International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 1, March 2005 pp. 1-6

A Multimodal Emotion Recognition Using the Facial Image and Speech Signal

Hyoun-Joo Go*, Yong-Tae Kim, Myung-Geun Chun***

*** Chungbuk National University, School of Electrical and Computer Engineering**
Research Institute for computer and Information Communication
**** Hankyong National University Department of Information and Control Engineering**

Abstract

In this paper, we propose an emotion recognition method using the facial images and speech signals. Six basic emotions including happiness, sadness, anger, surprise, fear and dislike are investigated. Facial expression recognition is performed by using the multi-resolution analysis based on the discrete wavelet. Here, we obtain the feature vectors through the ICA(Independent Component Analysis). On the other hand, the emotion recognition from the speech signal method has a structure of performing the recognition algorithm independently for each wavelet subband and the final recognition is obtained from the multi-decision making scheme. After merging the facial and speech emotion recognition results, we obtained better performance than previous ones.

Key words : Emotion Recognition, Face Recognition, Speech Recognition, Human Interface, Independent Component Analysis, Fuzzy Membership Function.

1. Introduction

The popularity of computers has rapidly increased due to the progress of information technologies. Accordingly, the researches on human and computer interface are gaining more interest. Human being usually recognizes others emotional state based on the language, voice, gesture, sight, etc.

Related to this, various researches on facial emotion recognition have been performed. These studied are usually based on the optical flow analysis[1], PCA (principal component analysis)[2], LFA(local feature analysis)[3], LDA(linear discriminant analysis)[4], and ICA(independent component analysis)[5][6] method.

On the other hand, Fukuda[7] attempted the classification of 6 basic emotions using tempo and energy of speech signal. Moriyama[8] studied a method of recognizing and synthesizing emotional content in speech. Also, Silva[9] performed the emotion recognition using the pitch and HMM (Hidden Markov Model) of the English and Spanish speech signal.

Fig. 1. Six basic emotions (happiness, sadness, anger, surprise, fear, dislike)

Manuscript received Feb. 22, 2005; revised Mar. 9, 2005.
This work was supported by grant No.(R01-2002-000-00315-0) from the Basic Research Program of Korea Science & Engineering Foundation.

Figure 1 shows the six basic emotions studied by Ekman and Friesen[10]. It is noted that the six emotions are all common even for different cultures. Hereafter, we deal with an emotion recognition for the six emotions using facial images and speech signals.

Facial expression recognition is performed by using the multi-resolution analysis based on the discrete wavelet. And then, the feature vectors are extracted by using the ICA(Independent Component Analysis) method.

On the other hand, the emotion recognition from speech signal method has a structure performing of the recognition algorithm independently for each wavelet subband and then the final recognition is obtained from the multiple decision-making scheme. Here, the matching values of each emotion state in speech signals and in facial images are merged using a fuzzy membership function, so called ZMF (Z-Membership function).

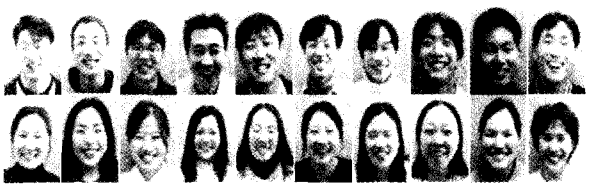


Fig. 2. Traing images for twenty persons (Happiness)

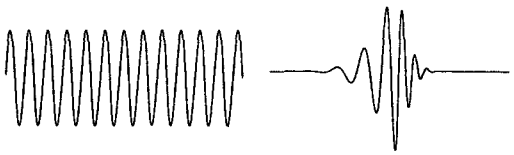
Figure 2 shows 640×480 facial images containing 6 basic emotions for 20 people (10 males, 10 females). We use 180 female images among 720 facial images(6 images per emotion for 20 people) to construct a code book. And then we used the remaining images as checking set.

The speech signal is acquired from an emotional sentence. This speech signal contains six basic emotions for 20 people (10 males, 10 females). We use female 120 speech signals among 360 speech signals (3 speech signals per emotion for 20 people) to construct a code book. And then we used the remaining speech signals as checking set.

This paper is organized as follows. Section 2 and Section 3 explain the emotion recognition method from facial image and speech signal, respectively. Section 4 describes our proposed fusion scheme and emotion recognition experiments. Finally we make some concluding remarks in Section 6.

2. Emotion Recognition From Facial Expression

Fourier analysis expands signals or functions in terms of sinusoids (or, equivalently, complex exponentials) which has proven to be an extremely valuable in mathematics, science, and engineering, especially for periodic, time-invariant, or stationary phenomena. On the other hand, wavelet is a "small wave", which has its energy concentrated in time to give a tool for the analysis of transient, nonstationary, or time-varying phenomena. It still has the oscillation wavelike characteristic but also has the ability to allow simultaneous time and frequency analysis. This is illustrated in Figure 3 with sinusoid over $-\infty \le t \le \infty$ and, therefore, having infinite energy and with the wavelet having its finite energy concentrated around a point[11].



(a) A sine wave (b) Daubechies' wavelet

Fig. 3. sine wave and wavelet

Using the wavelet transform, facial image can be decomposed into several subband frequency images. Figure 4. shows an example of a decomposed image into third level. Sub-images LL3, HL3, LH3, and HH3 are third level wavelet transformed results and correspond to LL, HL, LH, and HH frequency bands, respectively.

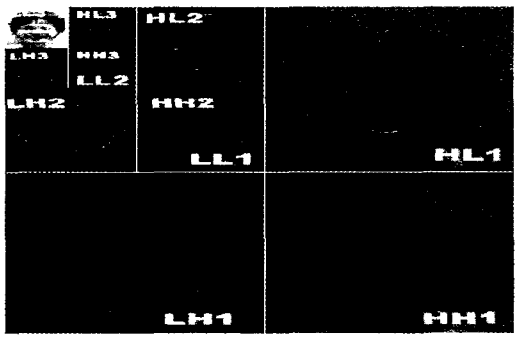


Fig. 4. A separated image by four band applying wavelet transform

After decomposing the facial images using the wavelet transform, we first perform a dimensionality reduction by applying PCA. We then search for the most discriminant projection along eigenvectors by successively selecting the independent components by using the ICA algorithm.

ICA is originally proposed to solve the blind source separation problem of recovering independent source signals after they are linearly mixed by an unknown matrix **A**. Nothing is known about the sources or the mixing process except that there are N different recorded mixtures. The task is to recover a version, Y, of the original sources, **S**, identical save for scaling and permutation, by finding a square matrix, W, specifying spatial filters that linearly invert the mixing process, i.e. $Y=W\mathbf{X}$ which ensures stability of the learning rule. Note that although a non-linear function is used in determining W, once the algorithm converges and W is found, the decomposition is a linear transformation, $Y=W\mathbf{X}$. Figure 5 shows the Block diagram for ICA algorithm[12].

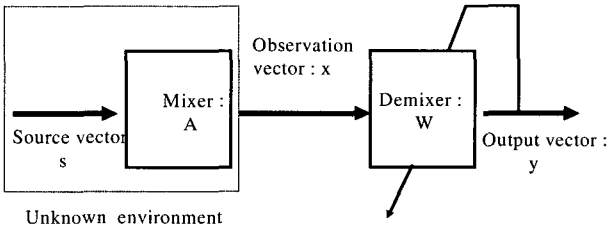


Fig. 5. Block diagram for ICA algorithm

Figure 6 shows the flowchart of independent component analysis after principal component analysis and Figure 7 shows the linear combination of feature vectors and there icafaces

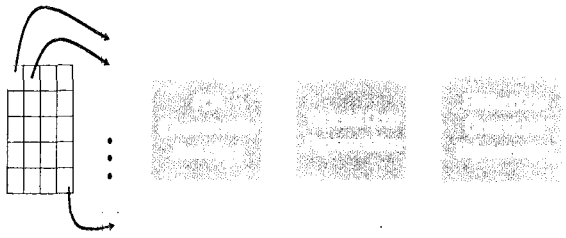


Fig. 6 Flowchart of independent component analysis

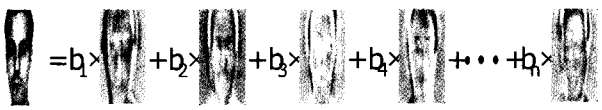


Fig. 7 Linear combination of feature vector and icafaces

In this paper, we applied the discrete wavelet transform three times for the training images of 640×480 and acquired the decomposed images of 40×30 from third band. These

images use the information of LL band. Thereafter, we apply the PCA method for the LL4. Here, we use 140 eigenfaces for male and 120 eigenfaces for female. After applying the PCA, we use ICA method to get the icafaces. Here, we use 20 icafaces shown in Figure 8.

Fig. 8. icafaces obtained by ICA

We apply the ICA for the image data set. Here we also get the best recognition rate at LL3 band which are 94.44% and 96.11% for male and female, respectively.

3. Emotion Recognition from Speech Signal

We adopt the multiresolution analysis of wavelet packet for a speech signal. From this, we obtain a discrete approximation and a detail signal at a particular resolution. The detailed signal corresponds to the coefficients generated from high-pass filter convolution of h(n) and the approximation signal correspond to the coefficients generated from high-pass filter convolution of g(n). The resulting convolution coefficients are subsampled or decimated, the output signals subband-4 and subband-3 are decreased by a factor of four as compared with original signal.

As you can see in Figure 9, output signals are generated through the structure of wavelet packets. For example, output signals subband-2 and subband-1 imply high-frequency signal and low-frequency each in lower frequency band.

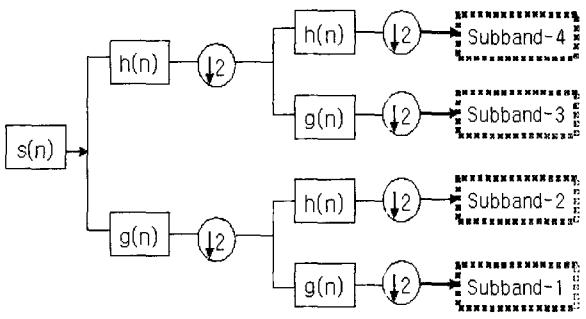


Fig 9. Structure of wavelet packet

The speech signal is divided equally into frequency bands through the wavelet packet. The input signal is analyzed by adopting Daubechies basis function of a mother wavelet. The detection of speech utterances in the presence of background noise is the first step of a speech processing. The efficiency of endpoint detection affects the performance of the entire recognition system. For the detection of the endpoints, several algorithms have been used. Among these algorithms, we use both short-time energy and zero crossing rate.

After detecting the speech segment, we devide the whole frequency band into several subbands. And then speech features from each subband are extracted. These speech features may be MFCC, LPCC or other features[8]. We also use the 13 dimentional MFCC of each subband as feature vector.

First, codebook is constructed by the K-means algorithm. Here, we make two codebooks for male and female respectively to improve the recognition rate. The size of codebook is very important factor. Increasing of codebook size improves recognition rate but it makes worse influence on the recognition time or required memory. Thus, we take the size of codebook as 384×13 per a subband.

The overall recognition process is shown in Figure 10. For an input speech signal, the start and end points are detected and then we apply the wavelet transform for the detected segments to extract the feature vectors. The obtained feature vectors are compared with the codebooks to compute the matching scores. Using these matching scores, we apply the multiple band decision-making scheme shown in Figure 11 to get a total score. Here, we compute the matching scores for each codebook at each subband and then the matching scores are summed at each emotioin, which will be the final matching score for the emotion. The final decision is made by taking the maximum score.

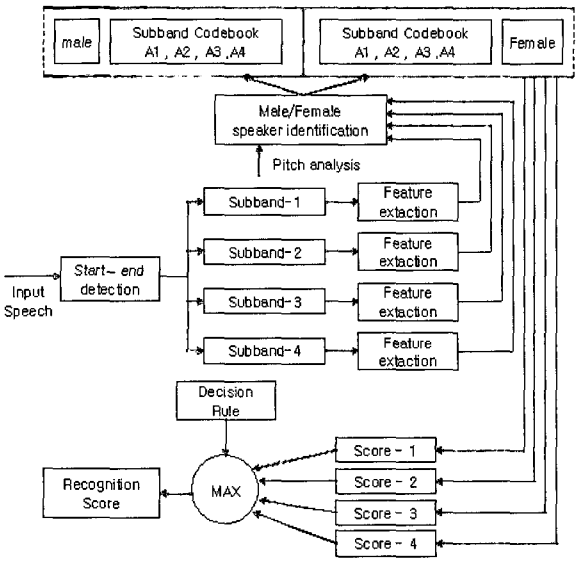


Fig 10. Emotion recognition system using the wavelet filter banks

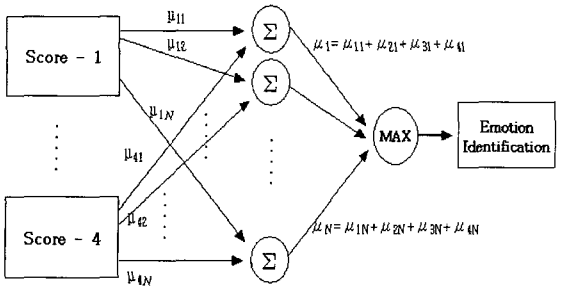


Fig 11. Multiple band decision-making method

We carry out various experiments to test the performance of the methods. For establishing the database, we obtain six emotional speech signals by pronouncing in Korean "뭐라구 (So what?)". Here, we take the sampling frequency as 11.025kHz and use the Hamming window whose width is 10 ms with 5 ms overlapping. After dividing the original signal by using the Hamming Window, we obtain 13 dimensional Mel-Cepstrum calculation per each frame. In this manner, we obtain a codebook whose size is 64 for each emotion.

We summarize the results of emotion recognition in Table 1. As you can see, A1 band shows the best recognition results for male and female. It is noticeable that the characteristic of recognition rates for each band is different with sexuality. This may reflect the fact that the frequency characteristics of male and female voice is different. In case 1, we used three subband A1,A2, and A3 for the multiple band decision making scheme shown in Figure 11. On the other hand, we used all subband for the multiple band decision making scheme in case 2. We can find that the best recognition rate of 93.3% is obtained in case 1 for male and also the best recognition rate of 93.3% is obtained in case 2 for female. So, we will reflect this results the further processing of the fusion scheme.

Table 1. Emotion recognition rate for each subband

Subject	Band				case 1	case 2
	A1	A2	A3	A4		
male	89%	78%	71%	57%	93.3%	85%
female	85%	72%	81%	68%	86%	93.3%

4. Emotion Recognition Using the Facial Image and Speech Signal

Various emotion recognition studies have been performed by using facial expression or speech signal independently. Recently, for obtaining high accuracy, it is interested in a fusion model which consisted of two or more features such as face, speech, etc. We also propose an emotion recognition method using facial expressions and speech signal in this paper. As mentioned before, the facial expression recognition is performed by using the multi-resolution analysis based on the discrete wavelet. Here, the feature vectors are extracted by using ICA method. On the other hand, the emotion recognition from speech signal method has a structure of performing the recognition algorithm independently for each wavelet subband and then the final recognition is obtained from the multi-decision making scheme.

Let us consider the emotion recognition scheme based on the facial images and speech signals. We make the codebooks such as describing in Section 2 and Section 3. Figure 12 shows the structure of codebooks.

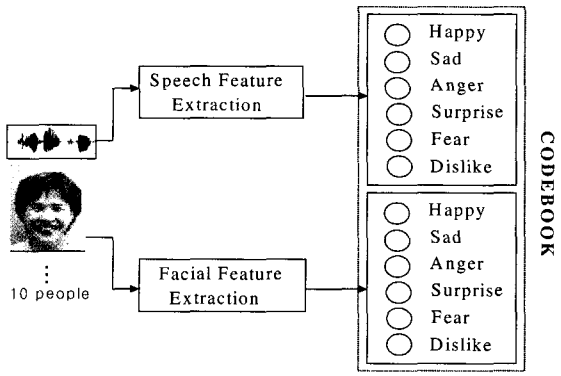


Fig 12. Process of making codebook for ten speakers

First, we classify the subject into male or female by the pitch of a speech signal. Then, the membership degree for each emotion is calculated by comparing the features in the input speech and reference features in codebook as described in Section 3. In this manner, the membership degree for face images is also calculated. Membership degree for each emotion can be expressed between number of 0 and 1 using ZMF(Z-Membership Function) where critical value are determined as 0.2 and 0.8 through various experiments[13].

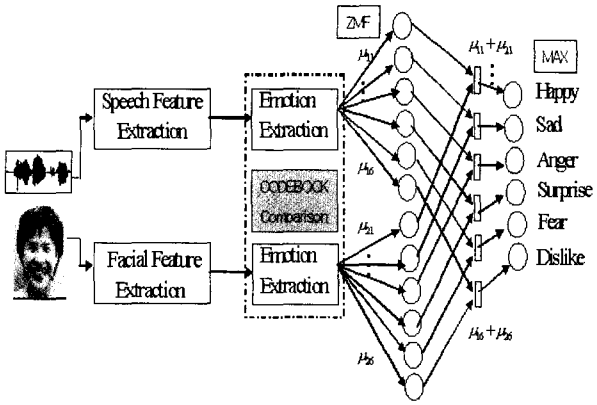


Fig. 13. Multimodal recognition method

Final emotional state can be decided by selecting the maximum membership value or adding two membership values. We take the method of selecting the maximum value. The final emotion recognition results are shown in Table 2 and Table 3 for male and female, respectively. The final recognition rates

are 97.7%(176/180) for male and 98.3%(177/180) for female, which show better results than previous ones. We find that the emotion of surprise makes most errors in male and the emotion of anger makes most errors in female. These show that the way of emotional expressions are different between male and female and also female usually express their emotion more explicitly in our case.

Table 2. Final recognition rate for male

	happiness	sadness	anger	surprise	fear	dislike
happiness	30	0	0	0	0	0
sadness	0	30	0	0	0	0
anger	0	1	29	0	0	0
surprise	0	0	1	29	0	0
fear	0	0	0	2	28	0
dislike	0	0	0	0	0	30

Table 3. Final recognition rate for female

	happiness	sadness	anger	surprise	fear	dislike
happiness	28	0	1	0	0	1
sadness	0	30	0	0	0	0
anger	0	0	30	0	0	0
surprise	0	0	0	29	0	1
fear	0	0	0	0	30	0
dislike	0	0	0	0	0	30

5. Conclusions

In this paper, we deal with a multimodal emotion recognition method using facial expressions and speech signal. Six basic human emotions including happiness, sadness, anger, surprise, fear and dislike are investigated. Facial expression recognition is performed by using the multi-resolution analysis based on the discrete wavelet. Then, the feature vectors are extracted by using ICA. On the other hand, the emotion recognition from speech signal method has a structure of performing the recognition algorithm independently for each wavelet subband and then the final recognition are obtained from the multi-decision making scheme. Finally we obtained better performance results than previous ones after merging the facial and speech emotion recognition methods by using the proposed scheme.

References

[1] J. Lien, T. Kanade, C. Li, "Detection, tracking, and classification of action units in facial expression", Journal of Robotics and Autonomous Systems, Vol. 31, No. 3, pp. 131-146, 2000.

[2] M. Turk, A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.

[3] P. Penev, J. Atick, "Local feature analysis: a general statistical theory for object representation", Network : Computation in Neural Systems, Vol. 7, pp. 477-500, 1996.

[4] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, 1997.

[5] Brown, G. and Satoshi, Y. and Luebben, H. and Sejnowski, T.J., Separation of optically recorded action potential trains in tritonia by ICA, In: Proc. 5th Annual Joint Symposium on Neural Computation, 1998.

[6] Marian Stewart Bartlett, Javier R. Movellan, "Face Recognition by Independent component Analysis" IEEE transactions on neural networks, Vol.

[7] V.Kostov and S.Fukuda, Emotion in User Interface, Voice Interaction System, IEEE Intl Conf. on Systems, Man, Cybernetics Representation, no. 2, pp. 798-803, 2000.

[8] T. Moriyama and S. Oazwa, Emotion Recognition and Synthisis System on Speech IEEE Intl. Conference on Multimedea Computing and Systems, pages 840-844, 1999.

[9] L.C. Silva and P.C. Ng, Bimodal Emotion Recognition, Proceeding of the 4th International Conference on Automatic Face and Gesture Recognition, pp. 332-335, 2000.

[10] P.Ekman and W.V. Friesen. Emotion in the human face System. Cambridge University Press, San Francisco, CA, second edition, 1982.

[11] Burrus, Gopinath, Guo, "Introduction to Wavelets and Wavelet Transforms A Primer" Prentice-Hall International, Inc, 1998.

[12] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager "Classifying Facial Actions" IEEE transactions on pattern analysis and machine intelligence, Vol. 21, No 10, 1999.

[13] Roger Jang, Chuen-Tsai Sun, Neuro-fuzzy and Soft computing. Prentice-Hall International, 1997.

Hyoun Joo Go

received the B.S. degree in control and instrumentation engineering from Hanbat National University, Korea in 1999 and M.S. degrees in electrical & electronics engineering from Chungbuk National University, Korea, in 2002, respectively. She is currently working toward th Ph.D. degree in electrical & electronics engineering at Chungbuk National University. Her current research interests include biometrics, multimodal emotion recognition, fuzzy-model-based control, and image signal processing.

Phone : +82_43_261_2388
Fax : +82_43_268_2386
E-mail : ghjswy@hanmail.net

Yong-Tae Kim
received the B.S. degree in electronics engineering from Yonsei University, Korea in 1991 and the M.S. and the Ph.D. degrees in electrical engineering from KAIST, Korea in 1993 and 1998, respectively. He is currently an assistant professor of Department of Information & Control Engineering, Hankyong National University, Anseong, Korea. His current research interests include intelligent robot, intelligent system, intelligent control, learning control, fuzzy control.

Phone : +82-31-670-5292
Fax : +82-31-670-5299
Email : ytkim@hknu.ac.kr

Myung Geun Chun
received the B.S. degree in electronics from Pusan National University, Korea, in 1987 and M.S. and Ph.D. degrees in electrical & electronics engineering from KAIST, Korea, in 1989 and 1993, respectively. From 1993 to 1996, he worked in automation laboratory of Samsung electronic company as a senior researcher. Since 1996, he has been a Professor in the electrical and computer school of Chungbuk National University. His research interests include biometrics, emotion recognition, speech signal processing.

Phone : +82_43_261_2388
Fax : +82_43_268_2386
E-mail : mgchun@chungbuk.ac.kr