Bimodal Emotion Recognition

Liyanage C. De Silva, Pei Chi Ng
The National University of Singapore
Department of Electrical Engineering
10 Kent Ridge Crescent, Singapore 119260
elelcds@nus.edu.sg, engp9794@nus.edu.sg

Abstract

This paper describes the use of statistical techniques and Hidden Markov Models (HMM) in the recognition of emotions. The method aims to classify 6 basic emotions (angry, dislike, fear, happy, sad and surprise [4]) from both facial expressions (video) and emotional speech (audio). The emotions of 2 human subjects were recorded and analyzed. The findings show that the audio and video information can be combined using a rule-based system to improve the recognition rate.

1. Introduction

With the coming of the computer age, human interactions with and through the computer are increasingly frequent. Examples include chat rooms on the Internet, online banking services and online tutorials. However, communication will only truly be possible if we are able to "see" the other party's responses.

Sending compressed video and audio on the Internet will place a huge burden on the network. However, if it is possible for the computer to recognize our emotions using a real-time emotion recognition device, to transfer this message across Internet to the other party will be trivial. With this data, regeneration of emotions will be possible at the receiver's end.

De Silva et al. [10] studied human subject's ability to recognize emotions from viewing video clips of facial expressions and listening to the corresponding speech audio clips. The result shows that the subjects under study could recognize anger, dislike, happy and surprise better from the video information while the emotions sadness and fear were better recognized from audio information. As a whole however, the human subjects achieved an average success rate of 47% in recognizing the emotions.

In this paper, we consider the problem of bimodal emotion recognition of the 6 basic emotions in three phases. In the video data phase, each image sequence is characterized by a pair of velocity and displacement feature vectors. Anandan's optical flow algorithm [1] is applied to track the feature movements from an image sequence. A facial expression recognition system developed by Chen [2] is used to classify the video data. For the audio phase, the pitch parameters are extracted the super resolution pitch determination algorithm [7] and trained using the HMM training function in the Speech Signal Processing and Recognition MATLAB Toolbox. In the final phase, a rule-based method is used to combine and classify the results from both the video and audio modes of data.

2. Emotion Database

It is desirable to obtain samples of speech and expressions as close as possible to the natural emotions. But due to the fact that there are no such freely available databases currently, we have to construct an experimental database for this project. Each subject was asked to perform 6 emotional outbursts for each basic emotion: anger, dislike, fear, happy, sad, surprise, and to repeat each emotion 12 times. Altogether, there are 144 image sequences and 144 audio files recorded from 2 subjects in the database.

2.1. Video Image Sequences

The image sequences were captured using a video camera connected to the computer. The monochrome images, of size 384 by 288 pixels are taken at the rate of approximately 20 frames per second. Each image sequence consists of 40 images, showing an emotional outburst of about 2 seconds. Each emotional outburst begins from a neutral expression to the emotion portrayed before ending in a neutral expression again. An example of an emotion sequence is shown in Figure 1

Figure 1. An example of the Angry emotion image sequence.

2.2. Audio Speech Files

During the emotional outburst, the subject is allowed to speak only one English word of his choice. The sound files contain only one utterance and are in windows canonical wave format, taken simultaneously with the image sequences. Each sound file is approximately 1-1.5 seconds long, recorded using a mouth piece microphone.

2.3. Observations of the subjects

Both subjects found that they could portray some emotions better than others, and the speech outbursts depend on the current mood of the speaker. Furthermore, subjects find it easier to portray an emotion if appropriate word is chosen to express their feelings. For emotions like fear and dislike, the subjects would shake their heads unconsciously, even though they have been told explicitly not to do so. The "shaking-head" gesture is part of an emotional expression even though they may not notice it. We also noticed that the two subjects voiced their emotions differently. The loudness and pitch that the subjects spoke with varies noticeably, even though both were male. Thus a user dependent characteristic profile would be required to classify the emotions of each individual, rather than a general profile for everyone.

3. Processing The Database

The video sequences and audio files were processed separately. Feature vectors were extracted from the video and audio data. The video samples were classified using nearest neighbor method while the audio samples were classified using HMM.

3.1. Video Data

The Anandan's optical flow algorithm [1] is used to track the edge movements. Two types of feature vectors, can be obtained using Chen's facial expression system [2], the maximum displacement and maximum velocity of the feature points. The six feature points, as shown in Figure 2 are:

- The vertical displacements (or velocities) or the upper and lower lips.
- The horizontal displacements (or velocities) of the left and right mouth corners
- The vertical displacements (or velocities) of the left and right inner eyebrows.

The results from the distance and velocity feature vectors are combined using weights giving:

$$d_{COM} = \alpha \times d_{DIS} + (1 - \alpha) \times d_{VEL} \qquad (1)$$

where d_{DIS}, d_{VEL} are the Euclidean distance for the distance and velocity feature vectors respectively and α is obtained from experimental results.

From the experiment, it was found that a value of $\alpha = 0.2$ for subject A and $\alpha = 0.3$ for subject B gave the highest success rates.

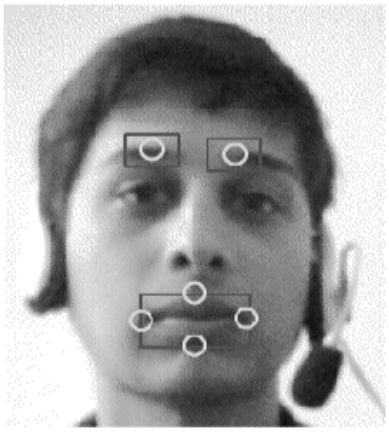

Figure 2. Picture of subject in the neutral expression with the 6 feature points indicated on the face.

3.2. Audio Data

The audio waveform is divided into frames of 0.01s each and the the super resolution pitch determination algorithm [7] is used to to extract the pitch values. An example

of the pitch contour of an emotion audio file is shown in Figure 3.

Two samples for each emotion are used for training. The pitch contours are fed into a HMM machine using a left-right model to train the emotion models, which are then used to recognize the rest of the audio samples.

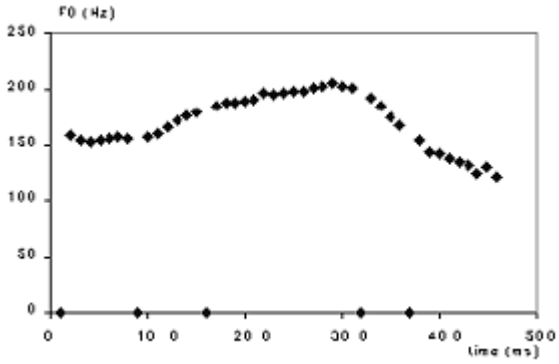

Figure 3. Example of the pitch contour for a fear emotion.

3.3. Classification Method

A rule-based method is used combine the audio and video information to give an emotion output. For example if a sample has been classified as angry by both the audio and video processing methods, then the final output will be the emotion angry. For samples that has been classified differently by the audio and video processing methods, the dominant mode will be used as the emotion classification.

4. Results

4.1. Video and Audio Classification

The results from video classifications are plotted against the audio classification to yield a graph for each subject as shown in Figure 4 and Figure 5. From the graphs we notice that the results for each subject are very individualized. For example, the anger audio samples are recognized correctly for subject A most of the time, but frequently misclassified as the emotion surprise for subject B. The video samples perform better on the overall, whereas the audio samples are dominated by certain emotions like fear and dislike for subject A, surprise and dislike for subject B. This results in emotions like anger being misclassified as surprise, or sad being misclassified as dislike consistently.

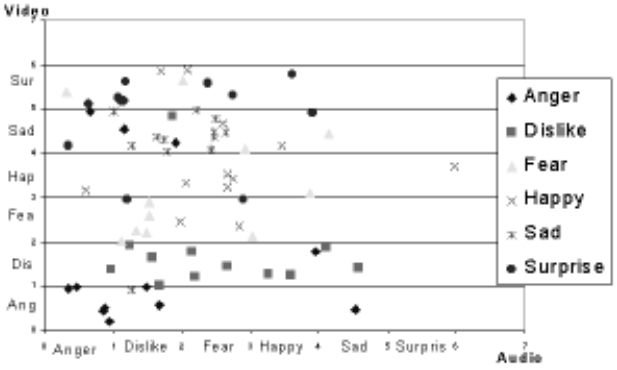

Figure 4. Subject A: Plot of Video emotion output vs Audio emotion output.

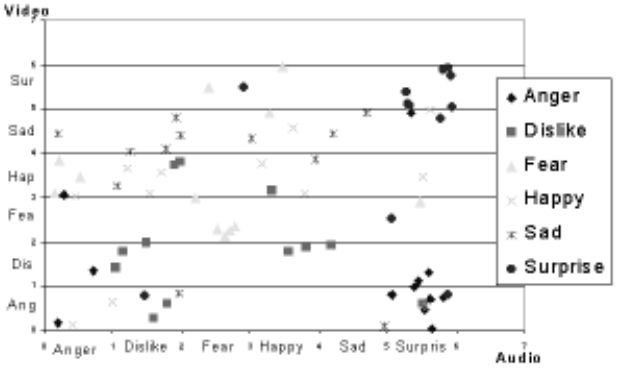

Figure 5. Subject B: Plot of Video emotion output vs Audio emotion output.

4.2. Bimodal Emotion Classification

From the graphs a rule-based system is used to output the emotion using both the audio and video data. However, with a generalized rule-based system, there are some samples that cannot be classified as any of the six emotions and these are known as the unclassified samples. These unclassified samples make up about 10% of the total sample space. This is one problem faced during the database acquisition process. Good actors are hard to come by and some of the emotions in these samples are not very well expressed. With either video or audio alone, deciding which samples are invalid requires much human intervention. Using both modes of data and certain rules, the system can automatically decide which samples to discard. After excluding these sam-

ples from the sample space, the bimodal recognition rate is about 72%.

The overall results are shown in Figure 6. The success rates using both video and audio information are better than that using either mode alone.

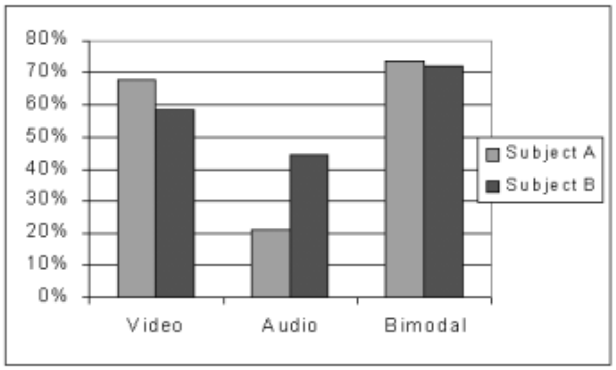

Figure 6. Overall Results in Emotion Recognition.

5. Conclusion

We have attempted to improve emotion recognition by categorizing emotions using both video and audio information using a rule-based algorithm. On the overall, the classification using video data performs better than the audio data, while the bimodal method yields better results than either video or audio modes alone. To further improve on emotion recognition in using audio information, other parameters like cepstral coefficients may be used as feature vectors to train the HMM. Future work involves using neural networks rather than the rules to train the system to recognize emotions. A larger database involving more subjects will also be gathered for more conclusive studies.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *Int. Journal of Computer Vision*, 1989.

[2] H. B. Chen. *Detection and transmission of facial expression for low speed web-based teaching*. Thesis for Degree of Bachelor of Engineering, National University of Singapore, 1998.

[3] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion / expression recognition. *IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 1998.

[4] P. Ekman. Strong evidence for universals in facial expressions: A reply to russel's mistaken critique. *Psychological Bulletin*, 115(2), 1994.

[5] A. Fridlund. *Human Facial Expression: An evaluationary view*. Academic Press, New York, 1994.

[6] G. Klasmeyer and W. F. Sendlmeier. Objective voice parameters to characterise the emotional content in speech. *Psychological Bulletin*, 1, 1995.

[7] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, 39(1):40–48, 1991.

[8] P. W. Picard and G. Cosier. Affective intelligence - the missing link? *BT Technical Journal*, 14(4), October 1997.

[9] K. R. Scherer. How emotion is expressed in speech and singing. *ICPhS 95 Stockholm*, 3, 1995.

[10] L. C. D. Silva, T. Miyasato, and R. Nakatsu. Use of multimodal information in facial emotion recognition. *IEICE Trans. Inf & Syst*, E81-D(1), January 1998.