깊이 영상을 활용한 조명 변화에 강건한 얼굴 표정 인식

예재원 ○, 윤국진
광주과학기술원 전기전자컴퓨터공학부
{yjw1861, kjyoon}@gist.ac.kr

요 약

본 논문에서는 깊이 영상 입력에서 사용자의 얼굴을 인식하고 표정을 판별하는 프레임워크를 제안한다. 기존의 표정 인식 기술은 2 차원 컬러 영상을 기반으로 연구가 이루어져 왔기 때문에 조명 변화에 취약하다. 제안한 방법은 조명 변화에 강건한 얼굴인식 및 표정판별을 위하여 깊이 영상을 입력으로 활용하였으며, 특징 기술자로 Histogram of Oriented Gradients(HOG) 특징 값(Feature)을, 학습 및 인식 과정에는 Support Vector Machine(SVM)을 사용하여 모델 기반의 표정 인식 프레임워크를 제안하였다. 제안하는 프레임워크의 타당성을 증명하기 위해 정성적, 정량적 평가를 수행하였으며, 그 결과 극심한 조명 변화에도 강건하게 얼굴을 인식하고 표정을 분류해낼 수 있음을 확인할 수 있었다.

1. 서론

영상 기반 감정 인식 기술은 사용자의 얼굴표정과 제스처를 읽어 들여 이에 따라 감정상태를 인지하는 기법으로 컴퓨터 비전 및 HCI 분야에서 많이 연구되어온 분야이다. 해당 기술은 최근 인간과 컴퓨터간의 상호작용이 중요해짐에 따라 그 중요성이 증대되고 있으며 영상광고 마케팅, 공연장 등 다양한 상황에서 사용자의 반응을 분석하여 이를 컨텐츠 제작에 활용하는 등 산업현장에서도 관련 연구가 활발하게 진행되고 있다.

가장 보편적인 영상 기반 감정인식 기술은 각기 다른 감정을 표현할 때 나타나는 얼굴 근육 변화 즉, 표정을 인식하는 것으로 대부분의 연구가 2 차원 컬러 영상에 대해 이루어져 왔다[1][2]. 2 차원 컬러 영상 기반 표정인식 기법은 주로 얼굴 영역의 명도 값을 활용하는 경우가 많으며 균일한 조명 환경에서는 높은 성능을 보이고 있다[3]. 하지만 조명 변화가 심한 경우에는 조명 변화에 따라 영상의 명도 값 변화 또한 매우 크기 때문에 얼굴영상 정보가 손실 또는 왜곡되어 표정 인식률이 크게 떨어지는 문제점이 있다.

본 논문에서는 이러한 컬러 영상 기반의 표정인식 기법들의 문제점을 해결하기 위해 3 차원 깊이 영상을 입력을 입력으로 하여 사용자의 표정을 인식해내는 프레임워크를 제안한다. 본 연구에서 깊이 영상 획득을 위해 마이크로소프트사의 Kinect 센서를 사용한다. Kinect 센서는 깊이 영상 획득을 위해 적외선 파장의 빛을 사용하는데, 이 대역의 빛은 조명에서 발생시키는 가시 영역의 빛과 다른 영역에 위치하기 때문에 조명의 영향을 전혀 받지 않고 균일한 품질의 얼굴 영상을 얻을 수 있어 이를 이용한 얼굴 인식 연구가 활발히 진행되고 있다[4][5].

깊이 영상 기반의 표정 인식 프로세스는 크게 깊이 정보 추출 단계, 얼굴 검출 단계, 표정 인식 단계의 세 가지로 구성되어 있으며, 매 프레임마다 단계적으로 수행된다. 깊이 정보 추출 단계는 Kinect 에서 얻어지는 깊이 영상에서 배경과 같은 불필요한 정보를 제거하고 정교한 얼굴 영상을 얻기 위한 필터링하는 단계이다. 다음으로 얼굴 검출 단계에서는 필터링 된 영상에서 얼굴의 위치를 검출한다. 이를 위해 먼저 머리를 검출하여 검색 범위를 줄인 후 정교하게 얼굴을 추출하는 과정을 거친다. 마지막으로 표정 인식 단계에서는 입력된 얼굴 영상과 미리 학습된 표정 모델을 비교하여 현재 관객의 표정이 어떤 것인지 인식하게 된다.

2. 깊이 영상을 이용한 표정 인식 방법

2.1 깊이 정보 추출

깊이 정보 추출 단계에서는 입력된 깊이 영상이 배경 제거 및 필터링 과정을 거쳐 표정 인식에 적합한 영상으로 보정된다. 먼저 배경 제거 과정은 Kinect 가 고정되어 있다는 것에 착안하여 일정 범위 밖에 존재하는 깊이 정보들은 모두 배경으로 판단하고 제거하였다. 다음으로 Kinect 의 플리커링(Flickering)으로 인한 깊이 영상의 빈 공간을 채우기 위해 Median 필터를 적용하여 Hole filling 을 수행한다. 마지막으로 정교한 깊이 영상을 얻기 위

해 필터링 된 깊이 영상에 정규화(Normalization) 과정을 수행하게 되면 얼굴 검출을 위한 입력영상이 얻어진다. 각 단계에서의 결과는 아래 그림에서 확인할 수 있다.

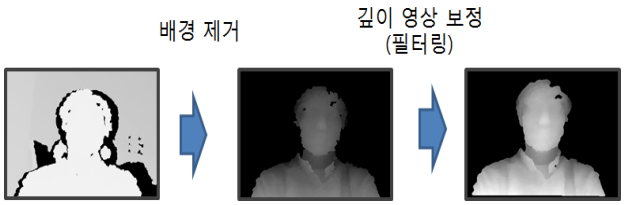

그림 1. 깊이 정보 추출 과정

2.2 얼굴 검출 단계

얼굴 검출 단계는 크게 학습 단계와 검출 단계로 구성되어 있다. 먼저 학습 단계는 미리 만들어진 학습 데이터를 활용하여 머리 및 얼굴 모델을 정의하는 단계로 얼굴 검출을 수행하기 전에 미리 학습해두어야 한다. 다음으로 검출 단계에서는 미리 학습된 머리 및 얼굴 모델을 바탕으로 입력 영상에서 머리와 얼굴을 검출하는 단계이다.

머리 검출: 입력된 깊이 영상에서 머리를 검출하기 위해서는 먼저 학습을 통해 머리 모델을 정의할 필요가 있다. 이를 위하여 본 논문에서는 머리 영상 1300 여장, 배경 영상 2800 여장을 수집하여 학습에 사용하였다.

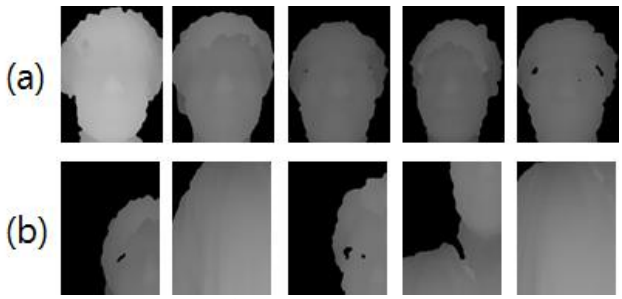

그림 2. 머리 검출을 위한 학습 데이터 (a)머리 데이터, (b)배경 데이터

머리 모델을 정의하기 위해 각 학습 영상들은 일정한 크기로 정규화한 후 Histogram of Oriented Gradients(HOG)[6]를 특징 값(Feature)으로 사용하였다. 최종적으로 HOG 로 표현된 각 학습 데이터들을 Support Vector Machine(SVM)[7] 프레임워크에 적용시켜 분류기를 작성하였다. SVM 은 Positive 데이터와 Negative 데이터 사이의 간격을 최대화하는 분류기를 생성하는 기법으로, 학습 단계에서 관찰할 수 없었던 변형된 데이터가 입력으로 주어져도 그에 강인하게 분류할 수 있는 기법이다. 따라서 표정 변화 및 머리 스타일, 안경 착용 유무 등에 따라 그 모양이 바뀔 수 있는 머리를 검출하는 데에 적합한 프레임워크라고 할 수 있다.

검출 단계에서 머리 검출은 슬라이딩 윈도우(Sliding window) 방식으로 전체 영역에 대해 수행하였는데, 이때 이미지 피라미드(Image pyramid)를 이용하여 다양한 크기의 머리를 검색할 수 있도록 하였다. 각 윈도우에서 얻어진 템플릿 영상을 미리 학습된 분류기에 넣어 각각 머리인지 아닌지를 검사한 후, 이를 이용하여 유사도 지도를 생성한다. 이 중에서 최고 유사도를 가지는 위치를 찾으면 그 위치가 검출된 머리의 위치가 된다.

얼굴 검출: 얼굴 검출 과정 역시 머리 검출 과정과 유사하게 수행된다. 먼저 학습을 위해 본 연구에서는 총 6000 여장의 얼굴 데이터를 선별하였다. 이때 추출된 입력 영상은 얼굴 내부에 깊이 차이가 거의 없어 변별력이 높은 특징 값을 생성하기 어려우므로, 얼굴 내부의 깊이 차이를 극대화하기 위해 깊이 정규화 기법을 적용해야 한다.

깊이 정규화는 주어진 얼굴 템플릿에서 최대값을 찾은 후 이 값이 255 가 되도록 1 차적으로 정규화를 한 뒤 임계 값(180~200) 이하의 값을 제거한 후 다시 2 차로 정규화하여 수행된다. 그 결과로 아래 그림의 (b)와 같이 원본 얼굴 영상에 비해 눈, 코, 입이 확연히 드러나는 얼굴 영상을 얻을 수 있었다.

최종적으로 정규화된 얼굴 깊이 영상들을 HOG 를 이용하여 표현한 후 이를 Principle Components Analysis(PCA)를 통해 학습하였다. PCA 는 여러 차원으로 표현되는 데이터 집합을 대상으로 해당 집합을 가장 잘 표현할 수 있는 축(주성분)을 찾는 기법이다. PCA 수행 후 만들어지는 주성분 공간에 데이터들을 투영하면 새로운 특징 벡터들을 얻을 수 있으며, 이 벡터들의 평균을 취하면 얼굴 모델을 생성할 수 있다.

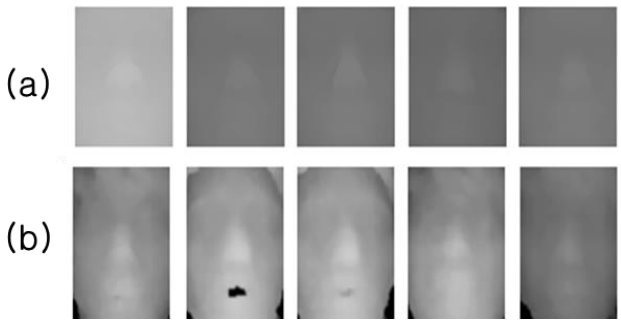

그림 3. 얼굴 검출을 위한 학습 데이터 (a) 원본 얼굴 영상 (b)정규화된 얼굴 영상

검출 단계에서는 머리 검출과 마찬가지로 슬라이딩 윈도우와 이미지 피라미드를 이용하여 여러 크기의 얼굴 템플릿을 만들고 정규화하여 이를 HOG 로 표현한다. 다음으로 HOG 로 표현된 특징 벡터들을 미리 학습된 주성분 공간에 투영한다. 이렇게 투영된 벡터들과 얼굴 모델과의 차이를 비교하여 가장 유사한 템플릿을 얼굴로 최종 선택하게

된다.

2.3 표정 인식 단계

본 연구에서는 공연 관람 시 자주 짓게 되는 6 가지 표정, 분노, 공포, 기쁨, 슬픔, 놀람, 무표정에 대한 모델을 만들고 이를 기반으로 관객의 표정을 인식하였다. 각 표정 모델을 학습하기 위해 총 10 명의 피험자를 대상으로 학습 데이터를 수집하였으며, 각 표정 별로 300 장의 데이터를 측정하여 총 18000 장의 학습 데이터를 획득하였다.

표정 별로 획득된 학습 데이터들은 얼굴 검출 단계에서와 마찬가지로 표정을 확연히 구분할 수 있도록 정규화 과정을 거친 후 HOG 를 이용하여 특징 벡터로 표현된다. 본 연구에서는 이렇게 얻어진 특징벡터를 Multi-class SVM 프레임워크에 적용하여 각 표정을 구분하였다.

기본적으로 SVM 은 선형 분류기이기 때문에 두 개의 클래스 구분만을 할 수 있다. 이를 여러 클래스를 구분하는 데에 활용하기 위해서는 다수의 SVM 분류기를 정의하고 각각의 결과를 취합하여 최종 결과를 선택하는 방식을 사용해야 한다. 여기서 다수의 SVM 분류기를 어떤 방식으로 구성하느냐에 따라 1-vs-all 방식과 1-vs-1 방식으로 구분할 수 있다. 1-vs-all 방식은 Positive 데이터에 구분하고자 하는 클래스의 데이터를 넣고 Negative 데이터에는 다른 클래스의 데이터를 넣어 분류기를 만드는 방식이다. 이 방식은 클래스 개수만큼의 분류기만 있으면 되므로 분류 횟수가 적다는 장점이 있지만 학습데이터가 선형으로 분류되지 않는 경우에는 정확도가 떨어지는 단점을 가지고 있다. 반면에 1-vs-1 방식은 두 개의 클래스를 짝으로 하는 클래스 조합에 대해서 모두 분류기를 만드는 방식으로, 정의되는 SVM 분류기의 수가 많지만 1-vs-all 에 비해 정확도가 높다는 장점을 가지고 있다.

본 연구에서는 분류해야 하는 클래스의 수가 한정적이고 정확도를 높이는 것이 중요하기 때문에 1-vs-1 방식의 Multi-class SVM 프레임워크를 사용하였다. 클래스의 수가 7 가지이므로 정의되는 분류기의 개수는 총 21 가지이며 각 분류기의 값을 시그모이드(Sigmoid) 함수를 이용하여 표현 후 취합하였다. 예를 들어 N 표정과 M 표정을 구분하는 N-M 분류기의 결과가 N 이라면 다음의 첫 번째 수식을 적용하고 반대라면 두 번째 수식을 적용하여 각 표정의 확률 값을 구했다.

$$S(x) = \frac{1}{1+\exp(-x)}$$: N-M 분류기 결과가 N 인 경우

$$S(x) = \frac{1}{1+\exp(x)}$$: N-M 분류기 결과가 N 인 경우

3. 실험 결과 및 분석

본 장에서는 성능 평가를 위해 직접 제작한 데이터 셋을 활용하여 제안한 얼굴 인식 및 표정 인식 알고리즘의 성능을 평가한다.

데이터 셋 생성: 본 논문에서는 각 단계별 알고리즘의 학습 및 성능을 평가하기 위해 암실 환경 및 일반 조명 환경에서 직접 촬영한 데이터 셋을 사용한다. 각 표정 별 1110 장씩을 확보하였으며, 관객이 안경을 착용한 경우와 아닌 경우에도 모두 적용이 가능하도록 각 피험자 마다 두 가지 영상을 촬영하였다. 데이터 셋에는 원본 영상 외에도 머리 검출을 위한 학습 영상들과 얼굴 검출 및 표정 인식 알고리즘의 학습에 적합하도록 얼굴 부분만을 잘라낸 영상들이 포함되어 있다.

얼굴 검출 모듈의 성능 분석: 본 논문에서는 개발한 사람 얼굴 검출 모듈의 성능을 분석하기 위해 미리 획득된 데이터 셋을 활용하였다. 머리 검출 및 얼굴 검출의 성능은 데이터 셋의 Ground truth 정보와 검출 알고리즘을 통해 추정된 얼굴의 개수와 위치, 크기가 얼마나 일치하는 지를 통해 판단하였다. 실험 결과 얼굴 검출 결과의 Ground truth 와의 영역 겹침 정도는 82.7%로 나타났으며, 다중 얼굴 검출 결과와 실제 사람 수의 일치 정도 역시 93.8%로 높게 나타났다. 장면 당 처리 속도는 영상에 사람 수가 몇 명이든 약 0.9 fps 로 동일하게 나타났다.

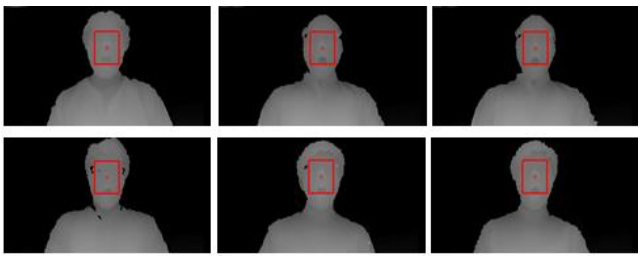

그림 4. 얼굴 검출 결과

표정 인식 모듈의 성능 분석: 표정 인식 모듈의 성능을 평가하기 위해 위와 마찬가지로 직접 제작한 데이터 셋을 이용하였으며, 전체 데이터 셋의 70%를 트레이닝을 위해 사용하고 나머지 30%를 이용해 정확도를 평가하였다. 실험 결과 평균 분류 정확도는 87.46%로 준수하게 나타났다.

표 1. 제안된 알고리즘의 성능분석

	중립	화남	혐오	공포	기쁨	슬픔	놀람
중립	96.73	0	0	0	0	3.03	0
화남	0	72.73	0	0	0	0	0
혐오	0	18.18	60.61	0	0	3.03	0
공포	3.03	0	21.21	100	12.12	0	0
기쁨	0	9.09	18.18	0	87.88	0	0
슬픔	0	0	0	0	0	93.94	0
놀람	0	0	0	0	0	0	100

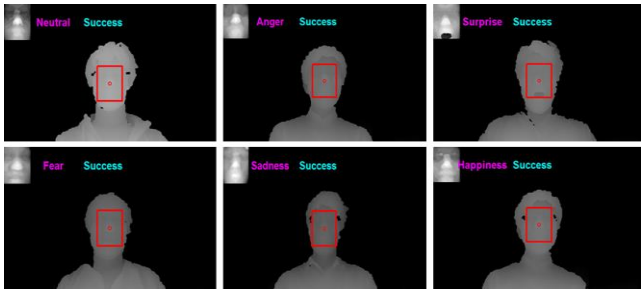

그림 5. 표정 인식 결과

4. 결론

본 논문에서 제안된 표정 인식 프레임워크는 조명의 변화가 심한 공연장과 같은 환경에서도 정확한 인식 성능을 보였으며, 기존의 얼굴 검출 및 표정 인식 기법들을 사용할 수 없던 환경에서도 적용이 가능하다는 점에서 그 의미가 크다고 볼 수 있다. 또한, 연구 수행 과정에서 확보한 데이터 셋은 추후 관련 연구들의 성능 평가를 위한 데이터로 활용할 수 있을 것으로 기대된다.

감사의 글

이 연구는 “문화체육관광부 및 한국콘텐츠진흥원의 2016 년도 문화기술연구개발지원사업”의 지원을 받아서 수행되었다.

참고문헌

[1] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, 1978

[2] Suk, Myunghoon, and Balakrishnan Prabhakaran. "Real-Time Mobile Facial Expression Recognition System--A Case Study." 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2014.

[3] Lo Presti, Liliana, and Marco La Cascia. "Using Hankel matrices for dynamics-based facial emotion recognition and pain detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015.

[4] Fanelli, Gabriele, Juergen Gall, and Luc Van Gool. "Real time head pose estimation with random regression forests." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[5] Papazov, Chavdar, Tim K. Marks, and Michael Jones. "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[6] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.

[7] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.