

TDMLP 를 이용한 댄스영상으로부터의 실시간 감정인식

박한훈, 박종일, 우운택*
[hanuni, jipark]@mr.hanyang.ac.kr, wwoo@kjist.ac.kr
한양대학교
*광주과학기술원

Realtime Emotion Recognition from Dance Image Sequence Using TDMLP

HanHoon Park, Jong-Il Park, Woontack Woo*
Hanyang Univ.
*K-JIST

요약

최근 신체움직임영상으로부터 감정을 인식하고자 하는 연구가 활발이 이루어지고 있다. Woo 등은 라반의 이론에 기반하여, 댄스 시퀀스에 간단한 영상처리를 수행해서 특징량(feature)을 추출한 후, 시간지연 다층 인식자(TDMLP : Time Delay Multi-Layer Perceptron) 신경망을 이용해서 댄스로부터 감정을 인식할 수 있는 기법을 제안했다. TDMLP 는 감정공간을 비선형적으로 분류하고, 입력 노드에 시간지연을 두어 특징량의 순간값이 아닌 변화 모습을 파악할 수 있도록 하고 있다. 그런데 이 방법에서는 신경망의 입력으로 이용되는 특징량의 분류 특성이 성능에 큰 영향을 주므로, 이를 적절히 분류하는 것이 매우 중요하다.

본 논문에서는 주성분 분석법(PCA: Principal Component Analysis)을 이용해서 특징량들의 기여도(contribution measure)를 계산한 후, 특징량들의 선형 결합을 시간지연 다층 인식자의 입력값으로 이용함으로써, 특징량의 정보 손실을 최소화하면서도 만족할만한 분류 특성을 갖게 하는 기법을 제안한다. 또한, 제안된 감정인식 방법을 응용한 예로, 사람이 다양한 장르의 음악에 맞춰 댄스 동작을 수행하면, 이를 평가해주는 “Express Yourself”라는 시스템을 소개한다.

1. 서론

말과 더불어 몸동작은 인간의 중요한 의사전달 수단이다. 특히, 댄스는 인간의 감정을 극대화시킨 몸동작으로, 댄스를 정량적으로 분석하여 감정을 인식하려는 연구는 꾸준히 수행되어 왔다. Woo 등은 라반의 이론에 기반하여[1] 댄스 시퀀스로부터 특징량(feature)을 추출한 후, 신경망을 이용해서 특징량을 비선형적으로 분류함으로써, 감정을 인식하는 기법을 제안했다[2]. 특히, 댄스의 흐름을 분석하기 위하여 시간지연 다층 인식자를 제안했다. 그러나, 신경망의 입력으로 사용되는 특징량이 근본적으로 쉽게 분류될 수 없다면, 시간지연을

이용한다고 하더라도 신경망의 인식율은 개선될 수 없으며, 오히려 저하될 수도 있다. 이를 해결하기 위해서 경험적인 판단을 이용하거나[2], 특징량의 기하학적인 특성을 이용해서[4] 분류하기 쉬운 특징량을 선정하는 방법이 제안되어 있으나, 단순히 특징량을 선정하는 것은 특징량에 포함되는 잡음에 의한 왜곡을 제거할 수 없으며, 근본적으로 선정되지 못한 특징량의 정보를 잃어버리게 되므로, 효율적이지 못하다. 따라서, 본 논문에서는 주성분 분석법을 이용하여 특징량들의 기여도를 계산한 후, 이를 이용하여 특징량들의 선형결합을 신경망의 입력으로 이용함으로써 특징량에 포함된

잡음 정보를 제거하면서도 정보 손실을 최소화하는 방법을 제안한다.

제안된 방법을 통해 인간의 감정을 4 가지(기쁨-happiness or cheerfulness, 놀람-surprise, 분노-angry or disgust, 슬픔-sadness or loneliness)로 분류한다[3].

본 논문에서는 제안된 감정인식 방법의 성능과 응용 가능성을 입증하기 위해 “Express Yourself”라는 엔터테인먼트 시스템을 구현한다. 이 시스템은 4 가지 감정을 표현하는 다양한 장르의 음악을 제공하는데, 사용자가 하나를 택한 후, 그에 맞춰 댄스로 표현하게 되면 시스템은 그 음악이 담고 있는 감정을 사용자가 잘 표현하는지를 평가해준다.

본 논문은 다음과 같이 구성된다. 제 2 장에서는 실시간 감정인식 방법에 대해서 간단히 설명하고, 제 3 장에서는 실험과정 및 결과를 제시한다. 제 4 장에서는 “Express Yourself” 시스템에 대해서 설명하고, 제 5 장에서는 결론을 제시한다.

2. 실시간 감정인식

각 감정을 표현하는 댄스 시퀀스로부터 특징량들을 추출[3,4]한 후, 주성분 분석법을 이용하여 특징량들의 주성분을 추출하여 시간지연 다층 인식자의 입력으로 하고, 각 감정을 출력으로 함으로써, 특징량들이 댄스 시퀀스가 표현하는 감정 정보로 정확하게 맵핑되도록 시간지연 다층 인식자를 학습시킨다. 학습된 시간지연 다층 인식자를 임의의 영상에 적용시켜서, 임의의 영상이 표현하는 감정을 인식한다.

2.1 시간지연 다층 인식자

흔히 사용되는 신경망의 하나로, 다층 인식자(MLP : Multi-Layer Perceptron)는 입력층(input layer)과 출력층(output layer) 사이에 하나 혹은 여러 개의 숨김층(hidden layer)을 가진 피드포워드(feedforward) 망이다. 다층 인식자는 내부 뉴런(neuron)들이 비선형적인 특성을 가지기 때문에, 이론적으로 임의의 모양을 가지는 영역으로 데이터를 분류할 수 있다. 하지만, 댄스와 같은 다이내믹한 변화를 가지는 데이터를 분류하는 데는 만

족할만한 성능을 가지지 못한다. 시간지연 다층 인식자는 입력 층에 버퍼를 두고, 입력 데이터를 일정 기간 동안 지연시킨 후, 지연된 데이터들을 함께 입력으로 이용한다. 이는 데이터의 변화 모습을 파악하는 효과를 가짐으로써 다이내믹한 데이터를 보다 정확하고 강건하게 분류할 수 있다.

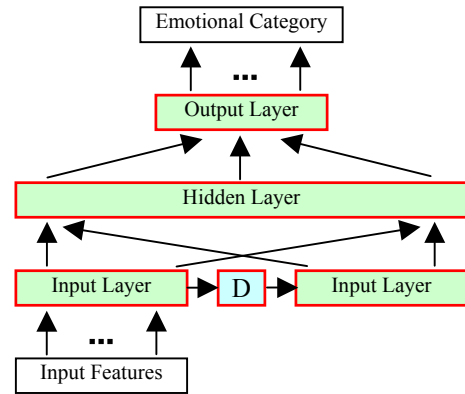


그림 1. 시간지연 다층 인식자.

2.2 주성분 분석법

시간지연을 이용하더라도 입력값의 특성에 따라 신경망의 성능은 크게 달라진다. 즉, 신경망의 입력으로 사용되는 특징량의 분류 특성이 좋지 못하다면, 시간지연을 이용하더라도 만족할만한 성능을 가질 수 없다. 주성분 분석법은 모든 특징량들의 정보를 압축함으로써 정보의 손실을 최소화하면서도 분류 특성이 좋은 특징량들의 기여도가 크게 나타나기 때문에 만족할만한 분류 특성을 유지할 수 있다.

m 프레임을 가진 댄스시퀀스로부터 모두 n 개의 특징량을 추출한 경우, 특징량들을 $m \times n$ 의 크기를 가지는 A 라는 행렬의 요소로 하여, A 에 특이값 분해(singular value decomposition)를 수행하면,

$$A = U \Sigma V^T$$

이 된다. 여기서 U 는 $m \times n$ 행렬, V 는 $n \times n$ 행렬이고, Σ 는 $m \times n$ 행렬로서

$$\Sigma = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \text{where } \mathbf{D} = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{pmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad r \leq m, n$$

이다. 여기서, σ_i 를 특이값(singular value)이라 하며, 행렬 \mathbf{U} 의 i 번째 열이 σ_i 에 대한 고유벡터 (eigen vector)가 된다.

σ_i 값이 큰 r 개의 고유벡터를 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ 이라 하고, 특징량들을 각각 $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n$ 이라 하면,

$$\begin{aligned} \mathbf{u}_1 &= \alpha_{11}\mathbf{f}_1 + \alpha_{12}\mathbf{f}_2 + \cdots + \alpha_{1n}\mathbf{f}_n, \\ \mathbf{u}_2 &= \alpha_{21}\mathbf{f}_1 + \alpha_{22}\mathbf{f}_2 + \cdots + \alpha_{2n}\mathbf{f}_n, \\ &\vdots \\ \mathbf{u}_r &= \alpha_{r1}\mathbf{f}_1 + \alpha_{r2}\mathbf{f}_2 + \cdots + \alpha_{rn}\mathbf{f}_n \end{aligned}$$

이 되고, 이를 다시 쓰면,

$$\mathbf{u}_i = \mathbf{F}\boldsymbol{\alpha}_i \quad \text{for } i = 1, 2, \dots, r$$

이 된다. 여기서 $\boldsymbol{\alpha}_i = [\alpha_{i1} \alpha_{i2} \cdots \alpha_{in}]^T$, $\mathbf{F} = [\mathbf{f}_1 \mathbf{f}_2 \dots \mathbf{f}_n]$ 이다. 이 식에서 $\boldsymbol{\alpha}$ 는 고유벡터 \mathbf{u}_i 에 대한 각 특징량의 기여도를 나타내며, 이를 구하면

$$\boldsymbol{\alpha}_i = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{u}_i$$

가 된다. 구해진 기여도를 이용해서 매 프레임마다 특징량들의 선형결합을 구하면,

$$\begin{aligned} p_1 &= \alpha_{11}f'_1 + \alpha_{12}f'_2 + \cdots + \alpha_{1n}f'_n, \\ p_2 &= \alpha_{21}f'_1 + \alpha_{22}f'_2 + \cdots + \alpha_{2n}f'_n, \\ &\vdots \\ p_r &= \alpha_{r1}f'_1 + \alpha_{r2}f'_2 + \cdots + \alpha_{rn}f'_n \end{aligned}$$

이 된다. 여기서 f'_i 는 매 프레임으로부터 추출된 특징량을 나타낸다. 최종적으로 p_1, p_2, \dots, p_r 은 특징량들의 주성분(principal component)으로서, 신경망의 입력으로 사용된다.

3. 실험 및 결과

3.1 댄스 시퀀스로부터 특징량 추출

신체 각 부위에 대한 복잡한 3 차원 정보를 실시간에 안정되게 추출하는 것은 현재의 기술 수준으로는 매우 어렵기 때문에, 본 논문에서는 일반 PC 에서도 실시간 처리가 가능한 시스템 구축에 중점을 두고, 3 차원 동작을 2 차원 영상으로부터 실시간으로 추출할 수 있는 사각박스나 무게 중심의 좌표와 같은 특징량으로 단순화시킨다.

댄스 시퀀스의 각 프레임에 간단한 영상처리 [3,4]를 수행하여 그림 2 와 같은 이진 이미지를 얻은 후, 이진 이미지로부터 무게중심의 좌표, 사각박스 중심의 좌표, 사각박스의 가로길이와 세로길이의 비, 실루엣 영역의 크기, 사각박스 영역의 크기 등의 특징량을 추출한다.

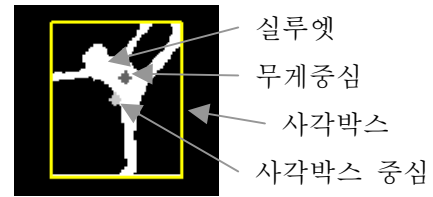


그림 2. 영상으로부터 특징량 추출.

3.2 필터링

매 프레임마다 추출된 특징량은 잡음 정보를 가지기 때문에 10 프레임 간격으로 평균한다. 여기서, 평균하는 간격이 중요한데, 너무 짧으면 잡음에 민감해지고, 너무 길면 특징량의 분류 특성이 나빠질 뿐 아니라, 시스템의 반응 속도가 느려진다[3].

3.3 특징량으로부터 주성분 추출

필터링을 거친 특징량들을 요소로 가지는 행렬을 만든 후, 주성분 분석법을 이용해서 특이값이 큰 7 개의 주성분($p_1, p_2, p_3, p_4, p_5, p_6, p_7$)을 추출한다.

3.4 시간지연 다층 인식자를 이용한 주성분 분류

추출된 7 개의 주성분은 d 개의 시간지연을 가지고 버퍼에 저장된 후, $7 \times d$ 개의 주성분이 신경망

의 입력으로 사용된다. 출력 층은 4 개, 숨김 층은 7×d×4 개가 된다. 신경망은 각 감정을 표현하는 댄스 시퀀스로부터 추출된 7×d 개의 주성분이 각 감정으로 정확하게 맵핑되도록 학습된다. 학습된 신경망을 임의의 댄스 시퀀스에 적용함으로써, 임의의 댄스 시퀀스가 표현하는 감정을 인식한다.

3.5 주성분 분석법을 이용한 성능 개선

주성분 분석법을 이용함으로써 신경망의 성능이 개선됨을 보이기 위해 경험적인 판단이나 기하학적인 분석을 이용했을 때와 비교해 보았다.

우선, 라반의 이론에 근거하여 댄스 시퀀스로부터 추출된 특징량들 중에서 신체 움직임을 잘 대변해 줄 수 있는 특징량들을 경험적으로 선정했다. 이 방법은 단순히 각 특징량이 얼마나 신체 움직임을 잘 대변할 수 있는지를 고려한 것이므로, 특징량의 분류 특성을 고려하지는 못한다. 즉, 실험 환경에 따라서 선정된 특징량의 분류 특성은 크게 달라질 수 있다. 무게중심의 좌표의 가속도량, 사각박스의 크기, 실루엣의 크기의 속도와 가속도량을 신경망의 입력으로 이용했을 경우, 시간지연의 개수에 따른 신경망의 성능은 그림 3 과 같다.

다음으로, 각 특징량의 기하학적인 분포를 보고 그룹화가 잘 되는 특징량들을 선정했다. 이 방법은 앞의 경험적인 방법과 달리, 특징량의 분류 특성에 초점을 맞춘 것이다. 경험적인 방법에 비해 신경망 입력의 분류 특성을 개선함으로써, 신경망의 성능은 다소 개선되었지만, 선정된 특징량에 포함된 잡음 정보에 의해 분류 특성은 저하되었다. 또한, 이 방법은 경험적인 방법과 마찬가지로 선정되지 못한 특징량의 정보를 이용할 수 없다. 무게중심의 좌표, 사각박스의 크기, 실루엣의 크기 및 속도, 가속도량을 신경망의 입력으로 이용했을 경우, 시간지연의 개수에 따른 신경망의 성능은 그림 4 와 같다.

마지막으로, 주성분 분석법을 이용하여 특징량으로부터 7 개의 주성분을 추출하여 이를 신경망의 입력으로 이용했다. 이 방법은 특정 특징량을

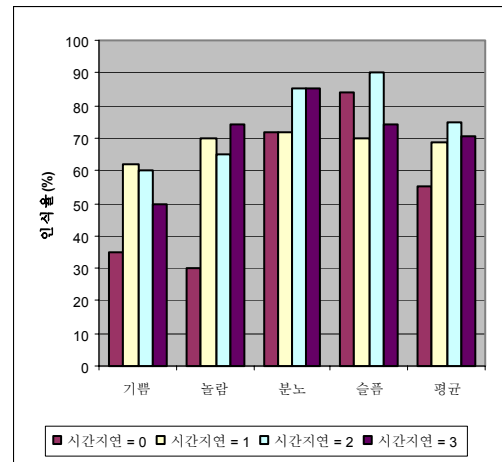


그림 3. 경험적인 판단에 의해 특징량을 선정했을 경우, 시간지연 수에 따른 신경망의 성능.

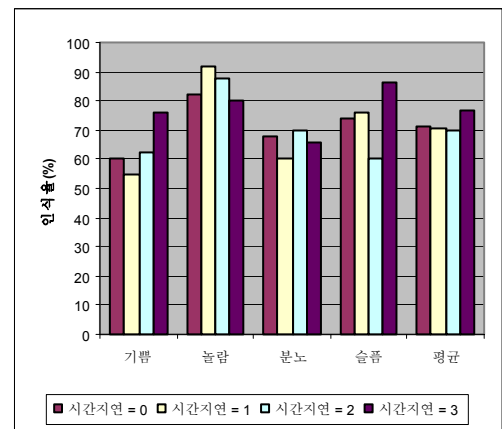


그림 4. 기하학적인 분석을 통해 특징량을 선정했을 경우, 시간지연 수에 따른 신경망의 성능.

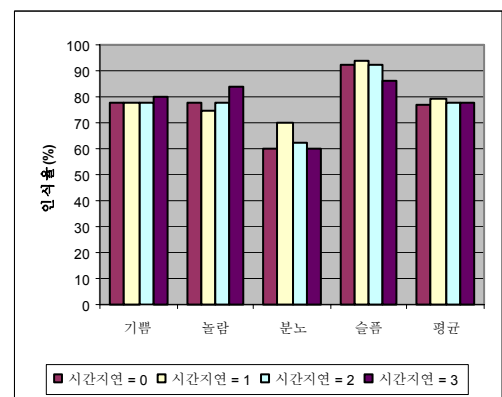


그림 5. 주성분 분석법을 이용했을 경우, 시간지연 수에 따른 신경망의 성능.

선정하는 방법과 달리 모든 특징량의 정보를 거의 잃어버리지 않으며, 분류 특성에 저해되는 잡음 정보를 제거함으로써 분류 특성을 크게 개선할 수

있다. 그림 5는 시간지연의 개수에 따른 신경망의 성능을 보여준다.

그림 3, 4, 5에서 보는 것처럼 경험적인 판단이나 기하학적인 분석을 이용했을 때에 비해, 주성분 분석법을 이용함으로써, 인식률은 크게 개선되었으며, 시간지연 수에 따른 성능 변화도 안정적이었다. 결과적으로, 주성분 분석법을 사용함으로써 평균적으로 80%의 인식률을 가졌다.

이 결과에서 주목할만한 내용은 시간지연에 따른 신경망의 성능 변화가 일관되지 않다는 것이다. 즉, 시간지연을 사용함으로써, 신경망의 성능은 개선되지만, 시간지연의 수가 늘어난다고 해서 신경망의 성능이 개선되는 것은 아니다. 평균 인식률을 고려해 볼 때, 하나의 시간지연만을 사용했을 때 신경망의 성능이 가장 좋았다.

4. 실시간 감정인식의 응용 - Express Yourself

4.1 개요

“Express Yourself”는 감정인식 기술의 응용성을 보여주기 위한 엔터테인먼트 시스템이다. 그림 6은 시스템의 개요를 보여준다.

각 감정 정보를 담고 있는 음악을 미리 준비한 후, 사용자가 음악 리스트를 보고 원하는 음악을 고르면, 그에 맞는 배경 영상과 함께 음악이 재생된다. 사용자가 음악을 들으면서 표현하고 싶은 동작을 자유롭게 취하면, 시스템은 사용자가 표현하는 감정이 무엇인지를 추정한 후, 음악이 담고 있는 감정과 비교함으로써, 사용자의 동작을 객관적인 입장에서 평가해준다.

여기서, 중요한 것은 각 감정을 담고 있는 음악을 선정하는 작업인데, 무용 전문가들의 조언을 바탕으로 선정함으로써, 객관성을 가지도록 노력하였다.

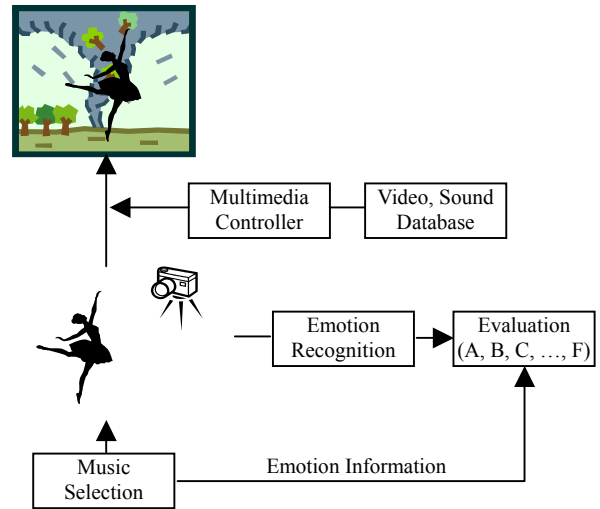


그림 6. 시스템 개요.

4.2 앞으로의 계획

“Express Yourself”는 본 논문에서 제안한 감정 인식 기법을 이용하여 구현되었으며, 사람의 자유로운 댄스 동작으로부터 감정을 추정함으로써 사람의 감정 표현 능력을 평가해 주었다. 그러나 보다 객관적인 성능 검증을 위해, 추후 과제로 주관 평가를 실시하여 시스템의 성능을 검증하는 작업이 필요하다.

5. 결론

본 논문에서는 댄스 시퀀스로부터 실시간으로 감정을 인식하기 위해 시간지연 다층 인식자를 이용해서 댄스 시퀀스로부터 추출된 특징량들을 비선형적으로 분류하는 방법을 소개했다. 추출된 특징량들은 기본적으로 잡음 정보를 가지고 있으며, 분류 특성이 좋지 않은 특징량들을 시간지연 다층 인식자의 입력으로 이용하게 되면, 시간지연 다층 인식자는 만족할만한 성능을 가질 수 없었다. 따라서, 본 논문에서는 주요소 분석법을 이용함으로써, 특징량들의 분류 특성을 크게 개선함으로써, 시간지연 다층 인식자의 성능을 극대화했다. 결과적으로, 시간지연 다층 인식자는 평균적으로 80%의 인식률을 보였다.

또한, 제안된 감정인식 기법의 응용성을 검증하기 위해 “Express Yourself”라는 엔터테인먼트 시스템을 구축하였다. 이 시스템은 각 감정을 담고 있는 음악을 듣고 사용자가 느끼는 감정을 댄

스로 표현했을 때, 사용자의 댄스를 객관적으로 평가해 주었다. 추후 과제로, 주관평가를 실시하여 시스템의 객관적인 성능을 검증하는 작업이 필요하다.

참고 문헌

- [1] R. Raban, *Modern educational dance*, Trans-Atlantic Publications, Inc. 1988
- [2] W. Woo, J. Park, Y. Iwadata, "Emotion analysis from dance performance using time-delay neural networks," *Proc. Of CVPRIP*, Vol. 2, pp. 374-377, 2000
- [3] 박한훈, 박종일, 우운택, "신체움직임에 대한 컴퓨터비전 기반 실시간 감정인식," *신호처리합동 학술대회논문집*, pp. 157-160, 2001
- [4] 박한훈, "신체 감정표현의 실시간 인식에 관한 연구," *한양대학교 석사학위 논문*, 2001년 12월
- [5] G. Strang, *Linear algebra and its applications*, Harcourt Brace Jovanovich, Inc. 1988
- [6] G. Nakos, D. Joyner, *Linear algebra with applications*, Brook/Cole Publishing Company, 1998

