Emotion Recognition based on Multiple Modalities

Dong-Ju Kim*, Hyeon-Gu Lee**, Kwang-Seok Hong*

Abstract

Emotion recognition plays an important role in the research area of human-computer interaction, and it allows a more natural and more human-like communication between humans and computer. Most of previous work on emotion recognition focused on extracting emotions from face, speech or EEG information separately. Therefore, a novel approach is presented in this paper, including face, speech and EEG, to recognize the human emotion. The individual matching scores obtained from face, speech, and EEG are combined using a weighted-summation operation, and the fused-score is utilized to classify the human emotion. In the experiment results, the proposed approach gives an improvement of more than 18.64% when compared to the most successful unimodal approach, and also provides better performance compared to approaches integrating two modalities each other. From these results, we confirmed that the proposed approach achieved a significant performance improvement and the proposed method was very effective.

Keywords : Emotion Recognition, Face, Speech, EEG, Multiple Modalities

I. Introduction

A challenging research issue and one that has been of growing importance to those working in human-computer interaction are to endow a machine with an emotional intelligence. Such a system must be able to create an affective interaction with users: it must have the ability to perceive, interpret, express and regulate emotions [1]. In this case, recognizing user's emotional state is one of the main requirements for computers to successfully interact with humans [2].

There have been a lot of researches in emotion recognition using various modalities such as face, speech and electroencephalography (EEG). It is virtually impossible to enumerate all of them. Multimodal emotion recognition has been a popular method recently. In addition to the basic approach that balances the bimodal information in the video based emotion recognition, there has been a lot of research done in the last a few years to improve the basic multimodal technique [3].

In this paper, we compose the multimodal emotion recognition system by combining the results of face, speech and EEG recognition. Especially, we expect to performance improvement of proposed system by adding the EEG modality in previous works based on face and speech. In addition, EEG signal can be acquired more conveniently according as headband-type devices are developed recently [4][5]. The distance measurement, i.e., the raw-score of face-based emotion recognition, is calculated by two-dimensional principal component analysis (2D-PCA) and nearest neighbor classifier, and also the likelihood measurement is obtained by Gaussian mixture model (GMM) algorithm based on mel-frequency cepstral coefficient (MFCC) in speech-based emotion recognition. In EEG-based emotion recognition, the distance value is calculated by triangular filter-based cepstral coefficient (TFCC) and k-nearest neighbour classifier. Then, these raw-scores, i.e., distance or likelihood value, are converted into normalized-scores. Since the raw-scores of face, speech and EEG have different distributions, we apply the sigmoid function to normalize these raw-scores from 0 to 1. Finally, we compose the multimodal emotion recognition system by fusing these normalized-scores using a

* 성균관대학교 ** 서일대학

투고 일자 : 2011. 5. 14 수정완료일자 : 2011. 9. 26
계재확정일자 : 2011. 11. 1

*Acknowledgements

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2010-0027789), MKE, Korea under ITRC NIPA-2010-(C1090-1021- 0008) (NTIS-2010-(1415109527)), and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0021411).

weighted-summation method. The fused-score is used to classify the unknown emotion states.

II. Emotion Recognition by Facial Expressions

A. 2D-PCA

Principal component analysis (PCA) is a well-known feature extraction and data representation technique widely used in the areas of pattern recognition, computer vision and signal processing. The central underlying concept is to reduce the dimensionality of a data set whilst retaining the variations in the data set as much as possible [6]. We first needs to obtain the projection axes having the largest variance of the projected images to obtain the eigenspaces. Then, we repeat this procedure in the orthogonal space until we realize that there is no more variance to take into account. Let an image that has m rows and n columns, be a two-dimensional matrix. In PCA, the corresponding x_i image is viewed as a vector with $m \times n$ coordinates that results from a concatenation of successive rows of the image. Denote the training set of M images by $X=\{x_1, x_2, \cdots, x_M\}$. Then, the corresponding covariance matrix is obtained given by

$$C = \frac{1}{M}\sum_{i=1}^{M}(x_i-\bar{x})(x_i-\bar{x})^T, \quad (1)$$

where $\bar{x} = 1/M\sum_{i=1}^{M} x_i$. Here, if $E=\{e_1, e_2, \cdots, e_D\}$, $D \le M$ are the eigenvectors of covariance matrix C that is sorted in descending order according to their corresponding eigenvalues, the feature vectors $Y=\{y_1, y_2, \cdots, y_M\}$ are obtained by projecting images into the eigenspaces following the linear transformation.

$$y_i = E^T(x_i - \bar{x}) \quad (2)$$

In the face-based emotion recognition, we use the 2D-PCA [7] parameter as feature. Generally, 2D-PCA is known as superior performance than PCA in face recognition. In face recognition, 2D-PCA was proposed to decrease the computational cost of conventional PCA. Unlike PCA, which treats 2D images as 1D image vectors, 2D-PCA views an image as a matrix. Consider a m by n random image matrix A. Let $X \in R^{n \times d}$ be a matrix with orthonormal columns, $n \ge d$. Projecting A onto X yields a m by d matrix $Y=AX$ In 2D-PCA, the total scatter of the projected samples was used to determine a good projection matrix X. Suppose that there are M training face images, denoted by m by n matrices A_k $(k = 1, 2, \cdots, M)$, and denote the average image as $\bar{A} = 1/M\sum_k A_k$ Then, the image covariance matrix, G can be evaluated given by

$$G = \frac{1}{M}\sum_{k=1}^{M}(A_k-\bar{A})^T(A_k-\bar{A}). \quad (3)$$

It has been proven that the optimal value for the projection matrix X_{opt} is composed by the orthogonal eigenvectors $X_1, X_2, \cdots, X_d$ of G corresponding to the d largest eigenvalues, i.e. $X_{opt} = [X_1, X_2, \cdots, X_d]$. Because the size of G is only n by n, computing its eigenvectors is very efficient. In addition, a nearest neighbor algorithm is employed to classify the unknown emotion state.

B. Nearest Neighbor

After feature extraction by 2D-PCA, a nearest neighbour classifier is used to measure the similarity between the training feature and test feature. The nearest neighbor classifier is a simple classifier that requires no specific training phase. Here the distance between two arbitrary feature vectors, Y_i and Y_j is defined by

$$d(Y_i, Y_j) = \sqrt{\sum_{p=1}^{P}(Y_i^p - Y_j^p)^2}, \quad (4)$$

where P denotes the feature dimension. Assume that the features of training images are $Y_1, Y_2, \cdots, Y_M$, where M is the total number of training images, and that each of these features is assigned a given identity (class) w_k. If a feature vector of test image is given Y, the recognition result is decided finding the class, w_k that has a minimum distance as follows.

$$d(Y, Y_r) = \min_j d(Y, Y_j), \text{and } Y_r \in w_k, \; then \; Y \in w_k \quad (5)$$

III. Emotion Recognition by Speech

A. MFCC

MFCC is the most widely used spectral representation of the speech signal in many applications, such as speech recognition and speaker recognition. To obtain MFCC, the speech signal is first pre-emphasized using a pre-emphasis filter to spectrally flatten the signal. The pre-emphasized speech signal is then separated into short segments called frames. In our research, we set the frame length to 32ms to guarantee that the signal was stationary within the frame. We also set a 16ms overlap between two adjacent frames to ensure that the signal was stationary between frames. A frame can be seen as the result of multiplying the speech waveform by a rectangular pulse that has a width equal to the frame length. This typically leads to significant high-frequency noise at the beginning and at the end of the frame due to the sudden changes from zero to signal and from signal to zero. Thus, a Hamming window is applied to each frame to reduce the familiar edge effects. In the next step, the pre-processed speech signal is converted into the frequency domain using a discrete Fourier transform (DFT), and a log magnitude of the complex signal in the frequency domain is obtained. Mel-scaling is then performed on the log magnitude signal using triangular filters that are linearly spaced from 0 kHz to 1 kHz, and placed non-linearly according to the mel-scaling approximations [8]. The mel-scale is obtained by

$$f_{mel} = 2595\log(1+\frac{f}{700}), \qquad (6)$$

where f_{mel} denotes the frequency in mels, and f is the input frequency in Hertz. The resultant signal of filtering is then transformed using an inverse DFT into the cepstral domain. The lower order coefficients are selected as the feature parameter, i.e. MFCC, of the speech signal. In this work, we introduce 39 orders of MFCC coefficients by computing the first and second derivatives.

B. Gaussian Mixture Model

Next, speech signal is modeled by GMM algorithm, in which MFCC is used as observation vecotor of GMM. GMM are commonly used to model conditional distributions of acoustic features in the utterance given emotion categories, so this paper employ the GMM to model the speech signal [9]. GMM is usually comprised of a weighted sum of M component densities given by

$$P(X|\lambda) = \sum_{i=1}^{M} c_i b_i(x) \ , \qquad (7)$$

where x is random vector of d-dimension, $b_i(x)$, $i = 1,2,...,M$, is component density, c_i, $i = 1,2,...,M$, denote the mixture weight, and M means mixture number. Each component density consists of the d-dimensional vector is represented as a d-variate Gaussian function as follows:

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{d/2}} \exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)], \qquad (8)$$

where μ_i is the mean vector and Σ_i is the covariance matrix. The mixture weights must satisfy the following constraint:

$$\sum_{i=1}^{M} c_i = 1 \qquad (9)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented as follows:

$$\lambda = \{c_i, \mu_i, \Sigma_i\}, \qquad i = 1,2,...,M \qquad (10)$$

These parameters are estimated by the expectation maximization (EM) algorithm using the training speech signals, and each emotion is represented by a GMM model λ. For a sequence of T test vector $X = x_1, x_2, ..., x_T$, the GMM likelihood of the log domain is calculated by the following equation, and is then utilized to recognize unknown emotion states.

$$L(X|\lambda) = \log p(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \qquad (11)$$

IV. Emotion Recognition by EEG Signal

A. TFCC

To present the EEG signal, we employ the TFCC as feature, and employ the k-nearest neighbor classifier to recognize emotion states of a user. In Fig. 1, we depict the whole architecture of the EEG-based emotion recognition approach. First, triangular filter-based cepstral coefficient of 12 dimensions is computed for each channel of EEG signal. For example, the triangular filter-based cepstral coefficients of channel 1 and 2 are presented as $m_1^1, m_2^1, \cdots, m_{12}^1$ and $m_1^2, m_2^2, \cdots, m_{12}^2$. After extracting the parameters of triangular filter-based cepstral coefficient in each channel, these parameters are combined as a new feature set by simple concatenation as shown in Fig. 1. Similar to MFCC in speech, the triangular filter-based cepstral coefficient is extracted as feature parameter of EEG signal. Remark that the different point between TFCC and MFCC is that triangular filter number is not in accordance caused by different sampling rates.

B. K-Nearest Neighbor

In this paper, we apply k-nearest neighbor classifier to recognize the emotion state using EEG signals. In k-nearest neighbor classifier, each training feature is represented in d-dimensional space according to the value of each of its d features. The test pattern is then represented in the same space, and its k nearest neighbors are selected. Neighbors are usually selected by computing Euclidean distance. The class of each of these neighbors is then tallied, and the class with the most "votes" is selected as the classification of the test pattern. In addition, the simplicity of the k-NN classifier makes it easy to implement.

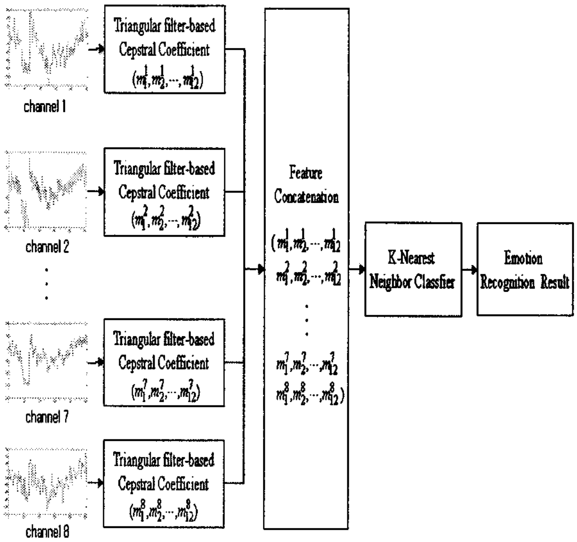

Triangular filter-based Cepstral Coefficient $(m_1^1, m_2^1, \cdots, m_{12}^1)$
channel 1
Triangular filter-based Cepstral Coefficient $(m_1^2, m_2^2, \cdots, m_{12}^2)$
channel 2
Triangular filter-based Cepstral Coefficient $(m_1^7, m_2^7, \cdots, m_{12}^7)$
channel 7
Triangular filter-based Cepstral Coefficient $(m_1^8, m_2^8, \cdots, m_{12}^8)$
channel 8
Feature Concatenation
$(m_1^1, m_2^1, \cdots, m_{12}^1$
$m_1^2, m_2^2, \cdots, m_{12}^2$
$m_1^7, m_2^7, \cdots, m_{12}^7$
$m_1^8, m_2^8, \cdots, m_{12}^8)$
K-Nearest Neighbor Classfier
Emotion Recognition Result

Fig. 1. Block diagram of EEG-based emotion recognition

V. Multimodal Emotion Recognition

In this paper, we compose the multimodal emotion recognition by combining the results of face, speech and EEG recognition to improve system performance, and show a block diagram of system in Fig. 2.

Before integrating the scores of the multiple modalities into a single score, it is needed the score normalization process because the score obtained from individual modalities may not be homogeneous. Score normalization usually involves mapping the scores, i.e., raw-scores, obtained from multiple modalities into the scores, i.e., normalized-scores, of common domain using various statistical techniques. Score normalization generally has various approaches such as z-score, min-max, decimal change, and sigmoid function. In this paper, we employ the approach using sigmoid function to normalize the raw-scores of face, speech, and EEG modalities. The score normalization using the sigmoid function maps the raw scores into the [0, 1] interval [10], and is defined by

$$o_i = \frac{1}{1+\exp(-\tau_i(o_{i,orig}))}, \qquad (12)$$

where $\tau_i(o_{i,orig})$ is defined $[o_{i,orig} - (\mu_i - 2\sigma_i)]/2\sigma_i$, $o_{i,orig}$ is the raw score of *i-th* modality, o_i is the normalized score, and μ_i and σ_i are the mean and standard deviation of the raw scores, respectively .

Once normalized, the normalized scores of multiple modalities can be combined by using a various score level fusion schemes. In this paper, we employ the fusion scheme using weighted-summation rule for fusing the normalized-scores of face, speech, and EEG modalities. The weighted-summation fusion has an advantage of no requiring any training phase in advance [11]. The weighted-summation fusion is expressed as

$$S = \sum_{n=0}^{N} w_n S_n \qquad (13)$$

where S_n is the normalized score of *n-th* modality, and S denotes the fused score. In addition, w_n denotes the weighting coefficient for the *n-th* modality, such that $\Sigma_n w_n = 1$

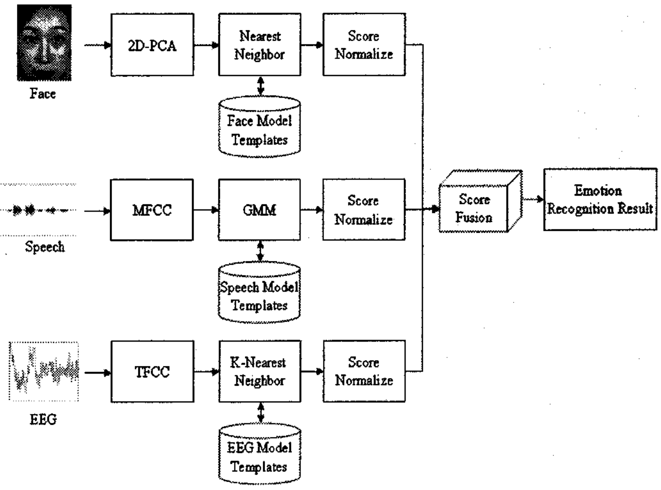


Fig. 2. System architecture of proposed system

VI. Experimental Results

A. Data Set

To evaluate recognition performance of proposed system, we used JAFFE facial database [12], speech database of Yonsei University [13], and EEG database constructed by Sungkyunkwan University as face, speech and EEG database, respectively. In addition, the emotion categories are classified four basic emotions: neutrality, happiness, sadness, and anger.

The JAFFE dataset contains 213 images of seven facial expressions posed by 10 Japanese female models. However, we employed four emotion such as neutrality, happiness, sadness, and anger to multimodal fusion approach. To evaluate performance of speech-based emotion recognition, the training data consists of 35 sentences for each emotion, while the test data consists of 10 sentences for each emotion excluding the training data. The speakers consist of 15 men and 15 women. This paper used a database produced by Professor C. Y. Lee, media and the Communication Signal Processing Laboratory of Yonsei University in Korea with the support of the Korea Research Institute of Standards and Science. This data covers the four emotions of neutrality, happiness, sadness, and anger.

EEG signals were recorded with a QEEG-8 system using a cap with 8 integrated electrodes located at standard positions for five person. The sampling rate was 256Hz, and signals were acquired at full DC. This database consisted of two data sets from five normal subjects during four emotion, i.e., neutrality, happiness, sadness, and anger in which subject set in a normal chair, relaxed arms resting on their legs. The self-assessment of a person may vary due to many different factors, e.g. daily conditions. Therefore, we needed a reliable and valid emotion induction method with the possibility to replicate the experiment. For this reason, we decided to use pictures from the international affective picture system (IAPS) [14] for emotion induction. The IAPS is a set of more than 1000 photographs with emotionally loaded content for the study of emotion and attention. In this paper, we selected 10 pictures from the four categories for each emotion induction. The eight EEG channels were placed at Fpl, Fp2, F3, F4, T3, T4, O1, and O2 positions respectively, according to the international 10/20 system [15]. The data set is acquired on the same day for a user, and the user performed a given task, i.e. EEG acquisition for 15 seconds after user is stimulated during 30 seconds as shown in Fig. 3. Then, the user peformed randomly to another task at the operator's request. Through such process, EEG database about four basic emotion is acquired.

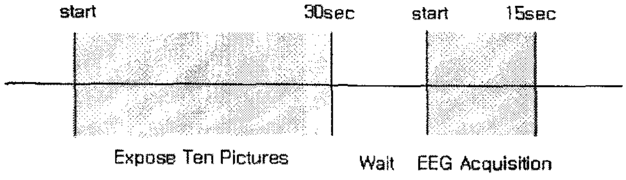
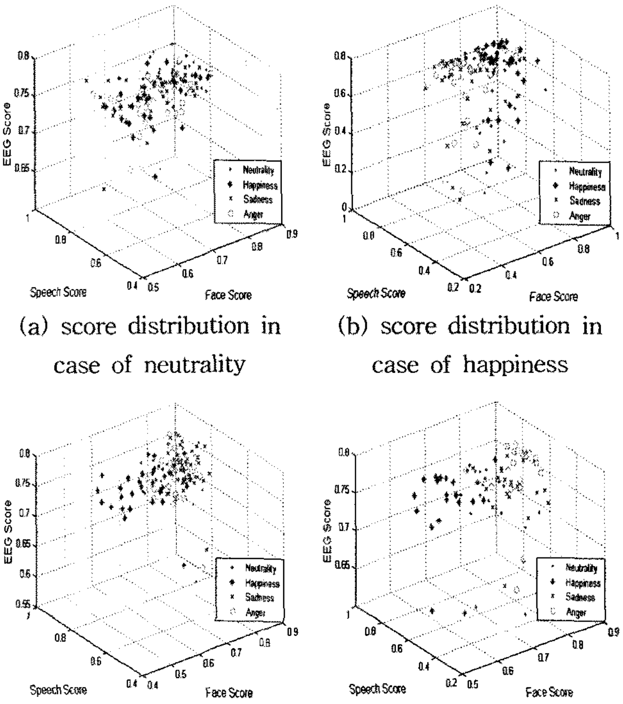


Fig. 3. EEG acquisition procedure

B. Performance Evaluation

First of all, we investigate a three-dimensional distribution of the scores for each emotion as shown in Fig 4. The scores of each modality have a range between zero and one, since the scores are normalized by the sigmoid function. In Fig. 4(a), the normalized scores such as happiness, sadness and anger emotions are distributed in a region of smaller values, while the scores of neutrality emotion has larger values. Similar to Fig. 4(a), we can observe that the scores are distributed in a region of larger values when having same emotion in Fig. 4(b), 4(c), and 4(d). Since the scores between same emotion and other emotion can be considerably separated in three-dimensional space, we can expect performance improvement using a weight-summation classifier. In addition, we perform a statistical analysis against scores of each modality by using mean and variance measurements, and the analysis results for face, speech and EEG modalities are illustrated in Table 1, 2, and 3, respectively. In Table 1, 2 and 3, we can be confirmed that the mean value of same emotion has larger than when is different emotion.



(a) score distribution in case of neutrality

(b) score distribution in case of happiness

(c) score distribution in case of sadness

(d) score distribution in case of anger

Fig. 4. Scatter plot showing the normalized scores for each emotion in three-dimensional space

Table 1. Statistical analysis of face score

Emotion	**Neutrality**			
	Neutrality	Happiness	Sadness	Anger
Mean	0.78645	0.70484	0.71583	0.71326
Variance	0.00213	0.00511	0.00879	0.00539
Emotion	**Happiness**			
	Neutrality	**Happiness**	Sadness	Anger
Mean	0.69281	0.79234	0.59155	0.61215
Variance	0.00687	0.00293	0.01178	0.01281
Emotion	**Sadness**			
	Neutrality	Happiness	**Sadness**	Anger
Mean	0.76487	0.65442	0.81552	0.75192
Variance	0.00347	0.00796	0.00166	0.00260
Emotion	**Anger**			
	Neutrality	Happiness	Sadness	**Anger**
Mean	0.72002	0.63080	0.78252	0.82877
Variance	0.00316	0.00347	0.00217	0.00151

Table 2. Statistical analysis of speech score

Emotion	**Neutrality**			
	Neutrality	Happiness	Sadness	Anger
Mean	0.789599	0.754444	0.769027	0.738220
Variance	0.003003	0.001310	0.002555	0.001160
Emotion	**Happiness**			
	Neutrality	**Happiness**	Sadness	Anger
Mean	0.651288	0.701040	0.651554	0.688288
Variance	0.015004	0.004191	0.012070	0.003491
Emotion	**Sadness**			
	Neutrality	Happiness	**Sadness**	Anger
Mean	0.799812	0.782453	0.817900	0.767296
Variance	0.001718	0.001450	0.002221	0.001832
Emotion	**Anger**			
	Neutrality	Happiness	Sadness	**Anger**
Mean	0.558190	0.646567	0.561949	0.659777
Variance	0.018235	0.006194	0.013780	0.005876

Table 3. Statistical analysis of EEG score

Emotion	**Neutrality**			
	Neutrality	Happiness	Sadness	Anger
Mean	0.75371	0.74423	0.74630	0.74714
Variance	0.00086	0.00091	0.00081	0.00060
Emotion	**Happiness**			
	Neutrality	**Happiness**	Sadness	Anger
Mean	0.65547	0.66222	0.65653	0.65583
Variance	0.03645	0.03366	0.03598	0.03506
Emotion	**Sadness**			
	Neutrality	Happiness	**Sadness**	Anger
Mean	0.74096	0.74397	0.74575	0.74512
Variance	0.00139	0.00103	0.00107	0.00120
Emotion	**Anger**			
	Neutrality	Happiness	Sadness	**Anger**
Mean	0.74974	0.75176	0.74632	0.75481
Variance	0.00224	0.00225	0.00215	0.00195

In the first experiment, the emotion recognition accuracy using a single modality such as face, speech is performed. In the face-based emotion recognition, we perform the classification experiments under different dimensions to further disclose the relationship between the recognition rate and dimension of feature vectors, Fig. 5 show the average recognition rates in case of using the algorithm the PCA and 2D-PCA. Generally, 2D-PCA feature set outperformed PCA. In result, the PCA reveal maximum rate of 64.25%, while 2D-PCA show maximum rate of 68.75%. In case of 2D-PCA, the confusion matrix is additionally shown in Table 4. The speech-based emotion recognition employes the MFCC as feature, and the GMM algorithm as classifier. In result, the emotion recognition result revealed 65.75%, and the result is shown in Table 5 by using confusion matrix. Also, the EEG-based emotion recognition employes the TFCC as feature, and k-nearest neighbor classifier to recognize emotion states of a user. Consequently, the emotion recognition result is shown in Table 6, and the average recognition rate reveal the 57.25%.

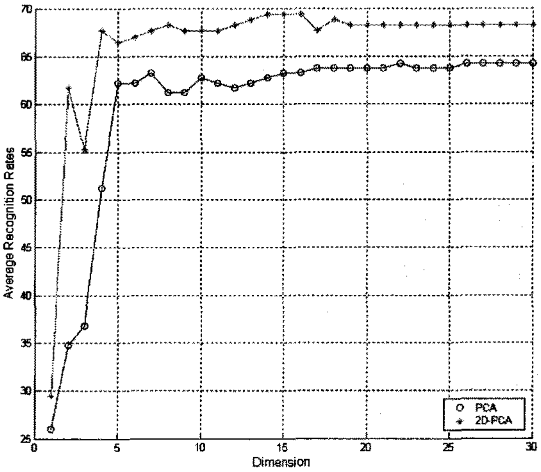


Fig. 5. Face-based recognition result using PCA and 2D-PCA

Next, we perform the various fusion experiments along with different weight: 1) fusion the face and speech modalities, 2) fusion the face and EEG modalities, 3) fusion the speech and EEG modalities, 4) fusion the face, speech, and EEG modalities. The fusion process of the proposed multimodal emotion recognition approach utilizes the sigmoid function-based normalization technology and weighted-summation rule. The weight values of each modality are based on the performance of their modalities. In score normalization procedure, the scores of each modality mean the probability values or distance values. The face-based and EEG-based emotion recognition module produces the distance similarity values, while the speech-based emotion recognition module produces the probability values. Thus, the scores of these modalities cannot be viewed in the same numerical range, and these scores follow different statistical distributions. For these reasons, it is essential that score normalization takes place prior to combining the scores of the individual modalities. Once normalized, the normalized scores of multiple modalities can be combined using various score level fusion schemes. In this paper, the normalized-scores obtained from face, speech and EEG are combined using a simple weighted-summation operation. This fusion method has an advantage that it does not require any training phase.

In the experiment results, the corresponding emotion recognition accuracy for face/speech, face/EEG, and speech/EEG show 83.0%, 75.35% and 73.59% as shown in Tables 7-9, respectively. Also, the emotion recognition rate for integrating three modalities reveals 87.14% as shown in Table 10. In contrast, the emotion recognition accuracy using a single modality such as face, speech and EEG are 68.75%, 65.75% and 57.25%, respectively. From these results, it is apparent that the multimodal approach employing the weight-summation rule in fusion phase outperform the unimodal approaches. Note that the recognition accuracy for integrating three modalities show maximum recognition accuracy as 87.14%. Consequently, we confirmed that the emotion recognition rate of the proposed approach was superior to approaches of single modality and approaches of integrated two modalities each other. From the experimental results. we demonstrate the effectiveness of the proposed multimodal emotion recognition approach by fusing three modalities.

VII. Conclusion

In this paper, we propose the multimodal emotion recognition system using face, speech and EEG. to improve system performance. The distance measurement of the face-based emotion recognition, is calculated by two-dimensional principal component analysis (2D-PCA) and nearest neighbor classifier, and also the likelihood measurement is obtained by Gaussian mixture model (GMM) algorithm based on mel-frequency cepstral coefficient (MFCC) in speech-based emotion recognition. In EEG-based emotion recognition, the distance value is calculated by triangular filter-based cepstral coefficient (TFCC) and k-nearest neighbour classifier. The individual matching scores obtained from face, speech, and EEG are combined using a weighted-summation operation, and the fused-score is utilized to classify the human emotion. From the experiment results, the proposed approach gives an improvement of more than 18.64% when compared to the most successful unimodal approach, and also provides better performance compare to approaches integrated two modalities each other. From these results, we confirmed that the proposed approach achieved a significant performance improvement and the proposed method was very effective.

Table 4. Confusion matrix for face-based emotion recognition using 2D-PCA

Emotion	Neutrality	Happiness	Sadness	Anger
Neutrality	**34.0%**	16.0%	22.0%	28.0%
Happiness	10.0%	**76.0%**	2.0%	12.0%
Sadness	2.0%	2.0%	**62.0%**	34.0%
Anger	5%	0.0%	10.0%	**85.0%**

Table 5. Confusion matrix for speech-based emotion recognition using MFCC

Emotion	Neutrality	Happiness	Sadness	Anger
Neutrality	**56.0%**	32.0%	12.0%	0.0%
Happiness	12.0%	**60.0%**	2.0%	26.0%
Sadness	12.0%	14.0%	**72.0%**	2.0%
Anger	5.0%	20.0%	0.0%	**75.0%**

Table 6. Confusion matrix for EEG-based emotion recognition using TFCC

Emotion	Neutrality	Happiness	Sadness	Anger
Neutrality	**58.0%**	2.0%	10.0%	30.0%
Happiness	0.0%	**47.0%**	25.0%	28.0%
Sadness	2.0%	10.0%	**54.0%**	34.0%
Anger	5.0%	15.0%	10.0%	**70.0%**

Table 7. Emotion recognition accuracy by integrating face and speech along with different weight

Weight of Face Modality	Weight of Speech Modality	Recognition Accuracy (%)
0.00	1.00	65.75
0.10	0.90	78.00
0.20	0.80	80.00
0.30	**0.70**	**83.00**
0.40	0.60	79.50
0.50	0.50	79.00
0.60	0.40	76.50
0.70	0.30	75.75
0.80	0.20	72.50
0.90	0.10	71.50
1.00	0.00	68.75

Table 8. Emotion recognition accuracy by integrating face and EEG along with different weight

Weight of Face Modality	Weight of EEG Modality	Recognition Accuracy (%)
0.00	1.00	57.25
0.10	0.90	67.05
0.20	0.80	71.05
0.30	0.70	73.81
0.40	0.60	73.11
0.50	0.50	73.60
0.60	0.40	74.85
0.70	**0.30**	**75.35**
0.80	0.20	73.34
0.90	0.10	72.32
1.00	0.00	68.75

Table 9. Emotion recognition accuracy by integrating speech and EEG along with different weight

Weight of Speech Modality	Weight of EEG Modality	Recognition Accuracy (%)
0.00	1.00	57.25
0.10	0.90	61.06
0.20	0.80	65.06
0.30	0.70	66.06
0.40	0.60	67.57
0.50	0.50	70.07
0.60	0.40	71.07
0.70	**0.30**	**73.59**
0.80	0.20	71.84
0.90	0.10	71.11
1.00	0.00	65.75

Table 10. Emotion recognition accuracy by integrating face, speech and EEG along with different weight

Face Weight	Speech Weight	EEG Weight	Accuracy(%)
0.00	0.00	1.00	57.25
0.00	0.10	0.90	61.06
⋮	⋮	⋮	⋮
0.20	0.50	0.30	85.16
0.20	**0.60**	**0.20**	**87.14**
0.20	0.70	0.10	87.06
⋮	⋮	⋮	⋮
0.90	0.00	0.10	72.32
0.90	0.10	0.00	71.32
1.00	0.00	0.00	68.75

References

[1] Picard, R., "Affective computing", MIT Press, Boston, 1997.

[2] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., "Emotion recognition in human-computer interaction". IEEE Signal Processing Magazine, 2001.

[3] M. Song, J. Bu, C. Chen, N. Li, "Audio-visual based emotion recognition-a new approach", Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1020-1025, 2004.

[4] Chang Ryu, Min An, Yoon Na, Jin Cho, Young Han, Kyoung Kim and Pyung Park, "A portable neurofeedback system and EEG-analysis methods for evaluation", World Congress on Medical Physics and Biomedical Engineering 2006, vol. 14, no. 8, pp. 1167-1169, 2007.

[5] Schaaff, K., Schultz, T., "Towards an EEG-based emotion recognizer for humanoid robots", The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 792-796, 2009.

[6] M.Turk, A. Pentland, "Eigenfaces for recognition", J. Cognitive Neuro-sci,, vol. 3, no. 1, pp. 71-86, 1991.

[7] Y. Jian, Z. David, F. Alejandro, and J. Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, pp. 131-137, 2004.

[8] D. O'Shaughnessy, "Speech Communication - Human and Machine", Addison-Wesley, New York, 1987.

[9] Reynolds, D. A., Quatieri, T., Dunn, R., "Speaker verification using adapted Gaussian mixture models", Digital Signal Process, vol. 10, pp. 19-41, 2000.

[10] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, "Acoustic-labial speaker authentication", Pattern Recognition Letter. vol. 18. pp. 853-858, 1997.

[11] A. Ross, A.K. Jain, "Information fusion in biometrics", Pattern Recognition Letter, vol. 24, pp. 2115-2125, 2003.

[12] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi & Jiro Gyoba, "Coding Facial Expressions with Gabor Wavelets", Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205, 1998.

[13] Kyung Hak Hyun, Eun Ho Kim, Yoon Keun Kwak, "Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept", IEEE International Workshop on Robot and Human Interactive Communication, pp. 312-316, 2005.

[14] P. Lang, M. Bradley, and B. Cuthbert. "International affective picture system (IAPS): Affective ratings of

pictures and instruction manual", Technical Report A-6, University of Florida, Gainesville, 2005.

[15] Papanicolaou A. C, Loring D. W, Deutsch G, Eisenberg H. M, "Task-related EEG Asymmetries: a Comparison of Alpha Blocking and Beta Enhancement", International Journal of Neuroscience, vol. 30, pp. 81-85, 1998.

Dong-Ju Kim

Regular member

Received the B.S. and M.S. degrees in Radio and Science Engineering from ChungBuk National University in 1998 and 2000, respectively. He received the Ph.D degree in Department of Electrical and Computer Engineering from Sungkyunkwan University in 2010.

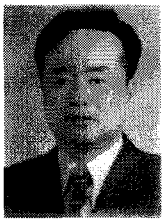
His current research focuses on digital signal processing, biometrics, and pattern recognition.

Hyeon-Gu Lee

Regular member

Received the B.S., M.S. and Ph.D degree in Department of Electronic Engineering from Sungkyunkwan University in 1989, 1991 and 2000. He is a professor of Department of Information and Communication Engineering at Seoil University.

His current research focuses on communication and signal processing, multimedia communication.



Kwing-Seok Hong

Regular member

Received the B. S., M. S., and Ph.D. in electronic engineering from the Sungkyunkwan University, Seoul, Korea, in 1985, 1988, and 1992, respectively. In March 1990, he joined the Seoul Health College, Seoul, Korea, where he was a full time lecturer of Computer Engineering. From March 1993 to February 1995, he was a full time lecturer at Cheju University, Cheju, Korea. Since March 1995, he has been a professor at Sungkyunkwan University, Suwon, Korea. His current research focuses on recognition and integration and representation of the five-senses.