International Conference on
Information, Communications and Signal Processing
ICICS '97
Singapore, 9-12 September 1997

Facial Emotion Recognition Using Multi-modal Information

Liyanage C. DE SILVA *[1] Tsutomu MIYASATO ** Ryohei NAKATSU **

* Dept. of Electrical Engineering
The National University of Singapore
10 Kent Ridge Crescent, Singapore 119260
Republic of Singapore
Email: elelcds@leonis.nus.sg
** Advanced Telecommunications Research Center (ATR)
Media Integration and Communications (MIC) Research Laboratories
Hikari-dai 2-2, Soraku-gun, Kyoto-fu 619-02,
Japan
Email:(miyasato/nakatsu)@mic.atr.co.jp

Abstract

Facial emotion recognition will become vitally important in future multi-cultural visual communication systems, for emotion translation between cultures, which may be considered analogous to speech translation. However so far the recognition of facial emotions is mainly addressed by computer vision researchers, based on facial display. Also detection of vocal expressions of emotions can be found in research work done by acoustic researchers. Most of these research paradigms are devoted purely to visual or purely to auditory human emotion detection. However we found that it is very interesting to consider both these auditory and visual information together, for processing, since we hope this kind of multi-modal information processing will become a datum of information processing in future multimedia era.

By several intensive subjective evaluation studies we found that human beings recognise Anger, happiness, Surprise and Dislike by their visual appearance, compared to voice only detection. When the audio track of each emotion clip is dubbed with a different type of auditory emotional expression, still Anger, Happiness and Surprise were video dominant. However Dislike emotion gave mixed responses to different speakers. In both studies we found that Sadness and Fear emotions were audio dominant. As a conclusion to the paper we propose a method of facial emotion detection by using a hybrid approach, which uses multi-modal information for facial emotion recognition.

1. Introduction

In this paper our main focus is on studying the effect of integration of audio and video in efficient facial expression recognition. We believe the results of this kind of study would help in machine recognition of facial expressions by correctly integrating audio and video data (De Silva et. al. [1]).

It is very important to recognise emotions of speakers apart from facial display (facial expression) to build a good interface for human-human communication via a computer generated agent. As humans we process auditory and visual information in parallel to recognise emotions. . Majority of studies in the area of facial emotion recognition by machines, treated these two vital sources of information separately.

First let us consider some research work in the area of visual information used in emotion detection. Coren and Russel [2] stated that the dominance of one expression over another (when presented stereoscopically - stereo scopic perceptual conflict) as the most common result, but some basic emotions (happiness, fear, etc.) failed to dominate some of the non basic emotions (excitement, calm). According to Ekman and Friesen [3] facial expression of just one "pure" emotion is a rare occurrence. If two conflicting emotional expressions are presented to an observer, the general tendency is toward one of the stimuli establishing perceptual dominance over the other [2].
In the paper by Kobayashi et. al [5], they have presented an approach of using neural network in visual emotion detection of 6 basic emotions. But in their paper they failed to represent emotion detection accuracy based on each individual emotion. Several other facial expression detection systems using visual

[1] The author was with ATR Media Integration and Communications Research Laboratories, Kyoto-fu 619-02 Japan till March 1997.

0-7803-3676-3/97/$10.00 © 1997 IEEE

information can be found in [8][9][10][11].

In the papers De Silva et. al. [12][13] stated the advantage of emotion enhancement in effective emotion transmission. Further they have stated the merits of incorporation of audio in detecting correct emotional state.

In the next section we describe the procedure of our evaluation experiment. Then in the following section a description of the results of the subjective evaluation experiment carried out in our research is given. Thereafter, system implementation and then conclusions to the paper is presented.

2. Procedure of the Experiment

In this section an overview of the whole experiment set up is presented. At first we discuss the method of data preparation and then a brief description of the evaluation procedure is given.

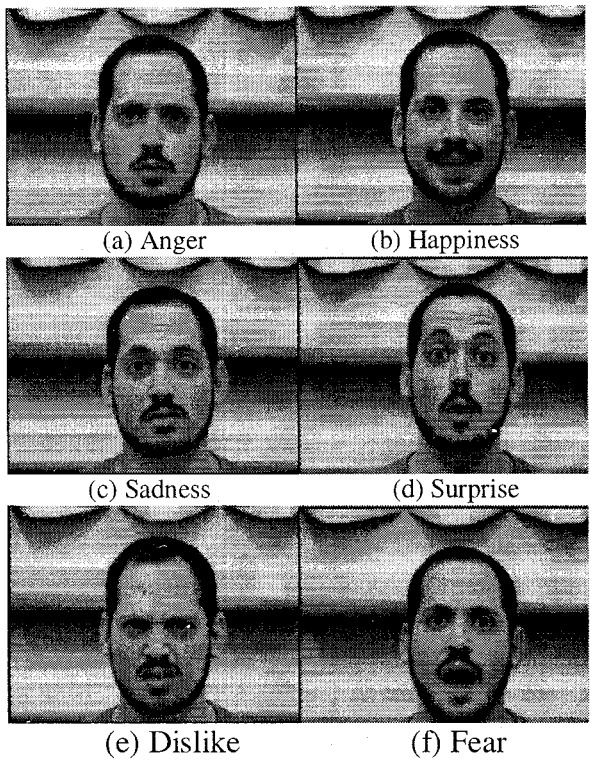

Figure 1 Some Images Taken from the Video Clips of the 6 Facial Emotions used in the Experiment

In this research we have selected two different speakers, who speak a language which is totally in comprehensible by the subjects. The first speaker selected speaks Spanish Language and the other speaker selected speaks Sinhala Language (a language spoken in Sri Lanka). Each speaker has been asked to make 36 different emotional sentences in front of a camera and their emotions are recorded. Then the recorded image sequences are edited in order to have 36 different emotional video clips of approximately the same length. (6 emotion sets with each set consists of 6 different emotional sentences: Angry, Happy, Sad, Surprise, Dislike, and Fear). Some images taken from the video clips are shown in Figure 1.

3. Experiment A – Original Emotions

In this experiment we have showed the recorded emotional video clips to the subjects, and asked their consent for each clip. The evaluation is carried out with 3 different types of media clips:

(a) audio information of the emotions only;
(b) video information of the emotions only;
(c) both audio and video information (original video clip).

In Figure 2, the cumulative correct percentage response of 18 subjects for 36 different emotion clips performed by a Spanish Language native speaker is shown. In Figure , the same for a Sinhala Language native speaker is shown.

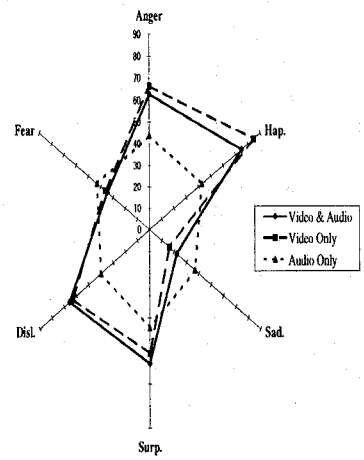

Figure 2 Cumulative Correct Percentage Response Obtained for the Spanish Language Speaker

Then we calculate the video and audio dominance for each type of emotion and the results are shown in the Fig. 4.

From the results of this subjective evaluation studies it can be noted that human beings recognize Anger, happiness, Surprise and Dislike by their visual appearance, compared to voice only detection. (Percentage of video dominance for each emotion are: 22%, 42%, 12% and 23% respectively for Spanish

Language Speaker and 12%, 8%, 2%, and 9% respectively for Sinhala Language Speaker.)

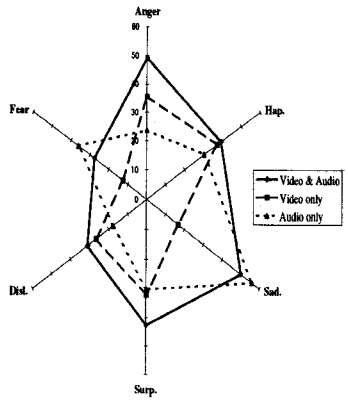

Figure 3 Cumulative Correct Percentage Response Obtained for the Sinhala Language Speaker

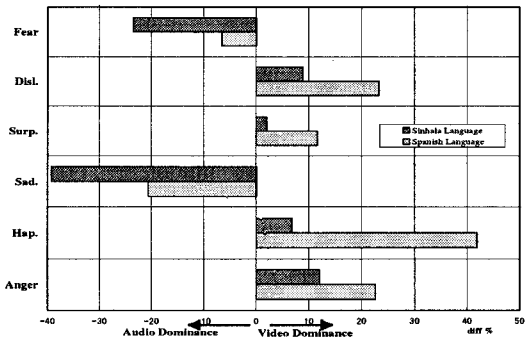

Figure 4 Video and Audio Dominance for the first Experiment

4. Experiment B – Mixed Emotions

A separate experiment was carried out (results are shown in Figure 5) in which the audio track of each emotion clip is dubbed with a different type of auditory emotional expression. From the results we still found Anger, Happiness and Surprise were video dominant. (Percentage of video dominance for each emotion are: 27%, 67%, and 13% respectively for Spanish Language Speaker and 1%, 30%, and 12% respectively for Sinhala Language Speaker.) However Dislike emotion gave mixed responses to different speakers. (19% more video dominant for Spanish Language Speaker and 11% more audio dominant for Sinhala Language Speaker) In both these studies we found that Sadness and Fear emotions were audio dominant. (Percentage of audio dominance for each emotion in the first experiment: 38% and 23% respectively for Spanish Language Speaker and 21% and 8% respectively for Sinhala Language Speaker. Percentage of audio dominance for each emotion in the second experiment: 16% and 23% respectively for Spanish Language Speaker and 1% and 9% respectively for Sinhala Language Speaker.

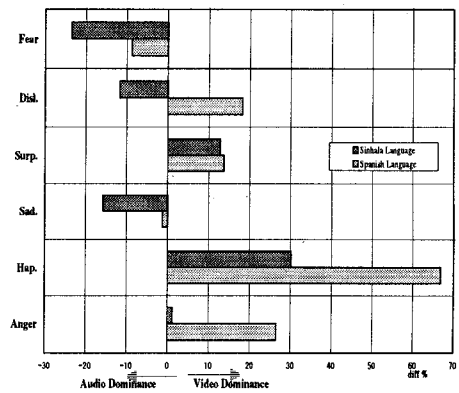

Figure 5 Video and Audio Dominance for the second Experiment

5. Emotion Mis-Classification Analysis

In this section we discuss some common emotion mis-classification errors for different speakers of different cultural backgrounds. Table 1 and 2 summarises them. One of the clear observations from Table 1 is that the percentage correct detection of a positive emotion (happiness and surprise) is higher for both audio and video based cases. However we observed conflicting results for Sinhala language video based case.

Table 1 Emotion Mis-classification table for Original Emotions

Type of Emotion Clip	Percentage of a Negative Emotion Detected as Negative	Percentage of a Positive Emotion Detected as Positive
Audio only (Spanish)	61.43%	72.19%
Video only (Sinhala)	73.29%	87.44%
Audio only (Spanish)	50.17%	56.16%
Video only (Sinhala)	75.66%	63.69%

Also according to Table 2 we can see that positive (non-negative) facial expressions are less effected by a opposite kind of auditory emotion for Spanish Language Speaker.

However the results obtained for Sinhala language speaker is contrary to the above.

This contradiction may be due to mis-interpretation of dislike emotion or may be merely a cultural effect. A further study is needed to confirm this fact.

Table 2 Emotion Mis-classification table for Mixed Emotions

Type of Audio Clip	Percentage of a Negative Emotion Detected as Negative	Percentage of a Positive Emotion Detected as Positive
Negative (Spanish)	81.42%	75.95%
Non-negative (Spanish)	58.13%	85.95%
Negative (Sinhala)	69.77%	66.16%
Non-Negative (Sinhala)	73.80%	68.58%

6. System Implementation

We are currently investigating the possibilities of implementing a prototypical system, which incorporates the findings of this research. For this we propose a kind of approach shown in Figure 6. In this figure the inputs to the weighting matrix are the results of emotion detection using available audio and video based facial emotion detection techniques.

$$(V_{Ang}, V_{Hap}, V_{Sad}, V_{Sur}, V_{Dis}, V_{Fea}, A_{Ang}, A_{Hap}, A_{Sad}, A_{Sur}, A_{Dis}, A_{Fea}) \in \{0,1\} \quad (1)$$

Inside the weighting matrix, the following logic is applied for selection of correct emotion, which uses both audio and video based detection results.

$$\begin{aligned} O_{Anger} &= W_{(1,Ang)}V_{Ang} + W_{(2,Ang)}A_{Ang} \\ O_{Happiness} &= W_{(1,Hap)}V_{Hap} + W_{(2,Hap)}A_{Hap} \\ O_{Sadness} &= W_{(1,Sad)}V_{Sad} + W_{(2,Sad)}A_{Sad} \\ O_{Surprise} &= W_{(1,Sur)}V_{Sur} + W_{(2,Sur)}A_{Sur} \\ O_{Dislike} &= W_{(1,Dis)}V_{Dis} + W_{(2,Dis)}A_{Dis} \\ O_{Fear} &= W_{(1,Fea)}V_{Fea} + W_{(2,Fea)}A_{Fea} \end{aligned} \quad (2)$$

For example the weights of the Weighting Matrix can be selected as shown below. These values are obtained using the results of Experiment A explained earlier.

Basically setting of the weights $W_{(1,xxx)}$,$W_{(2,xxx)}$ as shown in the Eqn. (4) will give reliable results compared to the results that would be obtained either by using audio only or by using video only for emotion detection.

Weighting Matrix

$$\begin{aligned} W_{(1,Ang)} &= 22.59 & W_{(2,Ang)} &= 0 \\ W_{(1,Hap)} &= 41.88 & W_{(2,Hap)} &= 0 \\ W_{(1,Sad)} &= 0 & W_{(2,Sad)} &= 20.65 \\ W_{(1,Sur)} &= 11.64 & W_{(2,Sur)} &= 0 \\ W_{(1,Dis)} &= 23.30 & W_{(2,Dis)} &= 0 \\ W_{(1,Fea)} &= 0 & W_{(2,Fea)} &= 6.54 \end{aligned} \quad (3)$$

$$\{W_{(1,Ang)}, W_{(1,Hap)}, W_{(2,Sad)}, W_{(1,Sur)}, W_{(1,Dis)}, W_{(2,Fea)}\} >>> 1$$

$$\{W_{(2,Ang)}, W_{(2,Hap)}, W_{(1,Sad)}, W_{(2,Sur)}, W_{(2,Dis)}, W_{(1,Fea)}\} \leq 1 \quad (4)$$

Finally:

$$Max\{O_{Anger}, O_{Happiness}, O_{Sadness}, O_{Surprise}, O_{Dislike}, O_{Fear}\} \quad (5)$$

could be used to detect the input emotion.

7. Conclusions

In conclusion we can state that human perception of facial emotions can be divided into three categories such as visually dominant emotions, auditory dominant emotions and mixed dominant emotions. However clearly we found that some emotions are strongly visually dominant and some are strongly auditory dominant. We can make use of this factor in efficient facial emotion detection by assigning a weighting function. We are hoping to implement a facial emotion detection system based on these observations, which may be used as a multi-cultural facial emotion translation system.

8. Acknowledgments

Authors wish to thank Mr. Eduardo J. Neeter for his kind co-operation in giving his full support to obtain his facial emotions and also to all the subjects who took part in the subjective evaluation. A special thank should go to Mr. Takahiro Otsuka and Ms. Naoko Tosa for their kind co-operation in the discussions related to using their systems for the final implementation of our multi-modal facial emotion recognition system.

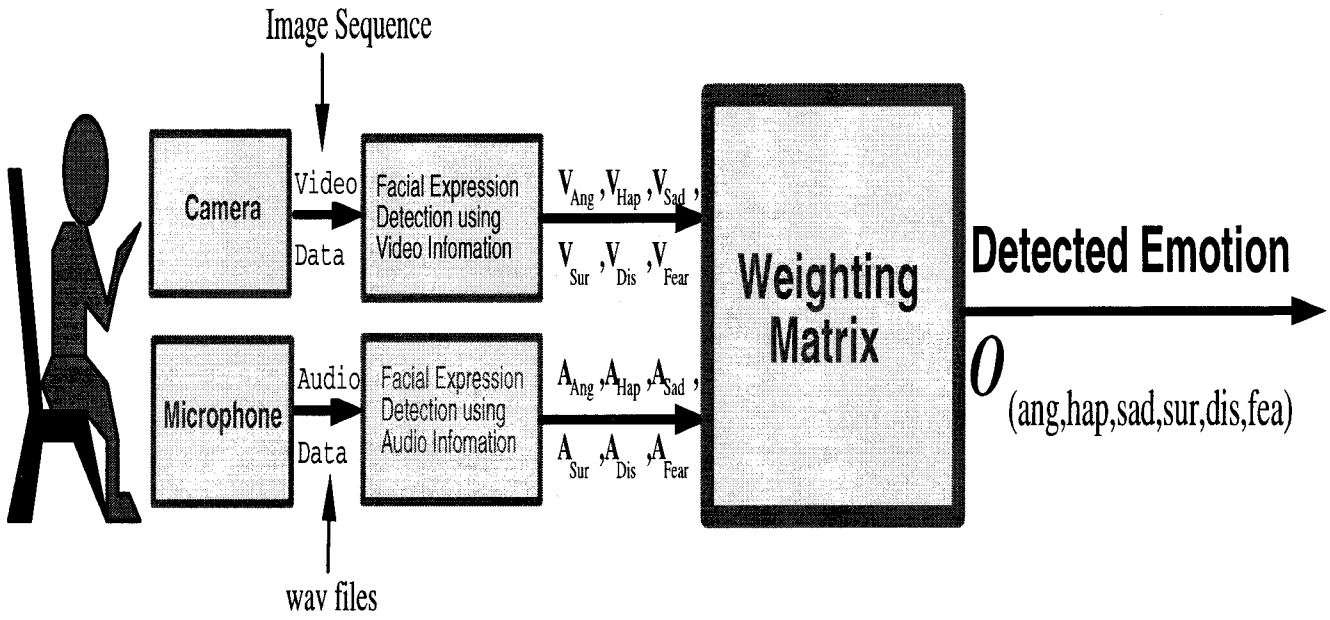

Figure 6 Proposed Emotion Detection Algorithm using Multi-modal Inputs

References

[1] Liyanage C. De Silva and Tsutomu Miyasato, "Degree of human perception of facial emotions based on audio and video information," in Procs. of Joint Workshop on Multimedia Communications (JWMMC'96), Taegu, Korea, pp 1-4-1 to 1-4-7, October 1996.

[2] Stanley Coren and James A. Russel, "The relative dominance of different facial expressions of emotion under conditions of perceptual ambiguity", Cognition and Emotion, Vol. 6, No. 5, pp. 339-356, 1991.

[3] P. Ekman and W. V. Friesen, "Pictures of Facial Effect," Palo Alto, CA: Consulting Psychologists Press, 1976.

[4] Alan~J. Fridlund, "Human facial expression: An evalutionary view,"New York: Academic Press, 1994.

[5] Paul Ekman, "Strong evidence for universals in facial expressions: A reply to russell's mistaken critique," Psychological Bulletin, Vol. 115, No. 2, pp. 268-287, 1994.

[6] James A. Russel, "Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies," Psychological Bulletin, Vol. 115, No. 1, pp. 102-141, 1994.

[7] Hiroshi Kobayashi and Fumio Hara, "Monitoring of facial expressions (in japanese). JSICE Journal of the Society of Instrument and Control Eng.s," Vol. 34, No. 4, pp. 248-254, April 1995.

[8] Kenji Mase, "Recognition of facial expressions from optiacl flow," IEICE Transactions, Vol. E74, No. 10, pp. 3474-3483, October 1991.

[9] Yaser Yacoob and Larry S. Davis, "Recognizing human facial expressions," Technical report, Center for Automation Research, Univ. of Maryland, May 1994.

[10] Irfan A. Essa and Alex Pentland, "Facial expression recognition using a dynamic model and motion energy," In Procs. of Int. Conf. on Computer Vision'95, Cambridge, MA, June 1995.

[11] Tatsumi Sakaguchi, Jun Ohya, and Fumio Kishino, "Facial expression recognition from image sequence using hidden markov model (in Japanese)," Transactions of Television Engineers, Japan, Vol. 49, No. 8, pp. 1060-1067, August 1995.

[12] Liyanage C. De Silva, Tsutomu Miyasato, and Fumio Kishino, "Emotion enhanced face to face meetings using the concept of virtual space teleconferencing," IEICE Trans. on Information Systems, Vol. E79-D, No. 6, pp. 772-780, June 1996.

[13] Liyanage C. De Silva, Tsutomu Miyasato, and Fumio Kishino, "Emotion enhanced multimedia meetings using the concept of virtual space teleconferencing. In Procs. of IEEE Multimedia Systems'96, Hiroshima, Japan," pp. 28-33, June 1996.