Proceedings of KFIS Spring Conference 2007 Vol. 17, No. 1.

음성 신호와 얼굴 표정을 이용한 감정인식 및 표현 기법

An Emotion Recognition and Expression Method using Facial Image and Speech Signal

주종태 · 문병현 · 서상욱 · 장인훈 · 심귀보

중앙대학교 전자전기공학부
kbsim@cau.ac.kr

요 약

본 논문에서는 감정인식 분야에서 가장 많이 사용되어지는 음성신호와 얼굴영상을 가지고 4개의 (기쁨, 슬픔, 화남, 놀람) 감정으로 인식하고 각각 얻어진 감정인식 결과를 Multi modal 기법을 이용해서 이들의 감정을 융합한다. 이를 위해 얼굴영상을 이용한 감정인식에서는 주성분 분석(Principal Component Analysis)법을 이용해 특징벡터를 추출하고, 음성신호는 언어적 특성을 배재한 acoustic feature를 사용하였으며 이와 같이 추출된 특징들을 각각 신경망에 적용시켜 감정별로 패턴을 분류하였고, 인식된 결과는 감정표현 시스템에 작용하여 감정을 표현하였다.

Key Words : 감정인식, 감정표현, 주성분 분석(PCA), Neural Network

1. 서 론

최근 들어 모든 가전제품이나 로봇, 컴퓨터 등의 급속한 발전으로 인해 인간과의 인터페이스가 중요한 문제로 대두되고 있다.

이러한 인간과의 인터페이스 문제를 해결하기 위해서는 감정인식 및 표현은 필수적이라 할 수 있으며 이와 관련된 연구들도 활발히 이루어지고 있다.

감정을 인식하는 매개체로 크게 음성신호와 얼굴 표정 영상으로 나누어질 수 있는데 먼저 음성신호를 통한 감정인식의 특징들은 피치, 에너지, 포만트, 말의 빠르기로 구성되어지며 연구자에 따라 모두 선택하기도 하고 일부만 선택하기도 한다. 그리고 각 특징 점들에서도 다양한 통계치를 추출하여 사용하는 경우가 일반적이다.

본 논문에서는 언어적 특성을 배제한 acoustic feature 중 피치의 통계치, 소리의 크기, 섹션개수, Increasing Rate(IR), Crossing Rate(CR)들의 특징들을 인공신경망(Artificial Neural Network)에 적용하여 감정을 인식하였다[1]. 다음으로 얼굴 영상을 통해 감정인식을 하는데 있어 특징들을 추출하는 방법으로는 광학적 흐름 분석, 홀리스틱 분석, 국부적인 표현 등이 있으며[2,3,4], 본 논문에서는 홀리스틱 분석법 중 가장 대표적인 PCA 방법[5]을 이용하여 특징을 추출하고 최소거리 분류 방법을 이용하여 감정을 인식하였다.

하지만 보통 사람들이 감정을 인식하는 경우 어떠한 한 가지 특징 정보만 가지고 감정을 인식하지 않고 다채로운 정보를 바탕으로 감정을 인식하게 된다. 그래서 이러한 것들을 시스템에 적용시키기 위한 방법으로 본 논문에서는 음성과 영상을 융합한 Multi-Modal 감성 인식 방법을 제안하고, 이 방법을 통하여 인식된 감정 결과는 본 논문에서 사용된 감정 표현은 본 연구실에서 개발한 감정 표현 방법을 이용하였다[8].

2. 음성신호를 이용한 감정인식

감정인식기는 특징을 추출하는 부분과 그 특징들을 이용하여 패턴을 인식하는 부분으로 나눌 수 있다. 본 논문에서는 autocorrelation approach를 사용하여 추출한 피치의 통계치, 소리의 크기는 magnitude estimation method에 의해서 구했고, 섹션 개수, Increasing Rate(IR), Crossing Rate(CR)들의 특징들을 추출했다[9]. 추출된 특징들을 입력 패턴으로 제시하고, 각 노드에 대해서 입력함수와 활성화 함수를 이용하여 출력을 산출한 후 출력 값이

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

목표 값과 일치하지 않을 경우 차이를 계산하여 오차를 구하고, 이를 역방향으로 연결강도를 갱신한다. 이와 같은 과정을 출력 값이 목표 값보다 작아질 때까지 반복하여 학습이 이루어진다. 이와 같은 방법을 Backpropagation (BP)이라고 하며 본 논문에서는 이 방법을 이용하여 감정을 분류하게 된다.

다음의 표 1은 신경망의 초기 파라메터 값들, 그림 1은 학습에 따른 에러 그래프, 표 2는 신경망 학습 결과를 각각 나타내고 있는데, 본 논문에서는 4가지 감정(기쁨, 슬픔, 놀람, 화)에 대해서 감정을 인식하므로 2개의 이진값 형태로 출력을 나타내었다.

표 1. 신경망 파라메터 설정

Parameter	Value
Input Units	5
Hidden Units	12
Output Units	2
Learning Rate	0.003
Tolerance	0.25
Sigmoid Function	$1/1+e^{-4x}$

표 2. 신경망 학습 결과

Emotion	Expression Patter
(0)0.001813 (0)0.017775	(0)0.063257 (1)0.917163
(1)0.981922 (1)0.999878	(0)0.069832 (1)0.955858

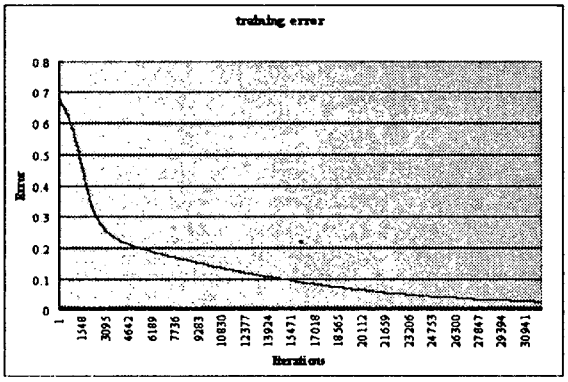

그림 1. 학습 에러 그래프

3. 얼굴표정을 이용한 감정인식

얼굴 영상은 다차원 특징 벡터로 이루어진 데이터이다. 이와 같은 것을 인식하기 위해서는 높은 차원에서의 정보를 유지하면서 낮은 차원으로 축소시키는 다변량 데이터 처리 방법이 필요한데 그중 가장 대표적인 방법이 PCA이다. PCA는 주축을 통계적인 방법에 의하여 구하고 구해진 주축 방향으로 특징 벡터를 사영 시킴으로서 차원을 축소하는 방법이다.

그림 1은 감정인식에 사용될 사진들을 보여주고 있으며 5명의 남성(25세-31세의 다양한 지역 출신의 대학원생)으로부터 5가지 표정(무표정, 기쁨, 화, 놀람, 슬픔)을 연기하도록 하고 사진을 찍었다. 이렇게 수집된 사진을 앞서 설명한 PCA 방법을 사용하였으며 고유 벡터의 수는 임의로 100개로 지정해 주어 실험한 결과 그림 4와 같은 고유 얼굴들이 나타났다.

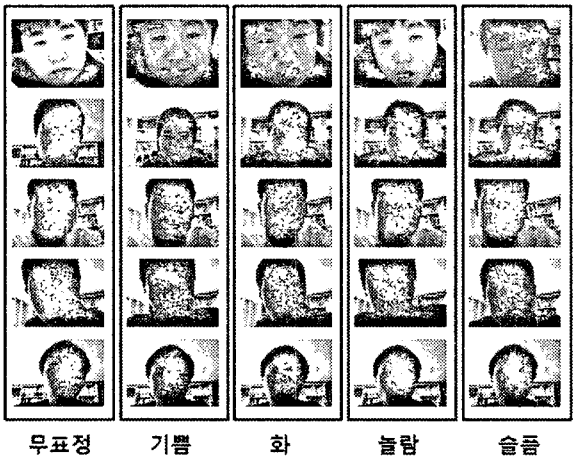

그림 2. 실험에 사용된 감정별 얼굴사진

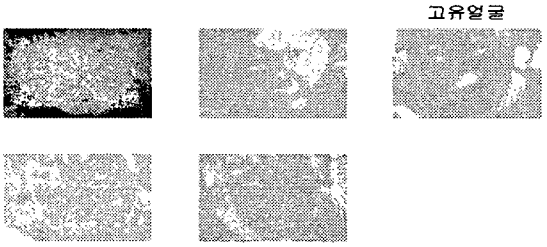

그림 3. PCA에 의해 얻어진 고유 얼굴들

위와 같이 학습을 마치고 검증 영상 중 하나의 얼굴 영상이 입력되면 고유 얼굴에 대한 사영을 취하여 성분값을 구한다. 이 값이 구해지면 후보 얼굴영상들의 고유 얼굴에서의 가중치와 유클리디안 거리를 비교하여 그 거리가 최소가 되는 표정이 입력과 가장 유사한 표정이므로 이 후보를 인식 결과로 결정하게 되며 그 결과는 그림 4 및 5와 같다.

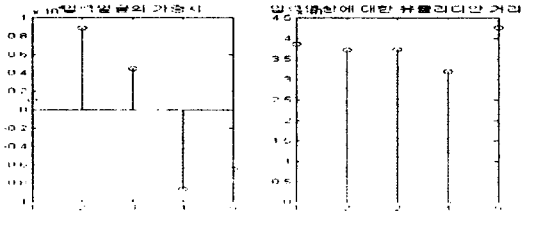

그림 4. PCA 결과 화면

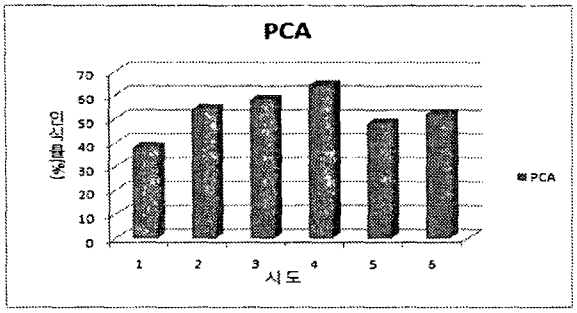

그림 5. PCA 인식율

4. Multi-Modal 감정인식 방법

Multi-Modal이란 여러 가지 정보를 융합하는 것을 말한다. 본 논문에서는 두 가지 특징 정보인 음성과 얼굴 영상의 인식을 융합하였다.

얼굴 영상 정보와 음성 신호 정보의 합성 방법으로는 크게 두 가지로 분류할 수 있다. 인식 전 얼굴 영상의 특징 벡터와 음성 정보의 특징 벡터를 합성하는 방법과 영상 정보와 음성 정보를 각각 인식한 후 가중치를 이용하여 인식 결과 값을 합성하는 방법이 있다.

이 중 전자의 방법의 경우 추출된 특징벡터를 합성하여 새로운 특징벡터로 생성하므로 이를 인식하기 위해서는 성능 좋은 유사도 측정 방법이 제안되어야만 한다. 그래서 이 방법은 특징벡터가 원래 감정의 특징을 가지고 있는지 알기가 어렵고 인식률도 낮다는 단점이 있다.

그림 6는 특징벡터 추출 단계에서 융합하는 방법을 나타낸다.

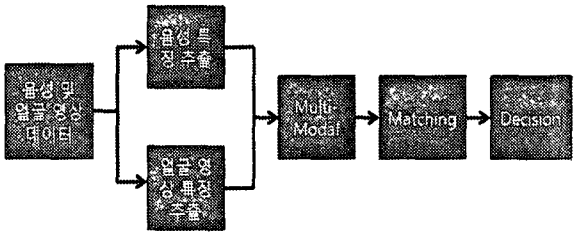

그림 6. 특징벡터 추출 단계에서 융합

한편 후자의 방법은 음성 신호와 얼굴 영상 정보에 대해 각각 감정인식을 한 후 이 값들을 감정별로 데이터베이스를 구축한다.

각각 입력 신호가 들어오면 이를 인식한 후 구축되어진 데이터베이스와 퍼지 소속 함수를 이용하여 감정별 소속도 값을 결정한다. 각각 음성 신호와 얼굴 영상 정보에 대한 소속도 값을 통해 감정인식 결과 값을 출력하게 되는데 이를 하기위한 방법으로는 동일한 감정에 대해 두 개의 소속도 중 큰 값을 선택하는 방법과 각 감정에 대해 두 개의 소속도 값을 모두 더하는 방법이 있을 수 있다.

그림 7은 각각 인식되어진 결과 값을 융합하는 방법을 나타낸다.

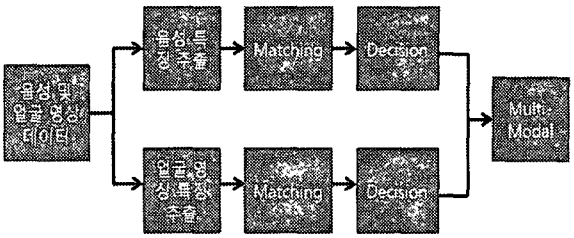

그림 7. 인식되어진 결과 값을 이용한 융합

5. 동적 감정 공간에서의 감정표현

5.1 감정표현 시스템

본 연구에서는 동적으로 변화하는 2차원 감정 공간 모델을 적용하여 인간의 감정 표현 시스템과 유사한 감정 표현 알고리즘을 사용하였다.

이와 같은 감정표현 시스템은 먼저 입력으로부터 각 감정의 가중치를 입력 받게 되며 이 값들을 이용하여 감정 공간을 구성하게 되는데 neutral 중심점으로부터 축적된 경험에 의해 그 길이와 사이각을 결정하게 된다. 이렇게 좌표축의 모양이 결정되면 입력된 5가지 감정의 가중치를 이용하여 해당하는 감정의 좌표축에 나타내고 그림 8과 같이 사각형 모양으로 형성한다.

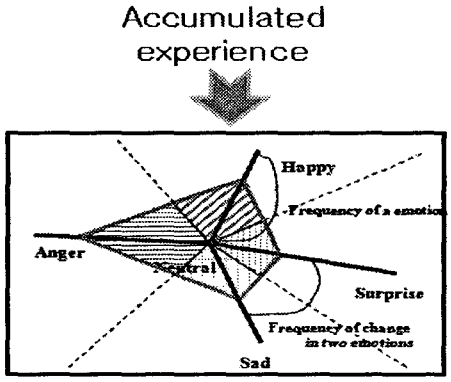

그림 8. 동적 감정 공간 모델

각각의 감정을 나타내는 두 직선사이를 이등분하는 직선을 구분하고, 각 감정에 속하는 영역의 크기로서 감정을 분류하게 된다. 본 논문에서는 이와 같은 영역의 크기를 다음과 같은 헤론의 공식을 통해 구했다.

$$s = \frac{a+b+c}{2}$$

$$S = \sqrt{s(s-a)(s-b)(s-c)} \qquad (1)$$

식에서 a, b, c는 삼각형 세 변의 길이, S는

삼각형의 넓이를 각각 나타낸다.

이렇게 구해진 영역의 크기를 이용하여 얼굴 표정을 표현하기 위해서는 얼굴의 각 특징 요소에 대한 파라미터를 설정하여 감정 영역의 크기를 가중치로 하여 조절함으로써 자연스러운 표정 변화 시스템을 구현할 수 있었다. 이에 관해서는 다음 절에서 자세히 설명한다. 마지막으로 이렇게 구해진 파라미터 값들을 가지고 얼굴 특성 요소인 눈, 눈썹, 입, 턱 등의 파라미터를 조정하여 최종적인 얼굴 표정을 표현한다.

5.2 감정표현을 위한 파라미터 설정

감정을 표현하기 위한 파라미터로 눈썹 3개 (p_0, p_1, p_2), 눈 3개(p_0, p_1, h), 입 3개(w, h, arc) 총 9개의 파라미터를 정의하였다. 여기서 p_n는 수평점, h는 수직 높이, w는 수평 넓이, arc는 휜 방향 및 정도를 각각 나타낸다.

이렇게 9개의 파라미터에 대해 각 감정이 나타내는 최대치에 영역의 크기 비율을 곱하여 감정을 표현하도록 하였으며, 각 감정이 속하는 영역의 크기를 다음과 같이 표현하면,

ar_{Happy}, ar_{Anger}, ar_{Sad}, $ar_{Surprise}$

$ar_{Total} = ar_{Happy} + ar_{Anger} + ar_{Sad} + ar_{Surprise}$

각 감정이 속하는 영역의 비율은 식 (2)와 같이 된다.

$$w_{Happy} = \frac{ar_{happy}}{ar_{Total}}, \quad w_{Anger} = \frac{ar_{Anger}}{ar_{Tatal}}$$

$$w_{Sad} = \frac{ar_{Sad}}{ar_{Tatal}}, \quad w_{Surprise} = \frac{ar_{surprise}}{ar_{Total}} \quad (2)$$

따라서 감정 표현 파라미터는

$P_i = A_i w_{Happy} + B_i w_{Anger} + C_i w_{Sad} + D_i w_{Surprise}$

가 된다. 여기서 A_i는 행복, B_i는 화남, C_i는 슬픔, D_i는 놀람만 나타날 때의 초기 설정 값이고 실험적으로 결정된다. 그리고 P_i는 각각의 파라미터, i는 각각의 파라미터 번호를 나타내므로, $i = 1,2,\cdots,9$가 되어 감성 표현 파라미터를 조절함으로써 얼굴 표정이 변화하게 된다[그림 9].

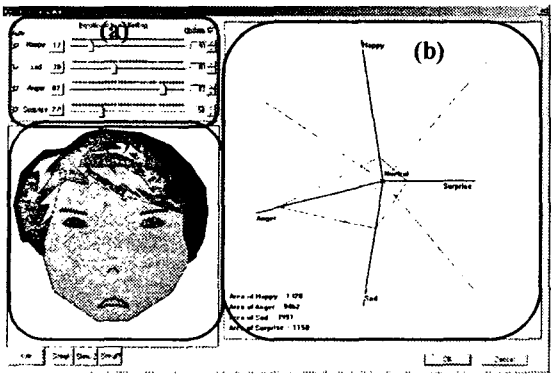

그림 9. 감정 표현 시스템

6. 결 론

본 논문에서는 4가지 감정에 대해서 음성신호와 얼굴 영상을 통해 감정을 인식하였으며, 인식한 결과 음성(평균 인식율 74%)이 얼굴 영상(평균 인식율 52%)보다 인식율이 우수함을 알 수 있었으며 음성신호와 얼굴 영상을 융합하여 감정인식을 할 수 있는 Multi-Modal 방법을 제안하였다.

또한 감정 인식된 결과를 이용하여 동적 감정 공간에서 감정을 표현하는 시스템을 구현하였다.

참 고 문 헌

[1] 박창현, 심귀보, "음성 신호를 이용한 감정 인식에서의 패턴 인식 방법," Journal of Control Automation and Systems Engineering , Vol. 12, No. 3, 2006.

[2] H. A Rowley, S. Baluja, T. Kanade, "Rotational Invariant Neural Network Based Face Detection," *Proc. of IEEE Conference on Computer Vision Pattern Recognition*, pp. 38-44, 1998.

[3] C. Padgett, G. Cottrell, "Representing face images for emotion classification," *Advances in Neural Information Processing Systems*. Vol. 9, MIT Press. 1997.

[4] 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.

[5] B. Menser, F. Muller, "Face detection in color Image using principal component analysis," Image Processing and Its Applications, *1999 Seventh International Conference on (Conf. Publ. No. 465)*, Vol. 2, pp. 620-624, 13-15 July 1999.

[6] C. C. Chibelushi, "Feature-Level Data Fusion for Bimodal Person Recognition", Image Processing and Its Application, Vol. 1, pp. 399-403, 1997

[7] Liyanage C. DE SILVA, Tsutomu MIYASATO, "Facial Emotion Recognition Using Multi-modal Information," International Conference on Information Communication and Signal Processing, pp. 397-401, 9-12 September 1997.

[8] 심귀보, 변광섭, 박창현, "동적 감성 공간에 기반한 감정 표현 시스템," *한국퍼지 및 지능시스템학회 논문지*, 제15권, 제1호, pp. 18-23, 2005.

[9] 심귀보, 박창현, "음성으로부터 감성인식 요소 분석," *한국퍼지 및 지능시스템학회 논문지*, pp. 199-201, 2001.