Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

음성 신호와 얼굴 영상을 이용한 특징 및 결정 융합 기반 감정 인식 방법

Emotion Recognition Method based on Feature and Decision Fusion using Speech Signal and Facial Image

주종태, 양현창, 심귀보

Jong-Tae Joo, Hyun-Chang Yang, and Kwee-Bo Sim

중앙대학교 전자전기공학부

(E-mail: kbsim@cau.ac.kr)

요 약

인간과 컴퓨터간의 상호교류 하는데 있어서 감정 인식은 필수라 하겠다. 그래서 본 논문에서는 음성 신호 및 얼굴 영상을 BL(Bayesian Learning)과 PCA(Principal Component Analysis)에 적용하여 5가지 감정(Normal, Happy, Sad, Anger, Surprise)으로 패턴 분류하였다. 그리고 각각 신호의 단점을 보완하고 인식률을 높이기 위해 결정 융합 방법과 특징 융합 방법을 이용하여 감정 융합을 실행하였다. 결정 융합 방법은 각각 인식 시스템을 통해 얻어진 인식 결과 값을 퍼지 소속 함수에 적용하여 감정 융합하였으며, 특징 융합 방법은 SFS(Sequential Forward Selection) 특징 선택 방법을 통해 우수한 특징들을 선택한 후 MLP(Multi Layer Perceptron) 기반 신경망(Neural Networks)에 적용하여 감정 융합을 실행하였다.

Key Words : Emotion Recognition, Speech Signal, Facial Image, Decision Fusion Method, Feature Fusion Method

1. 서 론

컴퓨터(기계) 기술이 점점 발전함에 따라 인간과 컴퓨터(기계) 사이의 상호교류에 대한 연구가 활발히 진행되고 있다. 이러한 연구는 인간에게 보다 더 편리하고 정확한 맞춤형 서비스를 제공하기 위해서 이루어지고 있으며, 그 중에 인간의 감정을 인식하고 표현해주는 기능들은 필수라 하겠다. 그리고 이 기능들을 통해 인간-컴퓨터 사이의 감정적인 교류가 가능해 질 것이라 생각된다.

인간의 신체에서 감정 인식을 할 수 있는 매개체로는 음성, 얼굴 영상, 제스처, 피부 온도 등이 존재하며 기존의 연구에서는 이러한 매개체들을 각각 이용하여 감정 인식 연구가 이루어졌다. 이와 관련 연구들로 Lee C.M. et al와 New T.L. et al은 음성 신호로부터 특징을 추출하는 방법으로 13차와 12차 MFCCs(Mel Frequency Cepstral Coefficients)를 사용하였으며 감정별 패턴 분류는 HMM(Hidden Markov Model)을 이용하였다[1][2].

Mase et al은 얼굴 영상에 지역별로 11개의 windows를 형성한 후 이 windows별로 근육의 움직임 정도를 파악하여 특징을 추출하였다. 그리고 K-nearest neighbor 규칙을 이용하여 감정별 패턴을 분류하였다[3].

이 밖에 제스처 및 피부 온도를 이용하여 감정 인식한 연구 사례는 다음과 같다[4][5]. 하지만 이런 연구들은 각각 매개체들의 단점들을 보완할 수 없으므로 최근에는 감정 융합 방법을 사용하여 감정 인식 실험이 많이 이루어지고 있다.

감정 융합 방법으로는 크게 결정 융합 방법과 특징 융합 방법이 존재하는데 각각의 인식 시스템을 통해 인식된 결과 값을 이용하는 것이 전자의 방법이고, 각각의 매개체에서 특징들을 추출한 후 감정 융합을 하는 것이 후자의 방법이다. 현재 이와 관련 연구 사례로는 다음과 같은 것들이 있으며 Mingli Song은 특징 융합 방법으로 Hidden Markove Model(HMM)을 이용하여 음성과 얼굴 영상에 대한 감정 인식 실험을 하였으며 De silva는 결정 융합 방

감사의 글 : 이 논문은 서울시 산학연 협력사업 (2005년 신기술 연구개발 지원사업, 과제번호 : 106876)에 의해 수행되었습니다. 연구비지원에 감사드립니다.

법으로 퍼지 룰 베이스를 이용하여 음성과 얼굴영상에 대한 감정 인식 실험을 하였다[6][7]. 그리고 Busso는 두 가지 방법에 대해 실험하고 비교 설명하였다[8].

그 결과 특정 한 가지 매개체를 이용하는 경우보다 다양한 매개체를 이용할 때가 감정 인식률이 높음을 알 수가 있었다. 그래서 본 논문 2절에서는 음성 신호 및 얼굴 영상에 대한 각각의 감정 인식 시스템에 대해 설명되어 있고 3절에서는 본 연구에서 사용된 두 가지 감정 융합 방법에 대해 설명되어 있다. 그리고 4절에서 실험 결과를 비교하였으며 5절에서 결론을 내렸다.

2. 음성 신호 및 얼굴 영상을 이용한 감정 인식

2.1 음성 신호

본 논문에서는 피치의 통계치 및 최대치, 소리의 크기, 섹션 개수, Increasing Rate(IR), Crossing Rate(CR)들의 특징들을 음성신호로부터 추출하였고 추출된 특징들을 Bayesian Learning(BL)을 이용하여 감정별 패턴을 분류하였다.

BL은 사전확률을 이용하여 어떤 가설의 확률을 계산하는 방법이다. 그래서 본 논문에서는 400개의 음성 샘플들을 이용하여 각 감정과 특징들 간의 확률 분포를 조사하여 사전 확률을 계산하였다. 그리고 사용자의 확률 분포와의 유사정도를 파악하여 5가지 감정(평활, 기쁨, 슬픔, 놀람, 화)으로 패턴 분류하였다[9].

그림 1은 이와 같은 일련의 과정들을 실험할 수 있는 시스템을 나타내는 그림이다.

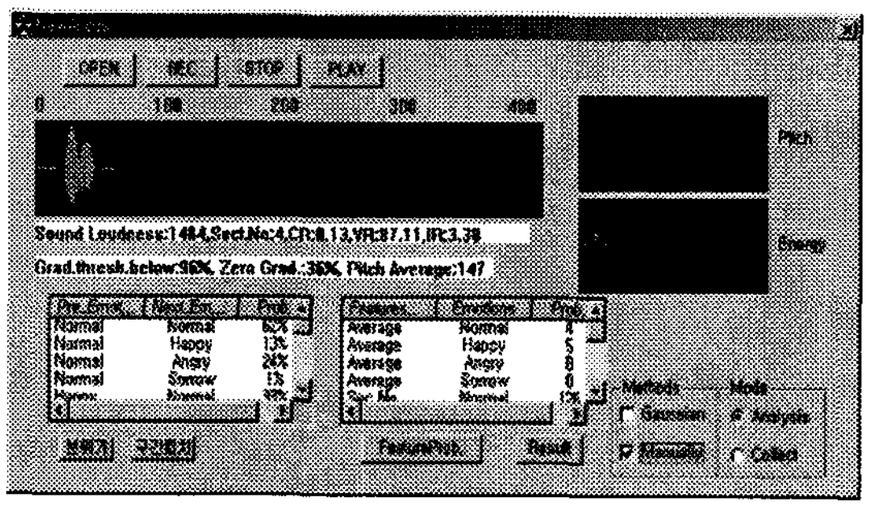

그림 1. 음성신호를 이용한 감정인식 시스템

2.2 얼굴 영상

얼굴 영상을 이용하여 감정 인식을 하기 위해서 특징을 추출해야 하는데 본 논문에서는 피부톤 축적 알고리즘과 GRAY 형태 변환을 이용하여 입, 눈과 눈썹의 특징을 추출하였다.

추출된 특징들은 다차원 특징 벡터로 구성되어져 있어서 패턴을 분류하기에 용이하지 않다. 그래서 정보를 유지하면서 저차원으로 특징 벡터를 축소시키는 방법이 필요한데 본 논문에서는 이 방법으로 Principal Component Analysis (PCA)을 사용하였다.

PCA 알고리즘을 통해 고유 데이터 벡터를 구한 후 유클리안 거리를 통해 학습 데이터와 입력 데이터간의 거리를 비교하여 그 거리가 최소가 되는 표정이 입력과 가장 유사한 표정이므로 그 학습데이터의 감정을 결과로 결정하게 된다[10].

이와 같은 일련의 과정들을 실험할 수 있는 시스템은 그림 2와 같다.

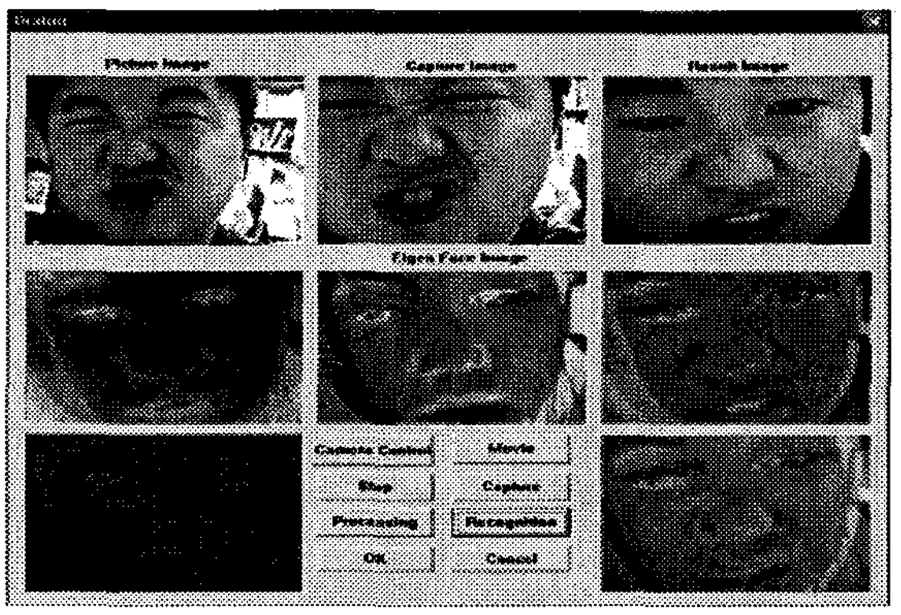

그림 2. 얼굴영상을 이용한 감정인식 시스템

3. 감정 융합 방법을 이용한 감정 인식

인간들은 정확하게 감정을 인식하기 위해서 특정한 한가지의 매개체만을 이용하지 않고 다양한 매개체와 상황들을 고려하여 감정을 인식한다. 그래서 인간-컴퓨터간의 감정 인식에서도 이와 같은 연구가 필요하며 현재 활발히 진행되고 있다.

본 논문에서는 감정 융합을 하기 위해서 퍼지 소속 함수와 Neural Networks를 이용하였다.

3.1 결정 융합 방법

결정 융합 방법은 각각의 감정인식 시스템을 통해 얻어진 결과를 이용하는 방법으로써 구현이 쉽다는 장점은 있으나 각각의 매개체의 단점을 보완하기에는 부족한 면이 있다.

본 논문에서는 결정 융합을 하기 위해서 S-모양의 퍼지 소속 함수를 이용하였다[11].

다양한 퍼지 소속 함수 중 S-모양을 사용하는 이유는 가중치를 통해 인식률을 높일 수 있기 때문이다.

S-모양의 퍼지 소속 함수식은 식(1)과 (2)와 같으며 5가지 감정별로 가중치를 구하게 된다.

$$w_s = \frac{1}{1+\exp[-a(x_s - c_s)]} \quad (1)$$

$$w_i = \frac{1}{1+\exp[-a(x_i - c_i)]} \quad (2)$$

여기서 w_s, w_i는 각각 음성 신호와 얼굴 영상에 대한 가중치이며, x_s, x_i는 각각 매개체의 입력 데이터들을 통해 감정을 인식한 결과이다. 그리고 c_i, c_s는 각각 매개체의 학습 데이터들을 감정을 인식한 후 감정별로 평균을 구한 결과이며 이 값은 실험을 반복함에 따라 입력 데이터가 학습 데이터로 등록됨으로써 변하게 된다. 마지막으로 a는 소속 함수의 기울기 정도를 나타내는데 이 값은 0.01~0.1사이로 값을 변화시켜 가중치의 결과 값이 가장 좋은 것을 선택하였는데 실험 결과 0.05가 가장 우수한 결과를 보였다.

이와 같은 방법으로 가중치를 구한 후 식(3)과 같이 각각의 매개체를 통해 얻어진 결과 값에 곱을 취하여 각각의 감정 상태에 대한 출력이 나오게 되며 식 (3)에서 I는 얼굴 영상에서의 감정 출력이고, S는 음성 신호에서의 감정 출력이다. 그리고 이렇게 인식된 감정들 중 최대값을 선택하여 감정 인식 결과로 나타낸다.

$$\begin{aligned} O_{normal} &= w_{i(normal)} I_{normal} + w_{s(normal)} S_{normal} \\ O_{happy} &= w_{i(happy)} I_{happy} + w_{s(happy)} S_{happy} \\ O_{surprise} &= w_{i(surprise)} I_{surprise} + w_{s(surprise)} S_{surprise} \\ O_{sad} &= w_{i(sad)} I_{sad} + w_{s(sad)} S_{sad} \\ O_{anger} &= w_{i(anger)} I_{anger} + w_{s(anger)} S_{anger} \end{aligned} \quad (3)$$

3.2 특징 융합 방법

특징 융합 방법은 구현은 어렵지만 각각 매개체의 단점을 보완할 수 있다는 장점을 가지고 있다.

본 논문에서는 이러한 특징 융합 방법을 실험하기 위해서 음성 신호와 얼굴 영상에서 특징들을 추출하였는데 그 결과 각각 6가지와 5가지의 특징 벡터를 추출할 수 있었다. 하지만 11가지의 특징 벡터를 모두 고려하면 차원의 저주에 빠질 위험성이 크고 인식 속도가 느려지는 단점이 생길 수 있으므로 특징 선택 방법을 통해 우수한 특징들을 선택하게 된다. 이와 같은 특징 선택 방법으로는 여러 가지가 존재하지만 본 논문에서는 Sequential Forward Selection(SFS) 방법을 이용하였다.

SFS는 비어있는 집합에 순차적으로 특징들을 추가한 후 목적함수에 대입하여 그 결과가 가장 우수한 것들을 특징들로 선택하는 방법이다. 본 논문에서 사용된 목적함수는 식 (4)와 같다.

$$\begin{aligned} y = {} & 3x_0 + x_1 + 4x_2 + 10x_3 - 5x_4 + 8x_5 + 7x_6 \\ & + 8x_7 - 10x_8 + 6.8x_9 + 7.3x_{10} - 5.2x_{11} \end{aligned} \quad (4)$$

이 식에서 y는 목적함수 결과 값이고 x_n은 특징들의 종류를 나타낸다. 그리고 목적 함수 파라미터들을 다음과 같이 표현한 이유는 학습 데이터로부터 각각의 특징들을 추출하여 그 크기가 5번째 안에 있는 것들은 reward(+1.0)을 주고 그 이후에 있는 것들은 penalty(-1.0)를 주었다. 이와 같은 실험을 100번 반복한 결과 값이다.

이와 같이 특징들이 결정되면 이 값들을 인공 신경망 중 Back-Propagation(BP)로 학습하는 Multi Layer Perceptron(MLP)에 입력으로 설정하여 감정별 패턴을 분류하였다.

다음 표1은 본 논문에서 사용된 초기 파라미터 설정 값들을 나타내고 있다.

표 1. 신경망의 초기 파라미터 설정

Parameter	Value
Hidden Units	13
Output Units	5
Learning Rate	0.005
Tolerance	0.1
Sigmoid Function	$1/1+e^{-3x}$

초기 파라미터를 설정한 후 학습데이터들을 이용해 오차 범위보다 작아질 때까지 학습을 시킨다. 그리고 입력 데이터를 입력한 후 감정별 패턴을 분류한다. 다음 그림 3은 결정 융합과 특징 융합을 모두 실험할 수 있는 시스템을 나타내는 그림이다.

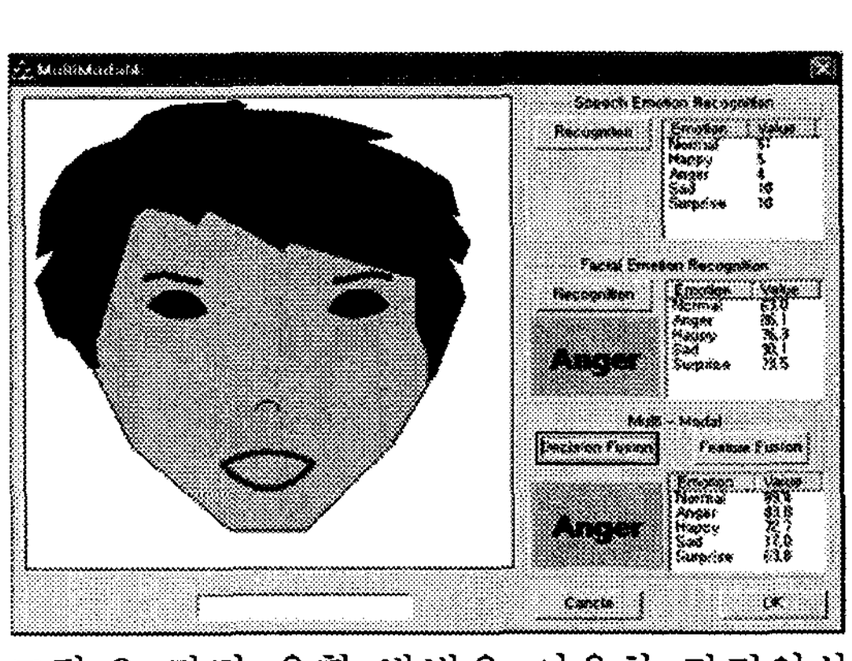

그림 3. 감정 융합 방법을 이용한 감정인식 시스템

4. 실험 결과

본 논문에서는 음성 신호와 얼굴 영상을 이용하여 감정 인식을 하기 위해 동일한 환경에서 감정별 Database를 구축한 후 실험을 하였다. 음성 신호의 경우 10명의 남성들에게 5가지 감정별로 40개의 음성 샘플을 얻었으며, 얼굴 영상의 경우 10명의 남성들에게도 5가지 감정별로 6개의 얼굴 영상을 촬영하여 구축하였다. 구축된 Database을 이용하여 30번의 감정 인식 실험을 하여 평균을 구한 결과는 그림 4와 같다.

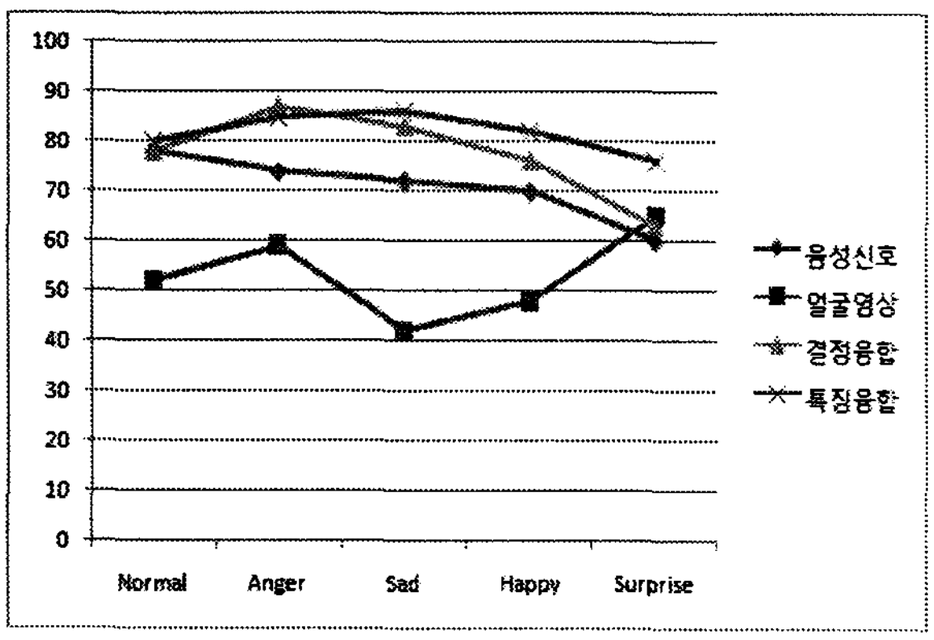

그림 4. 감정 인식 결과

실험 결과 음성 신호만을 이용하여 감정 인식한 경우 평균 인식률이 70.8%였으며, 얼굴 영상한 경우는 53.2%였다. 이와 같이 낮은 인식률을 높이기 위해서 본 논문에서는 결정 융합 방법과 특징 융합 방법을 이용하였으며 평균 감정 인식률은 77.4%, 81.8%였다.

이 결과를 통해 한가지의 매개체만을 이용하는 경우보다 다양한 매개체를 이용하여 감정 인식 하는 경우가 성능 면에서 우수함을 알 수 있었다. 그리고 결정 융합에 비해 특징 융합 방법이 각각의 매개체들의 단점들을 보완함으로써 우수한 성능을 보임을 알 수 있었다.

5. 결론 및 향후과제

본 논문에서는 음성 신호와 얼굴 영상을 이용하여 5가지 감정[평활, 기쁨, 화남, 슬픔, 놀람]으로 패턴 분류하였으며 결정 융합 방법과 특징 융합 방법을 통해 감정 융합을 실험하였다.

실험 결과 인간-인간 사이에서의 감정 인식처럼 인간-컴퓨터간의 감정인식에서도 다양한 매개체를 고려하는 것이 성능 적으로 우수함을 알 수 있었다. 그리고 각각의 성능을 비교해 봄으로써 감정 융합 방법의 특징들을 알 수 있었다.

차후 연구로는 획일화된 환경이 아니라 다양한 환경에서 실시간으로 감정을 인식하는 실험을 할 것이며 좀 더 다양한 결정융합 방법과 특징융합 방법을 실험하고 분석하여 우수한 감정 융합 방법을 제안할 것이다.

참 고 문 헌

[1] Lee C.M., Narayanan S.S. and Pieraccini. R., "Classifying emotions in human - machine spoken dialogs", *ICME'02*, vol. 1, pp. 737-740, 2002.

[2] New T.L., Wei F.S. and De Silva L.C., "Speech based emotion classification", *TENCON 2001*, vol. 1, pp. 297-301, 2001.

[3] Mase K., "Recognition of facial expression from optical flow", *IEICE Trans.*, vol. 74, no. 10, pp 3474-3483, 1991.

[4] H. Guncs and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gesture", *Journal of Network and Computer Application*, pp. 1-12, 2006.

[5] Yoshitomi Y, S. I. Kim, Kawano T and Kilazoe T, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face", *Robot and Human Interactive Communication 2000*, pp. 178-184, 2000.

[6] Mingli Song, Jiajun Bu, Chun Chen and Nan Li, "Audio-visual based emotion recognition", *CVPR'04*, vol. 2, pp. 1020-1025, 2004.

[7] D. Silval and P. C. Nag, "Bimodal emotion recognition", *Proc. of Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pp. 332-335, 2000.

[8] Carlos Busso and Zhigang Deng et al, "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", *ICMI 2004*, pp. 205-211, 2004

[9] C. H. Park and K. B. Sim, "Pattern Recognition Methods for Emotion Recognition with speech signal", *International Journal of Fuzzy Logic and Intelligent Systems*", vol. 6, no. 2, pp. 150-154, 2006.

[10] Ho-Duck Kim, Hyun-Chang Yang, Chang-Hyun Park, and Kwee-Bo Sim, "Emotion Recognition Method of Facial Image using PCA ", *Journal of Korea Fuzzy Logic and Intelligent Systems Society(KFIS)*, vol.16, no.6, pp. 772-776, Dec. 2006.

[11] Hyeun-Joo Go, Dae-Jong Lee, Myung-Geun Chun "An Emotion Recognition Method using Facial Expression and Speech Signal", *Journal of Korea Information Science Society(KISS)*, vol. 31, no. 6, pp. 799-807, 2004.