Visual/Acoustic Emotion Recognition

Cheng-Yao Chen, Yue-Kai Huang

Department of Electrical Engineering
Princeton University
{chengc, yuehuang}@princeton.edu

Perry Cook

Department of Computer Science
Princeton University
prc@cs.princeton.edu

Abstract

To recognize and understand a person's emotion has been known as one of the most important issue in human-computer interaction. In this paper, we present a multimodal system that supports emotion recognition from both visual and acoustic feature analysis. Our main achievement is that with this bimodal method, we can effectively extend the recognized emotion categories compared to when only visual or acoustic feature analysis works alone. We also show that by carefully cooperating bimodal features, the recognition precision of each emotion category will exceed the limit set up by the single modality, both visual and acoustic. Moreover, we believe our system is closer to real human perception and experience and hence will make emotion recognition closer to practical application in the future.

1. Introduction

Emotion recognition has been recognized as one of the most important non-verbal ways of people to communicate with each other [1]. However, present human-computer interfaces still don't fully utilize emotion feedback to create a more natural environment because the performance of the emotion recognition is still not very robust and far from real life experience.

Beginning in the 1990's, scientists around the world started to pay attention in the field of automatic emotion recognition. There are already several automatic emotion recognition systems based on either visual analysis [2, 3] or acoustic analysis [4, 5, 6]. However, only a few apply both modalities at the same time [7, 8]. Since it is human nature to consider both visual and acoustic feature together to perceive a certain emotion, the trend of cooperating both modalities in computer emotion recognition system to achieve a better performance compared to single modal system seems also inevitable.

Facial expression usually can recognize 6 emotion categories [2] (neutral state, joy, anger, sadness, surprise, and disgust) and acoustic analysis can usually perform 5 [4] (neutral state, joy, anger, sadness, and fear), we believe that the extension to 7 emotion categories (neutral state, joy, anger, sadness, surprise, and disgust) is possible if we can carefully apply both visual and acoustic features.

In Chen et al [9], they also tried to recognize emotion in six categories (excluded the neutral state) with bimodal information. Instead of directly combining the bimodal features together, they used a sequential weighting method to apply mainly one model at a time. However, we believe it will be more practical to real application if we can combine both modalities without any previous knowledge about the timing of the clips. In Busso et al [10], they tested bimodal system in both decision level and feature level, but only in 4 emotion categories (anger, sadness, joy, and neutral state). They compared the effect of different combination methods on each 4 category. They finally reached a neutral conclusion of choices for feature combination. We think it is more general to extend the bimodal system to 7 emotion states first, and compare the performance of different feature combination.

What we propose here is a comprehensive study in how to combine both visual and acoustic features together to extend the capability and performance of emotion recognition when only single modal works alone. This paper is organized as follows. First we explain how we build up our test videos and introduce our methodology for visual, acoustic, and bimodal analysis respectively. Continuing that, we demonstrate and compare the results for each single modal and bimodal system. Finally, we provide a discussion and a brief conclusion and future directions.

2. Methodology

The overall system can be divided into major three parts: facial expression analysis, acoustic analysis, bimodal feature analysis. The system block diagram is shown as Figure 1, and details of each part will be explained in the following subsections.

0-7803-9332-5/05/$20.00 ©2005 IEEE

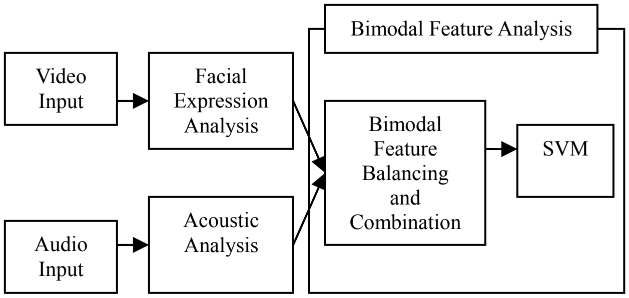
Bimodal Feature Analysis
Video Input
Facial Expression Analysis
Bimodal Feature Balancing and Combination
SVM
Audio Input
Acoustic Analysis

Figure 1. System block diagram

2.1 Facial Expression Analysis

In facial expression analysis, we first applied a facial feature tracking algorithm to track eyes, eyebrows, furrows (both permanent and transient), and lips as shown in Figure 2. After collecting all possible features and interpreting their movements, we employed FACS (facial action coding system) described in [11] to generate our facial feature vectors. We selected 27 features with most discriminant power in emotion recognition to form the final vector matrix. With this feature vector matrix and the ground truth emotion tags, we trained a support vector machine [12] to form the classifier. The above steps are summarized in Figure 3.

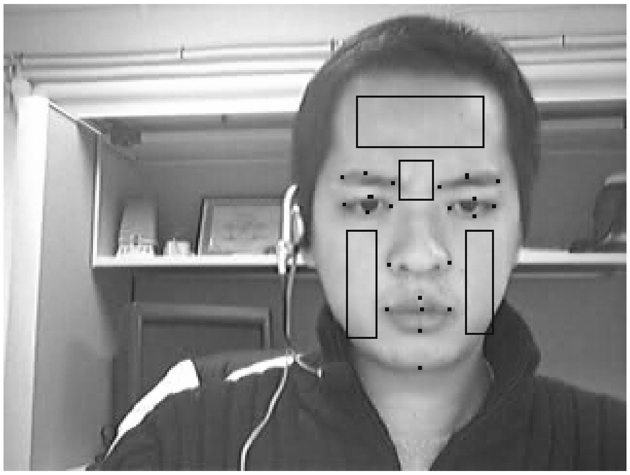

Figure 2. An example for facial feature tracking

Basically, facial expression analysis can be viewed as observing transient changes on face frame by frame to recognize emotions and acoustic analysis on the other hand needs more data to be statistical meaningful. As a result, we summarized the facial features within a period of time as the acoustic features. By doing so, we then had the same dimension of the feature matrix for all our experiments, visual, acoustic, and bimodal. We applied our visual analysis in two cases: one provided only 6 emotion categories from traditional facial expression analysis, and the other provided 7 with an additional but important "fear" state.

2.2 Acoustics Analysis

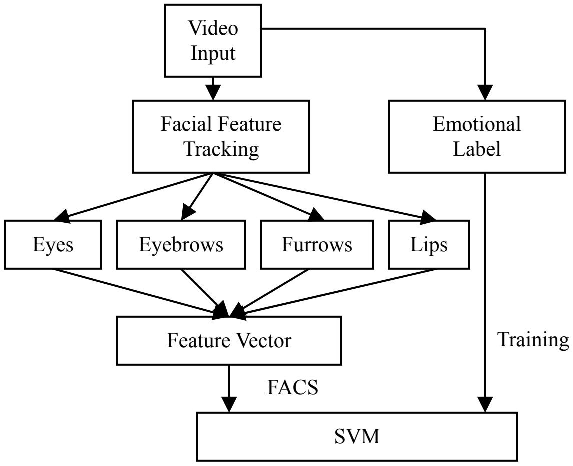

Figure 3. Steps for facial expression analysis

To extract the recognizable features from the recorded vocal data, we selected the standard speech cues which are known as global-level acoustics features. According to a recent study [4], more than 300 preliminary features that are not content-related can be selected for a single vocal passage. These features are mainly measures based on the F0 contours (the fundamental tune, or pitch, in voice) and intensity contours. Among them, 32 features were selected as robust markers of emotions using systematic reduction procedures in that work. Discriminant analysis was done to study and rank the contribution of each marker in emotional classification. In our experiment, we decided to extract the 8 features with the highest rankings discussed in [4] just for simplicity. Those 8 features can further be categorized as three types of information, pitch contour intensity contour, and energy spectrum. Then we extracted those acoustic features and again fed them into the SVM mentioned before. The whole system for acoustic analysis can be shown as Figure 4 in the next page. Again, we applied our acoustic analysis in two cases: one provided only 5 emotion categories from traditional acoustic analysis, and the other provided 7 class with additional "surprise" and "disgust" states.

2.3 Bimodal Feature Analysis

In [10], they also showed that combining in feature level outperformed combining in decision level. In order to explore a closer to human perception experience emotion recognition technique, we decided to incorporate both modalities at the feature level. However, since we had 27 visual features and only 8 acoustics features, we tried two types of feature combinations. In one way, we combined them directly, and thus generated a feature matrix with 35 features. In the other way, we tried to even the strength from both modalities, so we duplicated acoustics features to three times as compared to the original size. As a result, we had 27 visual features and 24 acoustic features, and we tested both ways to see whether the relative size of the feature from different modalities would affect the performance of emotion recognition.

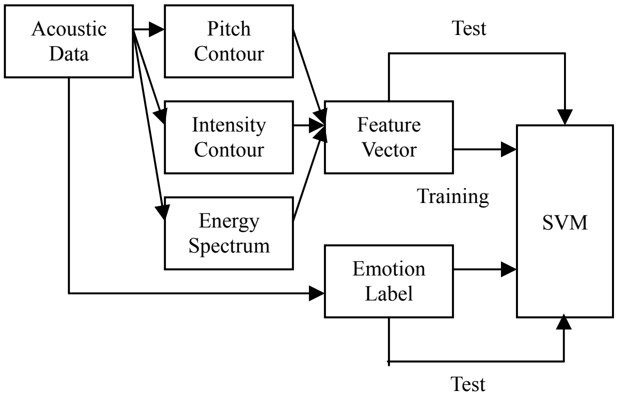

Figure 4. Blocks for acoustic analysis

3. Experiment Results and Discussion

The video/audio clips were recorded from two subjects reading 5 different sentences composed of common daily telephone conversations while expressing 7 emotional states with 5 times each, using a commercial digital camera and a headset microphone. Several volunteers were asked to judge the expressed emotion in 8 randomly-selected clips. Surprisingly, human judgment gave only around 50% accuracy. After carefully studying the judgment case by case, we found all errors were caused only by misjudgment between the similar emotional categories. (There are two circumstances: anger vs. surprise, sadness vs. fear vs. neutral). Since the detectable differences were indeed minor and subtle in these judgments, we proceeded forward using the video/audio clips in our system. We can also study how well our system can classify with and without taking these similar categories into consideration. We systematically selected 80% of the testing clips as train samples and 20% as test sample each time in every emotion category for our cross-validation. We then averaged all the cross-validation in order to have no favor in any of the single test and train set. The performance of all our experiments is summarized in Table 1.

First, we can observe that the average overall performance of visual analysis dropped when we extended the emotion categories from 6 to 7 if we excluded the performance of fear state. The performance of joy category was significantly low compared to other categories. The possible cause is that since joy recognition relies a lot on chin and lips movement, acting joy while speaking the lines at the same time generates a lot of false positives. However, this is inevitable in real life experiences and applications. Moreover, the overall performance is also lower than what were reported in the state-of-art system [13] based on the same reason mentioned above. Thus we hope that by incorporating acoustic features we can recover the loss of performance from acoustic “interference” and the extension of emotion categories.

Table 1. Performance comparison for all experiment setup

	Neutral	Joy	Anger	Surprise	Disgust	Sadness	Fear	Avg
Visual Only for 6	0.78	0.65	0.87	0.77	0.83	0.89	n/a	0.79
Visual Only for 7	0.78	0.63	0.83	0.77	0.76	0.82	0.86	0.75
Acoustic Only for 5	0.83	0.63	0.82	n/a	n/a	0.67	0.70	0.73
Acoustic Only for 7	0.67	0.46	0.71	0.69	0.69	0.61	0.54	0.63
Direct Bimodal for 7	0.77	0.73	0.91	0.77	0.86	0.84	0.91	0.82
Balanced Bimodal for 7	0.78	0.69	0.97	0.77	0.86	0.91	0.89	0.84

The overall performance of the acoustic classifier is 64 percent. As observed from our experiment, some pairs of emotions were mutually confused more. Sadness was highly confused with fear (17.8 % and 15.6 % of mutually misclassification rate, respectively), while anger was highly confused with surprise (8.9 % and 18.9 %, respectively). These results were the same as what we had predicted in the previous human judgment trial. Note that there is a huge performance decrease when we extended the emotion categories from 5 to 7.

When comparing the performance of direct and balanced feature combination, there is only a slight difference. However, bimodal system truly outperformed both visual and acoustic analysis when they acted alone. Almost all categories experienced a performance increase. We can see that the performance increase from acoustic analysis only is significant. The confusion matrix comparison for visual only, acoustic only and bimodal system (with balanced bimodal features) is shown in Figure 5 in the next page.

The balanced combination only has a slightly higher average recognition rate. This suggests that the key for performance enhancement in feature combination of emotion recognition may lie in a better feature selection. Since we just selected best visual and acoustic features from each modality without any consideration to each other, a more careful bimodal feature selection with more statistical and psychological support will be a good direction for future attempts.

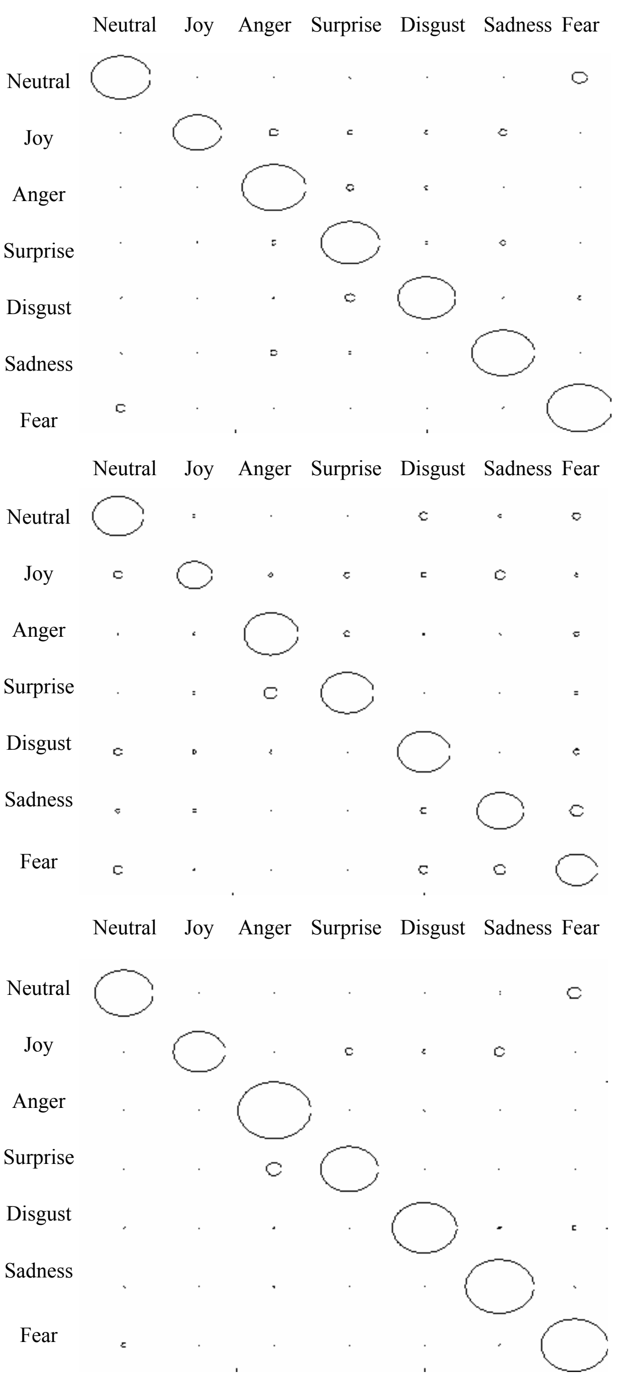

Figure 5. Confusion matrix comparison for visual only, acoustic only and bimodal (balanced) system (from top to the bottom).

4. Conclusion

In this research, we successfully showed that it is feasible and useful to incorporate both visual and acoustic feature analysis to recognize human emotions. With that combination we can also effectively extend the number of recognized emotion categories and increase the performance limit for each category when compared with only single modes. The methodology we generated our test video/audio clips and feature combination is closer to real-life experience and will be more natural to apply in future human computer interface. Our experiments also show that for future improvements to the performance of bimodal emotion recognition lie in better feature selection which can better incorporate both modalities.

5. Reference

[1] Cunningham MR, "Personality and the Structure of Nonverbal Communication of Emotion", *Journal of Perception*, Dec 1977; 45(4): 564-584.

[2] J. J. Lien et al. "Automated Facial Expression Recognition", Proc. of the Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 390-395.

[3] K. Mase, "Recognition of Facial Expression from Optical Flow", IEICE Transc., E., Oct. 1991, 74(10):3474-3483.

[4] S. McGilloway et al, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", Proceedings of the ISCA workshop on Speech and Emotion, 2000, pp. 207-212.

[5] Solbin C. and Alpert M. "Emotion in Speech: the Acoustic Attributes of Fear, Anger, Sadness, and Joy" *Journal of Psycholinguist Res.* Jul 1999, 28(4): 347-65.

[6] V. A. Petrushin, "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application", Proc. of Sixth International Conference on Spoken Language Processing, 2000.

[7] De Silva et al, "Bimodal Emotion Recognition", Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, March 2000, pp 332-335.

[8] J.F. Cohn and G.S. Katz, "Bimodal expression of emotion by face and voice", In Workshop on Face/Gesture Recognition and Their Applications, The Sixth ACM International Multimedia Conference, Bristol, England, 1998, pp. 41-44.

[9] L.S. Chen et al, "Emotion Expressions in Audiovisual Human Computer Interaction," Proc. of IEEE International Conference on Multimedia & Expo 2000, pp. 423- 426.

[10] Busso et al, "Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information", Proc. of the ACM International Conference on Multimodal Interfaces, 2004 pp. 205-211.

[11] P. Ekman and W.V. Friesen, *Facial Action Coding System*, Palo Alto: Consulting Psychologist Press, 1978.

[12] C.-W. Hsu and C.-J. Lin. "A Comparison of Methods for Multi-Class Support Vector Machines", *IEEE Trans. on Neural Networks*, 13(2002), pp. 415-425.

[13] Phillip Michel and Rana El Kaliouby, "Real Time Facial Expression Recognition in Video using Support Vector Machine", Proc. of ACM International Conference on Multimodal Interfaces,2003, pp. 258-264.