신체움직임에 대한 컴퓨터비전 기반 실시간 감정인식

박한훈, 박종일, 우운택*
한양대학교 전자전기컴퓨터공학부, *광주과학기술원 정보통신공학과

Emotion Recognition from Body Movement

HanHoon Park, Jong-Il Park, Woontack Woo*
Hanyang Univ., *K-JIST
{hanuni, jipark}@mr.hanyang.ac.kr, wwoo@kjist.ac.kr

Abstract

컴퓨터와 주변장치의 급속한 발전은 인간과 컴퓨터 사이의 인터페이스에 많은 변화를 가져왔다. 특히, 인간의 감정을 인식하는 기술은 컴퓨터를 보다 인간 친화적으로 만들기 위한 노력으로, 그 동안 꾸준히 진행되어 왔다.

본 논문에서는 한대의 카메라로 촬영한 영상으로부터 실시간으로 신체움직임이 표현하는 감정을 인식할 수 있는 방법을 제안하고, PC기반의 실시간 감정 인식 시스템을 구현한다.

1. 서론

신체움직임은 인간의 의사를 전달할 뿐 아니라 감정을 표현하는 중요한 수단이다. 특히, 댄스는 신체 움직임의 이러한 기능을 특성화한 것이므로, 댄스로부터 감정을 인식하려는 노력은 많은 연구의 초점이 되어 왔다. 하지만, 기존에 제안된 방식들은 번거롭게 복잡한 부착물을 필요로 하는 접촉식이거나, 컴퓨터 비전을 이용하는 비접촉식이라 하더라도 많은 계산을 필요로 하기 때문에 우리가 궁극적으로 추구하는 실시간 감정 인식에는 적합하지 않았다.

신체 움직임으로부터 감정을 실시간에 인식하기 위하여 본 논문에서는 2차원 영상에 간단한 처리를 수행하여 정량적인 정보를 추출한 후, 신경망을 이용해서 이를 감정으로 직접 맵핑하는 기술을 제안한다. 선행 연구에서 Woo 등은 영상으로부터 추출된 다양한 후보 특성량들 중에서 보다 정확하게 감정 정보를 대변할 수 있는 특성량을 무용 이론에 기반하여 선정한 다음, 각 특성량을 지연을 두어 신경망의 입력으로 이용해서, 특성량의 순간값 뿐아니라 특성량의 변화 모습까지 추적함으로써 상당히 높은 인식률을 달성하고 있다[1]. 그런데 이 방법에서는 영상에서 추출할 수 있는 다양한 특성량 중에서 일부만을 경험적으로 사용하고 있고, 시스템의 성능이 선정된 특성량에 따라 자칫 저하될 수 있어 보편성, 신뢰성 있는 시스템을 구축하기 위한 개선이 요구되고 있다. 따라서 본 논문에서는 가능한 모든 의미있는 특성량을 이용하여 인식률을 높이는 방법을 연구한다. 즉, 영상에서 고속으로 추출 가능한 저수준 특성량을 모두 신경망의 입력으로 사용한다. 단, 계산의 복잡도를 감안하여 시간 지연을 이용하지 않는다.

제안된 기술을 이용해서 4가지 감정 정보(기쁨-happiness or cheerfulness, 슬픔-sadness or lonelihood, 분노-angry or disgust, 놀람-surprise)를 인식하게 되는데, 정확한 판단을 할 수 없을 때는 불명확으로 판단한다.

본 논문은 다음과 같이 구성된다. 2장에서는 실시간 감정 인식 시스템에 대한 개략적인 설명을 한다. 3장에서 실험에 사용될 특성량을 소개하고, 4장에서는 특성량을 선정하는 방법에 대해 설명한다. 5장에서는 실험 환경 및 결과를 분석한다. 6장에서는 결론을 제시한다.

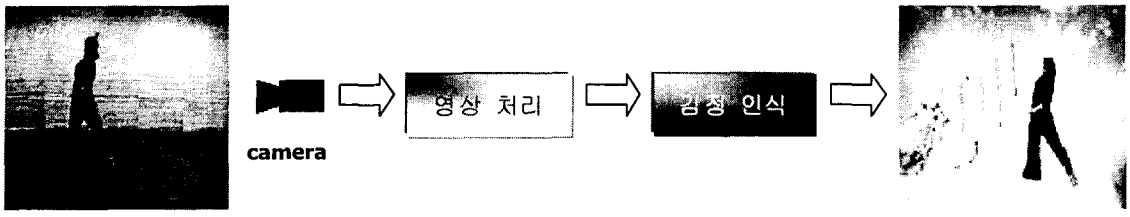

그림 1. 실시간 감정인식 시스템 개요.

2. 실시간 감정인식 시스템 개요

전체 시스템의 개요는 그림 1과 같다.

우선, 카메라를 통해 들어온 영상 시퀀스에 색차추출법(Difference-Keying)과 그림자 제거 처리[2]를 수행하여 배경을 제거한 이진 영상을 뽑아낸다. 이어, 추출된 이진 영상의 정보를 사각 박스의 크기나 무게 중심 좌표로 단순화 시키고[3], 이로부터 감정 인식에 적절한 특성량을 추출한다. 마지막으로, 추출된 특성량들을 MLP-신경망[4]을 이용해서 미리 정의된 감정으로 맵핑한다.

이렇게 감정이 인식되면 그에 알맞은 배경 영상, 특수효과 등을 구사하여 다양한 인터랙티브 시스템을 구축할 수 있다. MIDAS[3] 같은 시스템은 이를 실제로 구현한 좋은 예이다.

3. 특성량 추출 방법

신체 각 부위에 대한 복잡한 3차원 정보를 실시간에 안정되게 추출하는 것은 현재의 기술 수준으로는 매우 어렵기 때문에, 본 논문에서는 일반 PC에서도 실시간 처리가 가능한 시스템 구축에 중점을 두고, 2차원 영상에서 보여지는 3차원 동작을 사각박스나 무게 중심 등의 움직임으로 단순화 시켰다. 그림 2에서 보는 것처럼, 사각 박스의 크기나 무게 중심의 위치 좌표 등의 특성량을 이용해서 신체 움직임을 나타내고 있다.

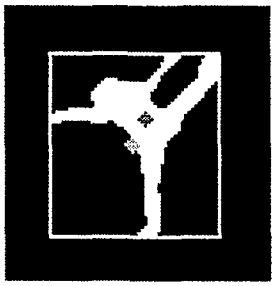

그림 2. 신체 움직임의 단순화.

신체 움직임을 표현하는 특성량으로서 사각 박스의 크기, 무게 중심의 좌표, 실루엣의 영역, 사각 박스와 실루엣 영역 사이의 비와 각 특성량의 속도, 가속도 등을 이용했다(표1). 여기서, 무게 중심이나 사각 박스의 좌표는 일반 공간(general space)을 추적하고, 나머지는 개인 공간(personal space)을 추적한다[5].

표 1. 실험에 사용된 특성량

사각 박스의 가로 길이와 세로 길이의 비	W
무게 중심의 좌표	(Cx, Cy)
사각 박스의 중심 좌표	(Rx, Ry)
실루엣 영역의 크기	Ss
사각 박스 영역의 크기	Sr
각 특성량의 속도	f(·)
각 특성량의 가속도	g(·)

4. 특성량 선정 방법

모든 특성량을 그대로 이용하는 방법과 특정 특성량들의 선형 결합을 이용하는 방법으로 나누어서 수행한다.

4-1. 모든 특성량을 이용하는 방법

모든 특성량을 이용하는 방법은 각 특성량의 크기, 속도, 가속도를 이용하여 신경망을 학습시킴으로써, 모든 분류 작업은 신경망이 전담하게 된다. 총 21개의 특성량을 이용하며 지연은 이용하지 않는다.

4-2. 선형 결합 방법(Heuristic 방법)

라반의 이론에 근거해서[8], 신체의 움직임을 공간, 속도, 가속도량을 가지고 정량적으로 분석한다. 추출된 특성량들의 선형 결합에 의해 공간, 속도, 가속도량을 대변한다.

P1 = a1*W + a2*Ss + a3*Sr

P2 = a4*f(W) + a5*f(Cx) + a6*f(Cy) + a6*f(Sr)

P3 = a7*g(W) + a8*g(Cx) + a9*g(Cy) + a10*g(Sr) + a11*g(Ss)

P1, P2, P3의 값이 신경망의 입력으로 이용된다. 지연은 이용하지 않는다.

5. 실험 및 결과

각 감정을 잘 표현할 수 있는 시퀀스를 얻기 위해 4명의 무용 전문가들의 동작을 캡쳐했다. 각 무용 전문가의 시퀀스는 신경망을 학습시키고, 또, 다른 사람의 시퀀스로 훈련된 신경망을 테스트하는 데 사용한다. 이는 같은 감정을 표현하는데 있어서도 사람마다 다를 수 있음을 감안한 것이다.

추출된 특성량은 그대로 이용되지 않고, 일정 간격 평균을 거친 후, 이용하게 된다. 이는 매 프레임에서 추출된 특성량을 그대로 이용하게 되면, 잡음에 매우 민감하게 되어 인식률이 나빠질 수 있기 때문이다. 또한 너무 긴 간격동안 평균을 하게 되면, 신경망의 학습율이 떨어지고 인식률 역시 나빠진다. 그림 4는 이러한 특성을 나타낸다.

본 논문에서는 10 샘플씩 평균을 취하여 각 특성량 별로 초당 3개의 평균값을 인식에 사용하였다.

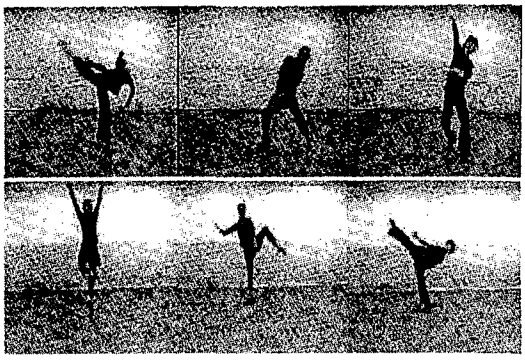

그림 3. 실험에 사용된 영상.

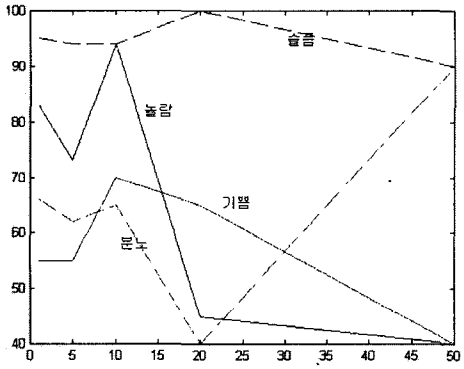

그림 4. 샘플 수에 따른 인식률.

표 2는 특성량을 선정하는 방법에 따른 결과를 보여준다. 특정 특성량을 이용할 경우, 시간 지연을 이용하지 않게 되면 인식률이 상당히 나빠진다. 이를 극복하기 위해서 특성량의 개수를 늘임으로써 인식률이 증가했음을 확인할 수 있다.

표 2. 특성량 선정 방법에 따른 인식률

[단위: %]

	기쁨	놀람	분노	슬픔
Full	70	94	65	94
Heuristic	70	45	50	94

실시간 감정인식 시스템을 구현하기 위하여 표 3과 같은 실험환경을 사용하였다. 시스템의 처리 속도를 높이기 위해, 모든 처리를 시스템에 최적화된 라이브러리를 이용하여 구현하였으며, 영상의 크기는 160*120 화소로 처리하였다. 구현된 시스템은 실시간으로 동작하면서 1/3

초에 한번씩 감정인식 결과를 발생시킨다.

표 3. 실험 환경

하드웨어	CPU: Intel Pentium-III 500 HDD: SCSI Ultra2 Wide (68pin) RAM: 256MB Cannon MV1 디지털 비디오 카메라 Canopus DV Raptor
소프트웨어	Intel Image Processing Library Intel Computer Vision Library Microsoft Vision SDK OpenGL Library Visual C++ 6.0 Adobe Premiere 6.0
영상	Uncompressed 160*120

6. 결론

본 논문에서는 한대의 카메라로 촬영한 영상으로부터 감정을 실시간에 자동 인식하는 시스템을 구축하고 성능을 확인하였다. 실험을 통해 실시간에 추출 가능한 저수준 특성량의 개수를 늘이면 인식률을 향상시킬 수 있음을 확인했다.

선행 연구에서 수행한 대로[1], 시간 지연을 고려해서 특성량의 변화 모습까지 추적한다면 인식률을 보다 개선할 수 있을 것으로 기대된다. 그러나 모든 특성량에 대해 시간 지연을 고려하면 계산복잡도의 증가로 인해 시스템의 실현이 어려워진다. 앞으로는 특성량의 시간 지연을 고려하되, 실현 가능한 특성량의 종류와 개수를 최적화하는 작업이 중요한 과제가 될 것이다.

감사의 글

실험에 협조해 주신 한양대학교 무용학과의 김정윤, 김현화, 이지은, 한류리 씨에게 감사드립니다.

참고 문헌

[1] W. Woo, J. Park, Y. Iwadate, "Emotion analysis from dance performance using time-delay neural networks", Proc. of CVPRIP, Vol. 2, pp. 374-377, 2000.

[2] N. Kim, W. Woo, M. Tadenuma, " Photo-realistic Interactive Virtual Environment Generation using multiview cameras" , in Proc. SPIE PW-EI-VCIP'01, vol. 4310, Jan. 2001.

[3] R. Suzuki, Y. Iwadate, M. Inoue, W. Woo, " MIDAS: MIC Interactive Dance System" , IEEE Int' l conf. on Systems, Man and Cybernetics, Vol. 2, pp. 751-756, 2000.

[4] P. Modler, F. Hofmann, I. Zannos, " Gesture Recognition by Neural Networks and the Expression of Emotions" , IEEE Int' l conf. on Systems, Man and Cybernetics, Vol. 2, pp 1072-1075, 1998.

[5] A. Camurri, M. Ricchetti, and R. Trocca, " Eyeweb-toward gesture and affect recognition in dance/music interactive system" , in Proc. IEEE Multimedia Systems, June 1999.

[6] F. Kawakami, M. Okura, H. Yamada, H. Harashima, S. Morishima, " An Evaluation of 3-D Emotion Space" , Proc. of 4th IEEE Int' l workshop on Robot and Human Communication, RO-MAN' 95 TOKYO, pp. 269-274, 1995.

[7] A. Camuri, S. Hashimoto, K. Suzuki and R. Trocca, " Kansei Anaysis of Dance Performance" , Proc. IEEE SMC' 99, IV, pp. 327-332, 1999.

[8] R. Raban, Modern Educational Dance, Trans-Atlantic Publications, Inc., 1988.