영상과 음성의 출력 데이터를 이용한 감정인식

오 재 흥* · 정 근호* · 주 영 훈* · 박 창현** · 심 귀 보**
*군산대학교 전자정보공학부, **중앙대학교 전자전기공학부

Emotion Recognition Using Output Data of Image and Speech

Jae-Heung Oh* · Keun-Ho Jeong* · Young-Hoon Joo* · Chang-Hyun Park** · Kwee-Bo Sim**
*School of Electronic & Information Eng, Kunsan National University,
**School of Electrical & Electronic Eng, Chung-Ang University,

Abstract - 본 논문에서는 영상과 음성의 데이터를 이용한 사람의 감정을 인식하는 방법을 제안한다. 제안된 방법은 영상과 음성의 인식률에 기반 한다. 영상이나 음성 중 하나의 출력 데이터만을 이용한 경우에는 잘못된 인식에 따른 결과를 해결하기가 힘들다. 이를 보완하기 위해서 영상과 음성의 출력을 이용하여 인식률이 높은 감정 상태에 가중치를 줌으로써 잘못된 인식의 결과를 줄일 수 있는 방법을 제안한다. 이를 위해서는 각각의 감정 상태에 대한 영상과 음성의 인식률이 추출되어져 있어야 하며, 추출된 인식률을 기반으로 가중치를 계산하는 방법을 제시한다.

1. 서 론

최근 컴퓨터의 급속한 발전과 함께, 사용자의 의지나 감정을 인지하는 인간 친화 적인 컴퓨터 인터페이스를 구축하고자 하는 연구들이 활발히 진행되고 있다[1]. 일례로, 최근에 개발, 발표되고 있는 로봇들은 두 발로 걷고, 춤을 추는 등 인간의 행동과 매우 유사한 동작을 할 수 있다. 그러나, 단순한 움직임만을 구현하는 것은 인간의 보조자로써 한계가 있다. 기계의 업무 영역을 넓히기 위해서는 좀더 인간과 유사해져야 한다. 이 때 필요한 것이 감정의 인식이다. 인간의 감정을 인식하게 됨으로써 좀더 유연한 방법으로 인간에게 도움을 줄 수 있을 것이다.

예를 들면, 어떤 사람이 로봇에게 음악을 틀라고 명령을 내렸을 때, 로봇이 사람의 감정을 인식한다면, 적절한 음악의 목록을 제시하고 명령을 수행할 수 있을 것이다. 그리고, 로봇 이외에도 게임 등의 소프트웨어에도 적용하여 더욱 재미있는 제품들이 만들어 질 수 있을 것이다. 감정 인식은 얼굴 표정 인식과 음성으로부터의 인식, 두 가지 접근 방법이 있다. 본 논문에서는 두 방법을 같이 사용하여서 감정의 인식을 연구하였다. 기존의 연구에서는 음성과 영상을 따로 분리하여서 사람의 감정을 인식하였다. 이렇게 사람의 감정을 인식하는데 있어서 영상과 음성을 따로 분리하여 연구할 경우, 출력된 결과에 있어서 잘못된 인식이 수행되었을 경우, 이를 보상해줄 방법이 없다. 그렇기 때문에 본 논문에서는 영상과 음성의 출력을 이용한 통합된 인식에 대한 연구를 실행하였다.

2. 영상과 음성의 감정 정보 추출

2.1 영상의 감정 정보 추출

2.1.1 얼굴 영역 추출

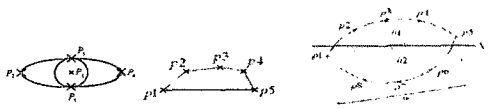
본 논문에서는 획득한 정면 얼굴 영상으로부터 특징 점을 추출하는데 사용되는 메모리 인식 시간을 줄이기 위해 320×240의 해상도로 축소하여 실험할 수 있도록 시스템을 구성하였다. 초기에 획득한 영상으로부터 영상 크기에 따른 거리별 오차를 보상하기 위해 각 특징의 크기를 일정한 크기 범위로 제한시키는 일반화 방법을 사용한다.

초기에 획득한 칼라 영상으로부터 얼굴 영역의 위치를 찾기 위해 컬러 공간의 색조 차이를 이용한다. CCD 카메라로부터 입력된 RGB 영상을 피부색 영역 추출에 널리 사용되는 YIQ 모델을 사용하여 피부색 영역을 추출한다. YIQ 모델을 사용하여 피부색 영역과 그 외 영역을 나눈 후, 피부색 영역의 중심점으로부터 얼굴 영역을 산출하게 된다.

2.1.2 형판 벡터 추출

입력 영상에서 얼굴 영역은 일정한 크기로 입력되어지는 것이 아니다. 정확한 형판을 추출하기 위해서는 추출된 얼굴 영역을 일정한 크기로 정규화 하지 않으면 안 된다. 어떠한 이미지의 크기가 변화할 때는 여러 가지 보간법을 사용하여 불필요한 데이터의 추가나 손실을 최소화해야 한다. 본 논문에서는 구현이 용이하고, 성능 또한 뛰어난 양선형 보간법을 사용한다. 이렇게 정량화 된 얼굴 영역 영상에서 눈, 눈썹, 입의 형판을 추출하기 위해서는 얼굴 영역을 얻을 때 사용된 방법인 YIQ-칼라 공간 하나만을 사용하여 각각의 형판을 정확히 분리하여 추출하기가 곤란하다[2]. 이를 보완하기 위해서 두 번째 과정에서는 HSV, YCbCr 칼라 공간을 추가적으로 적용시켜 데이터 손실을 최소화시킬 수 있게 된다.

인간은 감정에 따라서 눈썹, 눈, 입의 크기나 모양이 변화하게 된다[3]. 이러한 변화에 따라서 각각의 상태 정보를 가지고 있는 형판 벡터를 추출한다. 이를 위해서는 각각의 형판에 대해 같은 방법으로 형판 벡터를 추출하는 것보다는 각각의 형판에 대한 특징과 감정 인식에 있어서 얼마나 중요한지를 따져서 형판별로 다른 방법을 적용해 형판 벡터를 추출하는 것이 더 효율적이다. 그림 1은 각각의 형판에 대한 특징 점들을 표시하였으며, 그림 2에서는 그림 1을 사용해서 실제적으로 실험을 통해 얻어진 형판 벡터들을 나타낸다.



(a) (b) (c)

그림 1. (a)눈의 특징 점, (b)눈썹의 특징 점, (c)입의 특징 점

그림 2. 형판 벡터 추출 예

2.1.3 인식을 위한 학습 알고리즘

본 실험에서 최종적으로 얻고자 하는 것은 감정인식이다. 이를 위해서는 얻어진 형판벡터들에 대한 학습과정이 필요하게 되는데, 본 논문에서 사용한 학습알고리즘은 역전파 알고리즘이다. 이 역전파 알고리즘의 입력으로는 각각 추출된 형판 벡터들이 입력으로 들어가며, 결과적으로 우리가 얻고자 하는 출력 값인 감정상태를 얻게된다. 획득한 형판벡터를 이용하여 인간의 감정(무표정, 놀람, 화남, 기쁨)들에 대해 각각 50개씩 전체 200개의 학습이미지들에 대한 데이터를 산출한다. 이를 신경회로망을 이용하여 각각의 감정상태에 따라서 학습을 시킨다. 역전파 알고리즘은 학습이 수행되는 동안 가중치 정보가 향상되고 원하는 반복회수에 도달하면 학습을 멈추고 최종가중치 정보가 저장된다. 이 정보를 이용하여 감정인식을 수행하게 된다. 마지막으로 형판벡터의 추출과정에서부터 인식까지 일련의 과정을 그림 3에 나타내었다.

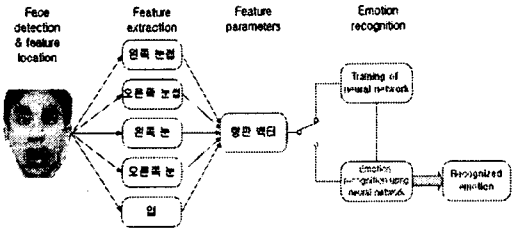


그림 3. 형판 벡터 추출에서 학습까지의 단계

2.2 음성의 감정 정보 추출

2.2.1 음향의 감성정보

소리의 정보 전달에는 자료의 전달뿐 아니라 정서적인 정보 또한 포함하고 있다. 동일한 정보를 전달하더라도 부드러운 목소리로 전달하는 경우와 불쾌한 목소리로 전달하는 경우의 정보전달 효과는 다르다. 소리 신호는 푸리에의 이론에 여러 주파수들의 조합으로 이루어져 있으므로 갖가지 감성을 갖는 소리들을 주파수 영역에서 분석하여 분류할 수 있다. 목소리의 기본음(Fundamental Frequency)은 125Hz~250Hz 이고, 목소리의 힘을 주는 음역대는 350Hz에서 2000Hz이다. 즉, 에너지가 가장 많은 부분이다. 표 1에서는 주파수 대역에 따라 다른 청각적 느낌을 나타낸 표이다.

표 1. 주파수 대역에 따른 청각적 느낌

주파수	청각적 느낌
125Hz-500Hz	증폭하면 묵직함이 생기고, 줄이면 평장히 약한 목소리가 됨
2000Hz-5000Hz	증폭하면 목소리의 명확도가 높아짐
4000Hz-8000Hz	지찰음, S, SH, CH, C등을 말할 때의 듣기 싫은 소리가 나옴. 줄이게 되면 목소리의 깨끗함도 감소됨.
8000Hz 이상	입에서 나오는 공기소리

2.2.2 음향 요소 분석

본 논문은 음향 요소를 날카로움, 저음, 굵음, 가늠, 큼, 작음의 6가지로 정의한다. 6가지 요소의 분석을 위하여 모음 '아'에 대한 실험을 하였다. 표 2에서 F : Formant, Mag : Magnitude, Int : Intensity, NU : Non Uniform, M1 : Man 1을 의미한다.

표 2. M1에 대한 음향적 분석

	M1_평	M1_날카로움	M1_저음	M1_중저음
1F(Hz)	824	800	600	650
2F(Hz)	1100	1200		
3F(Hz)	3000	NU	2800	2780
4F(Hz)	3480		3300	3400
5F(Hz)	4570		NU	4300
Mag	0.8	1.65	0.6	1.4
Int(dB)	78dB	88dB	75dB	80dB
Pitch(Hz)	134Hz	370Hz	109Hz	130Hz

Table 3. M2에 대한 음향적 분석

	M1_평	M1_날카로움	M1_저음	M1_중저음
1F(Hz)	812	700	630	743
2F(Hz)	1210	1200	1100	1160
3F(Hz)	2760	2700	2500	2600
4F(Hz)	3700	NU	3400	3400
5F(Hz)	NU		없음	3600
Mag	1.6	1.6	0.6	1.4
Int(dB)	83dB	87dB	75dB	80dB
Pitch(Hz)	117Hz	290Hz	109Hz	130Hz

표 2와 3은 각각 여러 실험 데이터 중 대표적인 결과이다. 위 결과에 따르면 음의 높낮이는 피치(Pitch)로 분명한 구분이 가능하다. 그리고, 날카로운 소리의 경우는 3, 4, 5 Formant가 순간순간 변하는 것을 알 수 있다. 날카로운 소리와 저음의 경우를 비교해보면, 배에서부터 나온 소리의 경우 Uniform한 Formant의 분포를 보이나, 날카로운 소리중 머리로부터 울리는 경우에는 3, 4, 5 Formant에서 Non Uniform한 특성을 보인다. 이 경우에는 소리가 맑지 않다. 그리고, 125 - 500Hz는 풍부한 느낌의 주파수가 분포한다. 또한, 1F가 500Hz 주변에서 많이 분포하는 경우에 더 굵은 목소리로 들린다. Table 2.에서 저음과 중저음 부분의 1F, 2F가 같은 대역에 분포하는 경우는 각각 다른 대역에 분포하는 경우보다 에너지의 크기가 더 크므로 더욱 굵은 소리를 낸다. 반대로 500Hz에서 멀수록 가는 소리이다.

2.2.3 DRNN을 이용한 감성인식을 위한 특징 추출

감성인식 시뮬레이터의 입력 값으로는 피치의 패턴들을 사용한다. 즉, 4가지 감정에 대한 대표 패턴들을 DRNN(Dynamic Recurrent Neural Network) 구조를 이용하여 학습, 인식한다. 피치를 추출하기 위해서 우선 Autocorrelation Approach using Center-Clipping Function을 사용한다.

$$A(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \quad (1)$$

식 (1)은 Autocorrelation 함수를 나타내고, Center Clipping function은 Autocorrelation 함수에 데이터를 입력시키기 전에 필요 없는 정보를 제거하는 역할을 수행한다.

$$y(n) = c[x[n]] \quad (2)$$

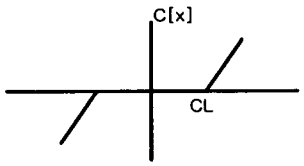


Fig 4. Center clipping function

식 (2)와 그림 4.가 Center clipping function을 나타낸다. 이 함수는 음성 신호가 일정한 레벨(C_L)내에 있으면 그 신호를 무시하고, C_L보다 크면 원래 신호에서 C_L을 뺀다. 이는 음성신호 중에서 피치에 해당하는 성분은 크기가 크게 나타나는 특징을 이용해서 잔여성분을 제거하는 방법이다. 그리고, C_L은 프레임 내의 가장 큰 음성신호 레벨의 64%를 기준으로 한다[4].

2.2.4 시뮬레이터

그림 5는 시뮬레이터의 구조를 나타낸다. 마이크를 통하여 음성이 입력되면, 음성의 피치를 추출하고 추출된 피치를 각 감정에 대응하여 학습을 시킨다. 학습을 통하여 신경망의 가중치를 획득 인식기를 통해 인식을 한다[5].

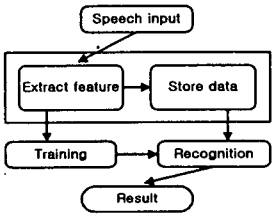


그림 5. 시뮬레이터의 구조

2.2.3 DRNN

- Fully connected.
- Input : 1, Hidden : 2, Out : 4
- Input : Pitch.
- Out : Normal, Angry, laugh, Surprise

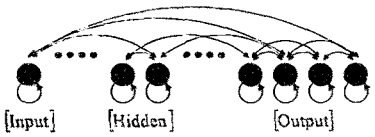


Fig 6. DRNN

DRNN의 구조는 그림 6과 같고, 음성의 입력이 시간에 따라 순차적으로 들어오므로 DRNN이 이러한 종류의 데이터에 적합하다. i번째 뉴런의 출력은 다음과 같다.

$$y_i = f(h_i(t-1)) + A(\sigma) \quad (3)$$

$$h_t = (\sum_j w_{ij} y_j(t) + x_i(t)) \quad (4)$$

$h_i(t-1)$는 시간 (t-1)에서 i번째 노드에 대한 입력이다. $x_i(t)$는 시간 t에서의 외부 입력이다. $f(\cdot)$는 nonlinear derivative activation function 이다[6].

$$f(x) = \frac{2}{1+\exp\left(\frac{-2x}{u_0}\right)} - 1 = \tanh\left(\frac{x}{u_0}\right) \quad (5)$$

3. 영상과 음성의 출력을 이용한 감정인식

영상과 음성 각각의 출력이 서로 상이한 경우에 우리는 어느 하나의 감정을 선택하여야 한다. 이를 위해서는 우선 영상과 음성에 대해서 각각의 감정 상태에 대한 인식률이 계산되어져 있어야 한다. 계산된 인식률을 사용하여서 각각의 감정 상태에 대한 가중치 값을 할당하게 된다.

각각의 감정 상태와 음성과 영상의 가중치는 아래와 같이 표현된다. 하나의 감정 상태에 대해서 가중치는 영상과 음성 2가지의 경우가 나오게 되는데, 이 두개 중 하나는 0에 가까운 값을 갖게 되고, 다른 하나는 그 감정의 인식률의 차의 곱을 취함으로써 가중치를 할당하게 된다.

$$\begin{aligned} Image\ Weight &: W_{1neutral}, W_{1happiness}, W_{1surprise}, W_{1anger} \\ Speech\ Weight &: W_{2neutral}, W_{2happiness}, W_{2surprise}, W_{2anger} \end{aligned}$$

이렇게 구해진 가중치 정보를 가지고 실제 시스템의 출력을 결정하게 되는데 여기에 사용된 방법은 식(6)과 같이 표현이 된다. 여기서 I는 영상의 감정 출력이고, S는 음성의 감정 출력이다. 이를 각각의 감정 상태에 대해서 가중치와 곱을 취하면, 각각의 감정 상태에 대한 출력이 나오게 되는데, 결과적으로는 값이 제일 큰 감정 상태가 출력으로 나오게 된다. 이는 식 (7)에 표현하였다.

$$\begin{aligned} O_{neutral} &= W_{1neutral} I_{neutral} + W_{2neutral} S_{neutral} \\ O_{happiness} &= W_{1happiness} I_{happiness} + W_{2happiness} S_{happiness} \\ O_{surprise} &= W_{1surprise} I_{surprise} + W_{2surprise} S_{surprise} \\ O_{anger} &= W_{1anger} I_{anger} + W_{2anger} S_{anger} \end{aligned} \quad (6)$$

$$System\ Output = Max\{O_{neutral}, O_{happiness}, O_{surprise}, O_{anger}\} \quad (7)$$

영상이나 음성의 입력으로부터 감정의 출력까지를 표시한 그림은 아래의 그림 7과 같다.

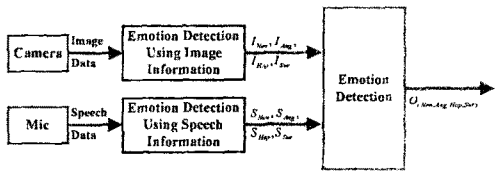


Fig 7. 감정 인식의 전체 시스템

4. 실험 결과 및 고찰

제안한 방법의 응용 가능성을 확인하기 위해서 실험을 통해 얻어진 결과들을 비교 분석해 보았다. 우선 음성과 영상의 인식률을 제안한 방법에 의해서 각각 추출한 후, 이를 바탕으로 음성과 영상에 대해서 각각의 감정 상태에 대한 가중치를 할당한다. 가중치를 할당한 후, 식(6)을 이용해서 영상과 음성의 출력 데이터를 이용해서 영상과 음성의 출력 데이터를 이용한 감정 인식을 행하게 된다. 영상과 음성의 오인식의 결과에 대한 몇 가지 실험 값들을 이용하여서 제안한 방법을 실험한 결과 약 60%정도의 오인식에 따른 결과를 해결할 수가 있었다. 평균적으로 영상과 음성의 감정 인식률은 75%정도인데, 제안한 방법을 추가하여 실험한 경우에는 약 80% 정도 이상의 결과를 산출할 수가 있었다. 이는 제안한 방법 자체가 영상이나 음성의 출력 중 어느 하나에 가중치를 줌으로써 각각의 감정 상태에 대한 오인식의 결과를 보완하기 위해서 제안되었기 때문이다.

5. 결 론

본 논문에서는 영상과 음성의 출력 데이터를 이용한 감정 인식 방법을 제안하였다. 기존의 연구에서는 감정 인식을 행하는데 있어서, 영상과 음성을 따로따로 분리하여 실험을 행하였다. 제안된 방법은 영상의 감정 상태 출력과 음성의 감정 상태 출력이 서로 상이 할 때, 어느 한쪽의 감정에 가중치를 줌으로써 인간의 감정 상태를 식별하였다. 여기에 사용된 가중치는 음성과 영상의 각각의 감정 상태에 대한 인식률을 기반으로 계산되어지게 된다. 영상에서의 감정 상태 출력은 형판 벡터를 이용하여 추출된 얼굴의 형판 벡터들을 역전파 알고리즘을 통해서 학습과 인식을 수행하였으며, 음성은 피치 정보를 이용하여 DRNN(Dynamic Recurrent Neural Network)으로 학습시켰다. 이러한 학습을 통해 출력된 결과를 이용하여 각각의 감정 상태에 대한 인식률을 계산하며, 이렇게 계산된 인식률을 바탕으로 영상과 음성의 출력 데이터가 상이할 때 어느 한쪽의 결과에 가중치를 줌으로써 감정상태를 출력하게 된다.

감사의 글: 본 연구는 산업자원부 차세대 신기술 개발사업(과제번호: N09-A08-4301-09)에 의해 지원되었습니다

[참 고 문 헌]

[1] Samal, A., Iyengar, P. A., "Automatic recognition and analysis of human faces and facial expressions : A survey", Pattern Recognition, Vol. 25, No. 1, pp. 65-77, 1992.

[2] Christophe Garcia and Georgios Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis", IEEE TRANSACTIONS ON MULTIMEDIA, Vol. 1, No. 3, pp. 264-277, 1999.

[3] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 2, pp. 97 - 115, 2001.

[4] J. S. Han., "Speech Signal Processing", Seoul, O-Sung-media, pp. 90, 2000.

[5] C. H. Park and K. S. Heo and D. W. Lee and Y. H. Joo, K. B. Sim., "Emotion Recognition based on Frequency Analysis of Speech Signal", Proc. Of the FIRA Robot 2002 Conference, Seoul, Korea, May. 27-29, 2002.

[6] K. B. Sim, "Methodology of Artificial Life", Seoul, Dream-Media, 2000.