Recognition of Emotional States using Voice, Face Image and Thermal Image of Face

Tetsuro Kitazoe, Sung-Ill Kim, Yasunari Yoshitomi, Tatsuhiko Ikeda
Department of Computer Science and Systems Engineering
Faculty of Engineering, Miyazaki University 1-1, Gakuen Kibanadai Nishi, Miyazaki, 889-2192 Japan

E-mail: kitazoe@cs.miyazaki-u.ac.jp

Abstract

A new integration method is presented to recognize the emotional expressions. We attempted to use both voices and facial expressions. For voices, we use such prosodic parameters as pitch signals, energy, and their derivatives, which are trained by Hidden Markov Model (HMM) for recognition. For facial expressions, we use feature parameters from thermal images in addition to visible images, which are trained by neural networks (NN) for recognition. The thermal images are observed by infrared ray which is not influenced by lighting conditions. The total recognition rates show better performance than each performance rate obtained from isolated experiment. The results are compared with the recognition by human questionnaire.

Key words: Emotional Pattern Recognition, Emotional Speech, Prosody, Facial Expression, Thermal Image

1. Introduction

It is useful and perhaps necessary to introduce a bit of emotional taste into the course of communications between human and robots. Our future society will be more enjoyable if a robot understands the emotional state of a user. Not only human beings but also some of animals express their own emotions through voices, face expressions and body gesture in total. Therefore, it is one of the important subjects for the ultimate goal to human like robot to use those kinds of nonverbal communication methods in combination, although there have been many attempts to recognize those nonverbal characteristics separately.

In the present paper, we attempt to perform a new method modeling the emotional features and recognizing emotional states. We present an integration method of human speech as well as visible and thermal facial expressions, aiming total understanding of the human mental states.

The recognition by thermal images is, among others, stressed in the present study. The thermal facial expressions obtained by infrared ray are not influenced by lighting condition in a room, while the face images by visible ray change very much depending on the surrounding circumstance.

2. Emotional Feature Extraction

For recognizing emotional states in both voices and facial expressions, we need to extract emotional feature parameters from them. We first analyze voices which contain emotional information including four kinds of feature parameters. As well as emotional feature extraction from voice, we also extract useful feature parameters from facial expressions of both visible and thermal images.

2.1. Emotional Feature Extraction from Voice

The prosody[1,2] is known as an indicator of the acoustic characteristics of vocal emotions. In our experiments, we used four kinds of prosodic parameters, which consist of fundamental pitch signal (F0), energy, and each derivative elements. The pitch signals in the voiced regions were smoothed by a spline interpolation. In order to consider the effect of a speaking rate, furthermore, we also used a discrete duration information when training Hidden Markov Models (HMM).

We analyze the feature parameters from the speech waveform shown in Figure 1, considering only the voiced regions as data points. All speech samples were labeled at the syllable level (/Ta/ and /Ro/) by manual segmentation in order to train each HMM using separated feature parameters. Taro is one of the most popular male name in Japan like John in English, which does not have any specific emotional meaning in itself. Figure 2 and 3 shows the pitch and energy signals extracted from each emotional speech of /taro/ spoken by a female announcer, respectively. In the figures, we can see that the feature signal of an emotion, anger, is the highest among other signals in each graph.

![Figure 1. Speech waveform labeled by two parts /ta/ and /ro/.](image-url)
2.2.  Emotional Feature Extraction from Visible and Thermal Image of Face

Many studies have been performed to tackle the issues of understanding mental states of human through facial expressions, using ordinary visible camera. However, those trials still seem to be tough jobs since there is a only slight difference among various facial expressions in terms of characteristic features for the gray level distribution of input image using visible ray (VR). Thus, we have attempted to apply thermal distribution images to facial expression recognition[3,4] using an infrared ray (IR). Figure 4 illustrates the examples of female face images by VR and IR. VR image has the shortcoming that the accuracy of facial expression recognition is greatly influenced by a lighting condition including variation of shadow, reflection, and darkness. However, it is perfectly overcome by exploiting IR.

When a face image is given into recognizer, it is necessary for better recognition performance to decide when to extract face images in accordance with voiced parts and where to take IR or VR images in the face part of the subject. We take two timing positions to extract face images, shown in Figure 1 as dotted lines where the first one and the second and ones are the maximum voice parts of /ta/ and /to/, respectively. Figure 5 and 6 show an extraction of face-parts areas which consist of three areas in the VR image and six areas in the IR image, respectively.

In the next step, we generate differential images between the averaged neutral face image and the test face image in the extracted face-parts areas to perform a discrete cosine
transformation (DCT). Figure 7 illustrates the procedure of extracting characteristic features from VR and IR images.

### 3. Recognition of emotion

In case of processing the emotional voice, the speech is sampled in the experimental conditions illustrated in Table 1 for pre-processing of emotional voice recognition, from which four dimensional emotional features are extracted.

We then train the discrete duration continuous Hidden Makov Models (DDCHMM) by using these features with three states, using label information, and run recognition tests.

**Sampling rate** | 16Khz , 16 Bit
---|---
**Pre-emphasis** | 0.97
**Window** | 16 msec. Hamming window
**Frame period** | 5 ms
**Feature parameters** | pitch signal (F0), energy, delta pitch, delta energy, discrete duration information

| Table 1. Analysis of speech signal |

In case of processing the facial images, on the other hand, 78 and 57 bits of feature parameters are used as input data for neural network (NN) with three layers for IR and VR facial expressions, respectively. The unit number of hidden layer is decided experimentally for improving the recognition accuracy for facial expression. The unit number of output layer is the number of facial expressions which should be recognized.

Figure 7 shows the overall procedure of recognizing emotional states using integration method of three different recognition results. The integration of three separate recognition accuracies is performed by the following equation.

\[
S_j = \sum_{i=1}^{3} x_{i,j}
\]

where \( x_{i,j} \) is an output value (1 or 0) for an emotional state \( i \) using a method \( j \). Therefore, recognition results are chosen when the \( S_j \) is maximum.

### 4. Experimental Results

We first examine a human performance on three different types of questionnaire. Next, we examine how effective our integration of three kinds of recognition methods is, when we compare the simulation performance results with human performance ones.

#### 4.1. Database

The samples consisted of semantically neutral utterance, Japanese name ‘Taro’, spoken and acted by an announcer. We capture voices and images simulating five emotional states such as neutral, angry, happiness, sadness, and surprise. We have simultaneously recorded voices and image sequences containing emotional states. We assembled 20 samples per each emotional expression as training data and 10 as test data.

#### 4.2. Questionnaire Results

The emotional states of speech, image samples, and both of them are estimated subjectively by total 21 students who consist of 14 male and 7 female students. Table 2 shows the three kinds of human performance results we obtained. As shown in this table, the average recognition rates for five emotional states are 84.0%, 82.4%, and 92.5%, when using emotional voices, facial expressions, and both of them, respectively. From these human performances, we could see that our data were included relatively correct emotional states and that the questionnaire result integrating both emotional voices and images gave better performance than those separately obtained from voices or images.

#### 4.3. Simulation Results

We next performed recognition of mental states over the same test data used in the questionnaire test, by integrating voices and facial expressions. Table 3 shows the recognition accuracies for each emotional state in case of emotional voices, VR and IR facial expressions, and total recognition accuracies using the integration method of the three different recognition results, respectively.
Table 2. Human performance on each emotional state

As shown in the table, the average recognition rates for five emotional states are 60%, 56%, and 48%, when using emotional voices, VR and IR facial expressions, respectively. In both cases of VR and IR facial expressions, the failure of recognition of emotion was mainly due to the difficulty to extract face-parts correctly because the subject changed her face-orientation. Overall results are shown in table 3(d) and the total recognition rates amount to 85% among five emotions (except for no answers).

Table 3. Recognition accuracies for each emotional state

This paper has described the new integration approach of recognizing human emotional states contained in voices and facial expressions. The emotional parameters were trained and recognized by HMM and NN for voices and images, respectively. The recognition results show that the integration method of recognizing emotional states in voices and images gives better performance than each isolated method in spite of the lower recognition rates compared to questionnaire counterpart.

References