Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face

Yasunari Yoshitomi[1], Sung-Ill Kim[2], Takako Kawano[3] and Tetsuro Kitazoe[1]

1:Department of Computer Science and Systems Engineering, Miyazaki University
1-1, Gakuen Kibanadai Nishi , Miyazaki , 889-2192 Japan
yoshi@cs.miyazaki-u.ac.jp, kitazoe@eagle.cs.miyazaki-u.ac.jp
2:Department of Gerontechnology,National Institute for Longevity Sciences
36-3 Gengo Morioka Obu Aichi 474-8511 Japan, kim@nils.go.jp
3:Department of Information System,Technosystem Kyushu
26-30 Morishitacho Yahatanishi-ku Kitakyushu 806-0046 Japan, t_kawano@tsk.yokogawa.co.jp

Abstract

A new integration method is presented to recognize the emotional expressions of human. We attempt 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 that obtained from each single experiment. The results are compared with the recognition by human questionnaire.

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 human. The goal of our research is to develop a robot which can perceive human feelings or mental states. The robot should be able to interact in a friendly manner with a human. For example, it could perhaps encourage a human who seems sad. Moreover, it could advise a person to stop working and rest for a while when the individual seems tired. We have divided a way to achieve this goal into three stages.

The first stage is to develop a method for the integration of information regarding the emotional expression of human. The basic elements for integration are visible-ray image, thermal image, voice, and sound. The present state of development of such a system is at the first stage. We can give a robot a type of information which can not be processed by the human brain. Thermal imaging is a good example because it is impossible for a human to perceive heat via the naked eye. The second stage is to develop an automatic, real-time, interactive system that has information integration as a processing characteristic. The third stage is to develop a robot which has a function developed from a synthesis of the first and second stages and can be used in our daily lives.

The presented investigation concerns the first stage of development wherein a robot acquires the ability to detect human feeling or inner mental states. Although the mechanism for recognizing facial expressions as one of the main, visible expressions of feeling has been received considerable attention in the course of computer vision research, its present stage still falls far short of human capability, especially from the viewpoint of robustness under widely ranging lighting conditions and in regard to the capability of a machine to understand emotional states. One of the reasons for the former is that nuances of shade, reflection, and local darkness influence the accuracy of facial expression recognition through the inevitable change of gray levels. The reason for the latter is that an input-image from an ordinary visible rays (VR) camera shows a very small difference between unconsciously and artificially smiling faces, for example, from the viewpoint of gray level distribution. In order to avoid these problems, an image registered by infrared rays (IR) which describes the thermal distribution of the face instead of ordinary visible light has been used for developing a robust method for recognizing facial expressions and emotional states which is applicable under widely varied lighting condition[1, 2, 3].

Recently, many researches have investigated some methods on recognizing the individual nonverbal characteristics such as emotional factors contained in the speech, facial expressions, and body gestures etc.. Therefore, it is most important to study the total aspects of the human emotional expressions.

In this paper, 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 these three elements, stressed in the present study.

2 Characteristics of IR Image for Human Skin

The principle of thermal image generation comes from well-known law by Stefan and Boltzmann, which is expressed as $W = \varepsilon\sigma T^4$, where W is radiant emittance (W/cm^2), ε is emissivity, σ is Stefan-Boltzmann constant (5.6705×10^{-12} W/cm^2K^4), T is Temperature (K).

For human skin, ε is estimated at 0.98 to 0.99[4]. In this study, however, the approximate value of 1 is used as ε for human skin. The values of ε for almost all substances except human skin are lower than that for human skin[4]. Accordingly, human face is easily extracted in the scene having its circumstances whose temperature are lower than that of human face, when the range of skin temperature is selected for producing the thermal image, using the value of 1 for ε [1]. Figure 1 shows the examples of male face images by VR and IR. In principle, the temperature measurement by IR dose not depend on skin color[4], darkness and lighting condition, resulting in that face and its characteristics are easily extracted in the input image containing face and its surrounding.

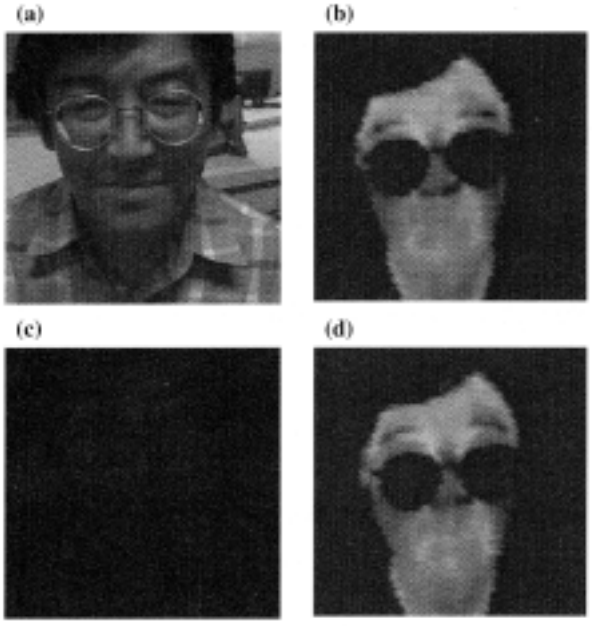

Figure 1: Examples of face-image at night;(a)visible ray with lighting, (b)IR with lighting, (c)visible ray without lighting, (d)IR without lighting[2].

3 Emotional Feature Extraction and Recognition of Emotion

For recognizing emotional information 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. Figure 2 illustrates the procedure for recognizing the emotional information contained in voices, VR and IR images. We explain the procedure in the following.

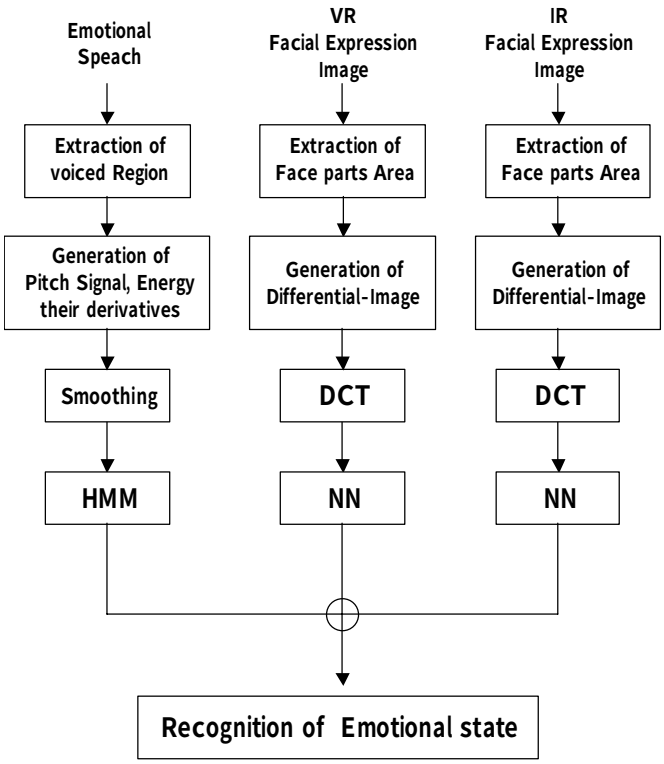

Figure 2: Procedure for recognizing the emotional information contained in voices, VR and IR images.

3.1 Emotional Feature Extraction from Voice

The prosody [5, 6] is known as an indicator of the acoustic characteristics of vocal emotions. In our experiments, we use four kinds of prosodic parameters, which consist of fundamental pitch signal (F0), energy, and their derivative elements. The pitch signals in the voiced regions are smoothed by a spline interpolation. In order to consider the effect of a speaking rate, furthermore, we use a discrete duration information when training Hidden Markov Models (HMM). We analyze the feature parameters from the speech waveform shown in Figure 3, considering only the voiced regions as data points. All speech samples are 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 4 and 5 show the pitch and energy signals extracted from each emotional speech of /taro/ spoken by a female subject who was a professional announcer, respectively. In the figures, for example, we can see that the feature signal of an emotion, anger, is the highest among all signals in each graph.

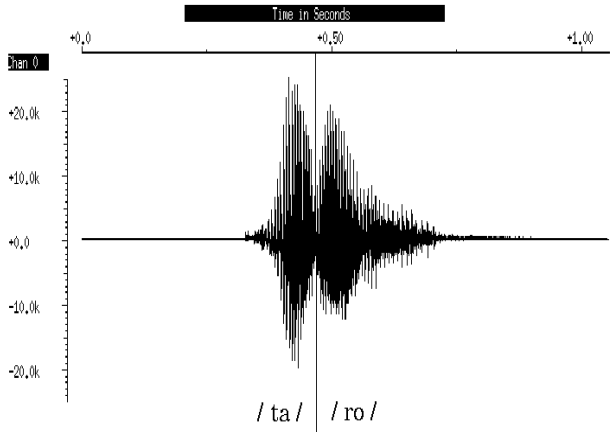

Figure 3: Speech waveform labeled by two parts /ta/ and /ro/.

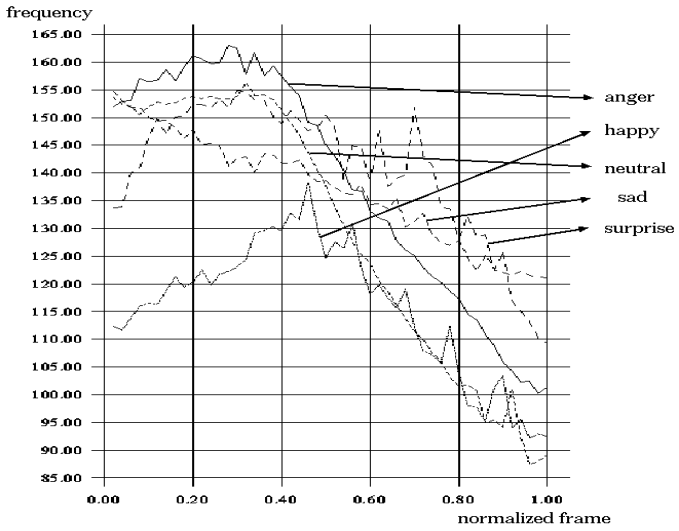

Figure 4: Pitch signals in each emotional state

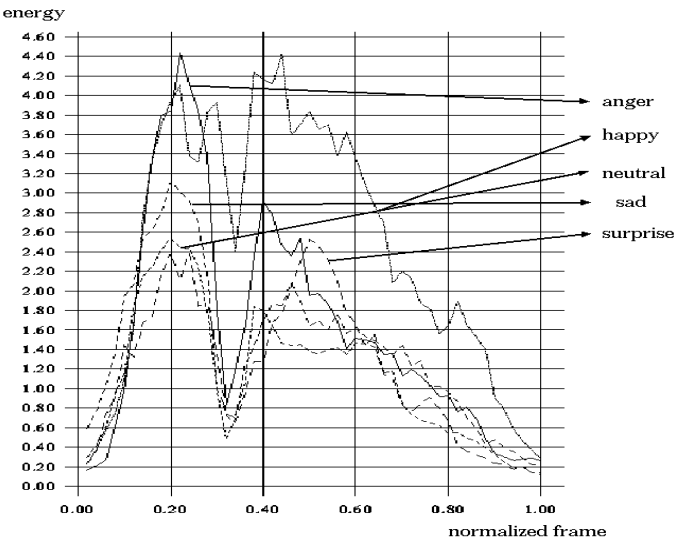

Figure 5: Energy signals in each emotional state.

3.2 Emotional Feature Extraction from Visible and Thermal Images of Face

Many studies have been performed to tackle the issues of understanding mental states of human through facial expressions, using ordinary VR camera. However, those trials still seem to be tough jobs since there is a slight difference among various facial expressions in terms of characteristic features for the gray level distribution of input image using VR. Thus, we have attempted to apply thermal distribution images to facial expression recognition using IR.

We take two timing positions to capture face images, shown in Figure 3 as dotted lines where the first and the second ones are the maximum voice parts of /ta/ and /ro/, respectively. When a face image is given into computer, it is necessary to extract face-parts areas correctly, which will be important for better recognition performance. Some face-part areas in VR image are extracted by exploiting both the information on corresponding thermal image and the simulated annealing method for template matching of the block-region including eyes. Figure 6 and 7 show blocks for extracting 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). For IR image, some face-parts areas are also extracted through segmentation of the image, then DCT is performed.

3.3 Recognition of Emotion

In case of processing the emotional voice, the speech is sampled in the experimental conditions illustrated in Table 1

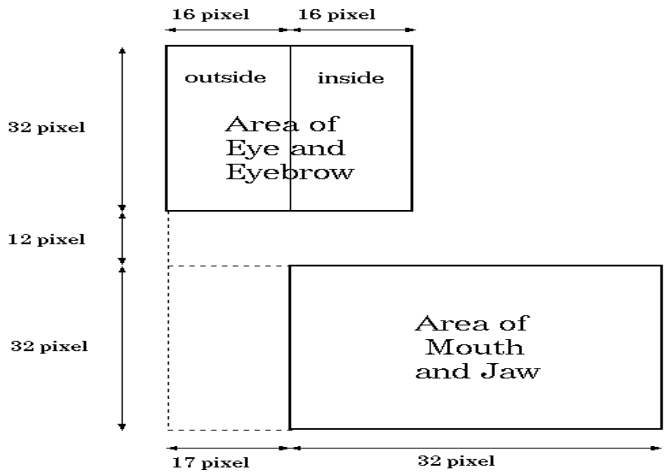

Figure 6: Blocks for extracting face-parts areas in the VR image

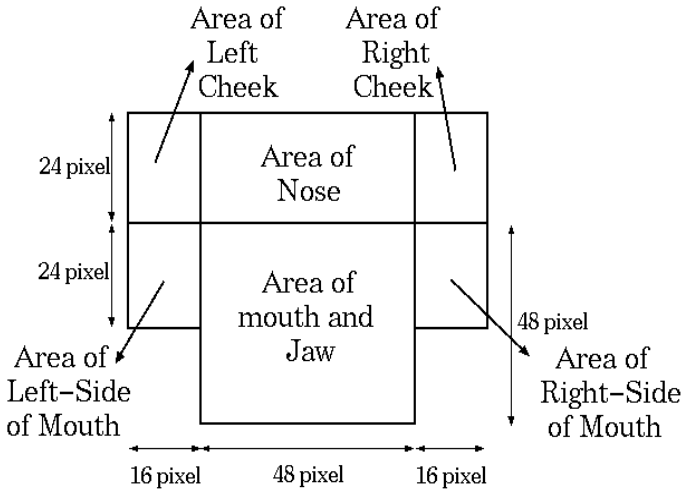

Figure 7: Blocks for extracting face-parts areas in the IR image.

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.

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.

The feature vector automatically made from the DCT coefficients for VR image is used as input to a neural network (NN) with Back Propagation method. The other feature vector automatically produced from thermal image with the way similar to the above method is used as input to another NN. In case of processing the facial images, 48 to 78 and 22 to 57 bits of feature parameters are used as input data for NN with three layers for IR and VR facial expressions, respectively.

For integration of information from VR and IR images, the weighted summation $S2_i$ for the emotional state i is calculated as $S2_i = \sum_{j=1}^{2} R_j x_{i,j}$, where R_j is a reliability of method j , and $x_{i,j}$ is an output intensity of i unit, corresponding to an emotional state i , in output layer in NN, using a method j . The identification accuracy in case of applying each method to learned data themselves is used as the value of reliability for each method with NN. The value of reliability is used for the method as the weight in the above weighted summation. Therefore, recognition result i is chosen when the $S2_i$ is maximum.

For integration of information from voices, VR and IR images, the weighted summation $S3_i$ for the emotional state i is calculated as $S3_i = \sum_{j=1}^{3} y_{i,j}$, where $y_{i,j}$ is an output value (1 or 0) for an emotional state i using a method j . Therefore, recognition result i is chosen when the $S3_i$ is maximum.

4 Experiment and Discussion

First, we capture IR and VR images simulating four emotional states such as neutral, happiness, sadness, and surprise, acted by two male subjects. We assemble 10 samples per each emotional expression per a subject as training data and 5 as test data.

The average recognition rates for four emotional states are 85%and 75%, when using VR and IR facial expressions, respectively. The average recognition rate obtained by information integration of VR and IR facial expressions is 95%, which is better than that from each single information.

Second, the samples consisted of semantically neutral utterance, Japanese name 'Taro', spoken and acted by one female subject who was a professional announcer. We capture voices and images simulating five emotional states such as neutral, angry, happiness, sadness, and surprise. We simultaneously record voices and image sequences containing emotional information. We assemble 20 samples per each emotional expression as training data and 10 as test data.

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**86.2**	7.7			5.7
	Hap.	2.8	**75.7**	1.2	0.5	4.3
	Neu.		1.9	**97.1**		2.9
	Sad.	1.4	2.9	1.7	**99.5**	25.7
	Sur.	9.6	11.8			**61.4**

(a) Human performance on emotional voices

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**56.7**	4.0	2.7	2.7	5.3
	Hap.	3.3	**90.0**		3.3	2.7
	Neu.	15.3	2.0	**94.0**	2.0	11.3
	Sad.	21.3	2.0	2.7	**92.0**	1.3
	Sur.	3.3	2.0	0.7		**79.3**

(b) Human performance on facial expressions

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**95.3**	1.3			3.3
	Hap.		**92.0**		0.7	3.3
	Neu.			**100.0**	0.7	4.0
	Sad.		0.7		**98.7**	12.7
	Sur.	4.7	6.0			**76.7**

(c) Human performance on both emotional voices and images

Table 2: Human performance on each emotional state.

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**80**	20		20	10
	Hap.		**10**			
	Neu.		20	**100**	10	20
	Sad.				**50**	10
	Sur.	20	50		20	**60**

(a) Recognition accuracy for emotional voices

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**40**	40		20	30
	Hap.	50	**60**		10	
	Neu.			**70**		
	Sad.	10			**40**	
	Sur.			30	30	**70**

(b) Recognition accuracy for VR facial expressions

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**40**				
	Hap.		**0**		20	
	Neu.			**70**		
	Sad.		100	30	**80**	50
	Sur.	60				**50**

(c) Recognition accuracy for IR facial expressions

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**60**			10	
	Hap.		**10**			
	Neu.			**90**		
	Sad.				**50**	10
	Sur.	10			20	**70**
	No Ans.	30	90	10	20	20

(d) Total recognition accuracies using integration method

Table 3: Recognition accuracy for each emotional state.

The emotional information in speech, image samples, and both of them is subjectively recognized by 14 male and 7 female students. Table 2 shows the three kinds of human performance results we obtain. 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, VR facial expressions, and both of them, respectively. From the table, we can see that our data include relatively correct emotional information and that the questionnaire result integrating both emotional voices and images gives better performance than that separately obtained from voices or images.

We next perform recognition of mental states over the same experiment data used in the questionnaire test, by integrating voices, VR and IR facial expressions. Table 3 shows the recognition accuracy for each emotional state. The average recognition rates for five emotional states are 60%, 56%, and 48%, when using voices, VR and IR facial expressions, respectively. In both cases of VR and IR facial expressions, the failure of recognition of emotion is mainly due to the difficulty to extract face-parts correctly because the subject changes her face-orientation to express her emotion. Overall results are shown in Table 3(d) and the total recognition rates amount to 85% among five emotions (except for no answers).

The present result was for each person. The individuality of feeling or states of mind is an important issue which we approach from two directions. One is to produce a data base made from averaged feature-vectors acquired from many subjects. The other is to produce a data base made from feature vectors acquired from individual subjects. In the second approach, face recognition is indispensable before understanding feeling or states of mind. Since the face identification method with IR image analysis has been developed[7, 8, 9], the data base of characteristic parameter for facial expression which can be made for each person will be available. Namely, after identification of the person, the facial expression can be recognized with the data base.

5 Conclusions

This paper has described the new integration approach to recognizing the emotional information of human contained in voices, VR and IR facial expressions. The emotional parameters are trained and recognized by HMM and NN for voices and images, respectively. The recognition results show that the integration method for recognizing emotional states gives better performance than any of isolated methods.

Acknowledgments

The authors wish to thank Dr.E.Hira of Mechanics Department, Miyazaki Prefectural Industrial Research Institute for his valuable suggestions for the experiment using IR apparatus.

References

[1] Y. Yoshitomi, S. Kimura, E. Hira and S. Tomita, "Facial Expression Recognition Using Infrared Rays Image Processing", *Proc. of the Annual Convention IPS Japan*, No.2, pp. 339-340, 1996.

[2] Y. Yoshitomi, N. Miyawaki, S. Tomita and S. Kimura, "Facial Expression Recognition Using Thermal Image Processing and Neural Network", *Proc. of 6th IEEE International Workshop on Robot and Human Communication*, pp. 380-385, 1997.

[3] Y.Sugimoto, Y.Yoshitomi and S.Tomita, "A Method for Detecting Transitions of Emotional States Using a Thermal Facial Image Based on a Synthesis of Facial Expressions",J. of Robotics and Autonomous Systems, Vol.31, pp.147-160, 2000.

[4] H. Kuno, Sekigaisen Kougaku, IEICE, Tokyo, 1994. (in Japanese).

[5] A.Waibel,"Prosody and Speech Recognition", *Doctoral Thesis*, Carnegie Mellon Univ, 1986.

[6] C Tuerk, "A Text-to-Speech System based on NETtalk", *Master's Thesis*, Cambridge University Engineering Dept, 1990.

[7] Y. Yoshitomi, T. Miyaura, S. Tomita and S. Kimura, "Face Identification Using Thermal Image Processing", *Proc. of 6th IEEE International Workshop on Robot and Human Communication*, pp. 374-379, 1997.

[8] Y. Yoshitomi, A. Tsuchiya and S. Tomita,"Face Recognition Using Dynamic Thermal Image Processing ", *Proc. of 7th IEEE International Workshop on Robot and Human Communication*, pp. 443-448, 1998.

[9] Y. Yoshitomi, M. Murakawa and S. Tomita, "Face Identification Using Sensor Fusion of Thermal Image and Visible Ray Image", *Proc. of 7th IEEE International Workshop on Robot and Human Communication*, pp. 449-455, 1998.