines is also implemented for the classification problem Occurrence Detection. And the construction of prediction system with DBM is the same with the system described in Fig. 1. The original code of DBM is from K. Poon-Feng et al. [34], which construct a two layer DBM.

In order to design an appropriate classification system for facial emotion recognition, we propose a classification system as shown in Fig. 1.

There are two stages for classification: training and testing. During the training process, the features are extracted from training data set, the classifier is used to train model, e.g., SVM or DBM classifier. During the testing process, the features are extracted from testing data set, the trained models either by SVM or DBMs are used to predict the value that is used to identify the class of the test data.

For simple fusion, the information of SVC result and DBM result are merged together to get a better prediction results [37].

V. Implementation and Results

A. Evaluation Criteria

Here the evaluation criteria are illustrated in two-class cases. The TP, FN, TN, FP [38] are defined in Table I.

TABLE I. Evaluation Criteria of Two-Classes Cases

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

Then for the precision and recall of first class, we have:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

For emotional facial classification problem, the evaluation criterion of AU occurrence is the F1-measure, which is the combination of recall and precision. For an AU with precision *P* and recall *R*, it is calculated as:

$$F_1 = \frac{2PR}{P + R} \tag{4}$$

B. Results of Emotional Facial Data

The results of classification system of Occurrence Detection of AUs with SVM, DBM and fusion is shown in Table II.

TABLE II. CLASSIFICATION RESULTS OF OCCURRENCE DETECTION OF AUS WITH SVCAND DBM

	Precision	Recall	F1	UAR	Accuracy
Average of SVC	0.456	0.191	0.237	0.573	0.857
Average of DBM	0.103	0.025	0.030	0.504	0.893
Fusion	0.481	0.236	0.359	0.602	0.910
Var of SVC	0.044	0.026	0.035	0.005	0.012
Var of DBM	0.007	0.002	0.002	0.00008	0.008
Var of Fusion	0.052	0.032	0.040	0.008	0.016

The experiment performance of some AUs is much lower than others. It is because that the frame samples are randomly selected from all video frames and the distribution of dataset is extremely unbalanced. The information selected from the whole dataset is not enough to learn the template from the video of different volunteers. For future work, the distribution of dataset need to be further improved and dimension reduction of PCA might also need to be implemented for processing more information of the dataset.

As the precision of recall computed by class of label 1, when there is extremely little label 1 in training dataset, the learning machine cannot learn the characteristics of this class. And also, we can see from the different experiment results between SVC and DBM in this challenge, SVM is more robust for extremely unbalanced data. The fusion result is better than any of classification result by any method (SVM or DBM).

VI. CONCLUSION AND FUTURE WORK

This report has implemented several learning methods: SVM and DBM, which are all excellent methods in general and the aim is to construct the prediction system which is most suitable to the challenge. Comparing the experiment results of different prediction systems, the best performance of Occurrence Detection of AUs is obtained by emotional facial classification system with SVM. This paper mainly focuses on the implementation and caparison of different learning methods. For future work, the prediction results of facial data and audio data should be merged, and also the prediction results of different learning methods should also be merged to enhance the performance of the prediction system. Besides, there are numerous decision fusion methods, and only the simple fusion is implemented in this paper. For future work, more fusion methods need to be learned for their principles, and more fusion methods need to be implemented and compared.

REFERENCES

[1] R. Picard. Affective computing. MIT Press, 1997.

[2] Arman Savran, Ruben Gur, Ragini Verma. Automatic detection of emotion valence on faces using consumer depth cameras. In IEEE (ICCV'13), 2013.K.

[3] Yale Song, Louis-Philippe Morency, and Randall Davis. Learning a sparse code-book of facial and body micro-expressions for emotion recognition. In Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI '13), pages 237-244, ACM, New York, NY, USA, 2013.

[4] Zhihong Zeng, Maja Pantic, Glenn I.Roisman. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, VOL. 31, NO. 1, January 2009.

[5] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold Based Analysis of Facial Expression," J. Image and Vision Computing, vol. 24, no. 6, pp. 605-614, 2006.

[6] M.S. Bartlett, G. Littlewort, P. Braathen, T.J. Sejnowski, and J.R. Movellan, "A Prototype for Automatic Recognition of Spontaneous Facial Actions," Advances in Neural Information Processing Systems, vol. 15, pp. 1271-1278, 2003.

[7] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05), pp. 568-573, 2005.

[8] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06), pp. 223-230, 2006.

[9] G.C. Littlewort, M.S. Bartlett, and K. Lee, "Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain," Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07), pp. 15-21, 2007.

[10] G. Guo and C.R. Dyer, "Learning from Examples in the Small Sample Case: Face Expression Recognition," IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 35, no. 3, pp. 477-488, 2005.

[11] J. Whitehill and C.W. Omlin, "Haar Features for FACS AU Recognition," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06), pp. 217-222, 2006.

[12] K. Anderson and P.W. McOwan, "A Real-Time Automated System for Recognition of Human Facial Expressions," IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 36, no. 1, pp. 96-105, 2006.

[13] M. Valstar, M. Pantic, and I. Patras, "Motion History for Facial Action Detection from Face Video," Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04), vol. 1, pp. 635-640, 2004.

[14] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form

[15] Face Profile Image Sequences," IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 36, no. 2, pp. 433-449, 2006.

[16] H. Meng and N. Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov

models. In Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II, ACII'11, pages 378–87, Berlin, Heidelberg, 2011. Springer-Verlag.

[17] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. Image Vision Comput., 31(2):153–163, 2013.

[18] A. Sayedelahl, P. Fewzee, M. S. Kamel, and F. Karray. Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In Proceedings of the 4th international conference on Affective computing and intelligent interaction-Volume Part II, ACII'11, pages 407–14, Berlin, Heidelberg, 2011. Springer-Verlag.

[19] A. C. Cruz, B. Bhanu, and S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In ACII (2), pages 341–350, 2011.

[20] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 42(4):993–1005, 2012.

[21] Breiman Leo, Random Forests, Machine Learning 45 (1): 5–32,2001.

[22] Geoffery E. Hinton, Salakhutdinov RR. Reducing the dimensionality of data with neural networks, Science, 313(5786):504-7, 2006.

[23] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In Proceedings of the fourth IEEE International conference on automatic face and gesture recognition (FG'00), pages 46–53, Grenoble, France, 2000.

[24] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Paouzaiou, and K. Karpouzis, "Modeling aturalistic Affective States via Facial and Vocal Expression Recognition," Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06), pp. 146-154, 2006.

[25] F. Fragopanagos and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," Neural Networks, vol. 18, pp. 389- 405, 2005.

[26] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues," Proc. 18th Int'l Conf. Pattern Recognition (ICPR '06), pp. 1136-1139, 2006.

[27] M. Song, J. Bu, C. Chen, and N. Li, "Audio-Visual-Based Emotion Recognition: A New Approach," Proc. Int'l Conf. Computer Vision and Pattern Recognition (CVPR '04), pp. 1020-1025, 2004.

[28] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang, "Training Combination Strategy of Multi-Stream Fused Hidden Markov Model for Audio-Visual Affect Recognition," Proc. 14th ACM Int'l Conf. Multimedia (Multimedia '06), pp. 65-68, 2006.

[29] Z. Zeng, J. Tu, P. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, and S. Levinson, "Audio-Visual Affect Recognition through Multi-Stream Fused HMM for HCI," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05), pp. 967- 972, 2005.

[30] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling Naturalistic Affective States via Facial, Vocal, and Bodily Expression Recognition," LNAI 4451, pp. 91-112, 2007.

[31] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues," Proc. 18th Int'l Conf. Pattern Recognition (ICPR '06), pp. 1136-1139, 2006.

[32] SVMCorinna Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20, 1995.

[33] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing, 3:5–17, 2012.

[34] D. M. W. Powers. Evaluation: From precision, recall and f measure to roc, informedness, markedness, and correlation. Journal of Machine Learning Technologies, 2(1):37–63, 2011.

[35] K. Poon-Feng, D.-Y. Huang, M. Dong and H. Li, "Acoustic Emotion Recognition based on Fusion of Multiple Features-Dependent Deep Boltzmann Machines", ISCSLP, 2014.

[36] Schuller, B., Steidl, S., Batliner, A., Hantke, S., Honig, F., ¨ Orozco-Arroyave, J. R., Noth, E., Zhang, Y., and Weninger, ¨ F. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition. In Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany. ISCA, ISCA. 5 pages, 2015.

[37] D.-Y. Huang, Z. Zhang, and S. S. Ge, Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines, Special Issue of Computer and Speech Language on "Broadening the View on Speaker Analysis" 28(2), pp. 392-491, 2014.

[38] Fawcett, Tom (2006), An Introduction to ROC Analysis, Pattern Recognition Letters 27 (8): 861–874, 2005.