실감 음향 재생을 위한 영상기반의 실시간 화자 위치 검출

임재현, 이철희
연세대학교 전기전자공학과

Real-Time Vision Based Speaker Location Detection for Realistic Audio Reproduction

Jaehyun. Lim, Chulhee. Lee
Dept. of Electrical and Electronic Engineering, Yonsei University
Email: sigenl@hanmail.net, chulhee@yonsei.ac.kr

요약

일반적으로, 화상회의에서 화자의 위치를 검출하는 것은 음향 신호를 기반으로 이루어져 왔다.[1] 그러나 물리적인 환경의 제약이나 화자 검출 시스템의 한계를 벗어나는 노이즈가 발생하는 경우에는 검출 시스템의 성능저하를 초래하게 된다. 본 논문에서는 음향 기반의 검출 시스템과 독립적으로, 혹은 상호 보완적으로 사용될 수 있는 영상 기반의 화자 검출 알고리즘에 대하여 제안하고자 한다. 화자의 위치에 관한 정보는 화상회의에 한층 사실감을 부여하는 3 차원 오디오 재생에 사용될 수 있다.

1. 서론

화자 검출과 관련한 기존의 연구들은 대부분 마이크로폰 어레이(Microphone array)를 이용하는 등의 음향 기반 검출 알고리즘이었다[6]. 그러나 음향 기반의 알고리즘은 물리적 환경에 크게 의존하고, 노이즈에 의한 성능 저하가 두드러진다는 약점을 가지고 있다. 따라서, 최근의 화자 검출은 영상으로부터 얻은 화자 검출 정보를 이용하여 검출 정확도를 향상시키는 오디오 비주얼(Audio-Visual) 알고리즘에 초점을 맞추고 있다[1]. 본 논문에서는 화자 위치를 찾기 위해 영상 시퀀스에서 움직이는 입을 찾는 알고리즘에 대해 논하고자 한다. 특히 영상 정보를 실시간으로 처리하여 화자를 검출하는 데 초점을 맞추었다. 여러 명의 실험자를 대상으로 획득한 영상에 대한 실험을 통해서 화상회의와 유사한 상황 하에서 제안된 알고리즘이 성공적으로 화자 검출을 수행하는 것을 확인하였다.

2. 얼굴 검출 이론

제안된 알고리즘에서, 먼저 입력 영상으로부터 얼굴영역을 분리해내게 된다. 만약 얼굴 영역을 성공적으로 추출해 내면, 눈이나 코, 입과 같은 얼굴의 특징점들 역시 얼굴 영역으로부터 손쉽게 추출이 가능하다. 얼굴검출과 관련해서는 많은 이론들이 소개되어 있다[5]. 여기서는 그 중에서도 일반적인 피부색 모델을 사용하기로 한다[5]. 이 이론은 아래에 보다 자세히 기술되어 있다.

2.1 일반적 피부색 모델

사람의 피부는 2 차원 컬러 공간상에서 특정한 색도신호 값을 갖는 좁은 영역으로 군집되는 특징을 갖는다[2, 3]. 색도변환에 의한 색도신호 값은 다음과 같다.

$$r = R/(R+G+B)$$
$$g = G/(R+G+B) \quad (1)$$

위 등식에서 R, G, B 는 각각 red, green, blue 를 의미하고 r, g 는 붉은색 색도신호와 녹색 색도신호를 나타낸다. 다양한 인종의 차이에 무관하게 사람의 피부에 해당하는 색도신호 값들은 특정 구간에서 높은 값을 보이는 것으로 알려져 있다[3]. 본 논문에서는 이 피부색 모델을 사용하여 입력 영상으로부터 얼굴영역을 추출하였다.

2.2 얼굴 영역의 구분

제안된 알고리즘의 첫 단계는 위에서 설명한 바와 같이 주어진 입력 영상으로부터 색도 정보를 이용하여 얼굴 영역을 추출해내는 것이다. 특히, 입력 영상의 각 점들을 컬러 공간에서 피부 영역과 그렇지 않은 두 개의 영역으로 구분해 내었다. 이와 같은 분류 결과로부터 피부 영역을 표시하는 이진영상을 구성할 수 있다.

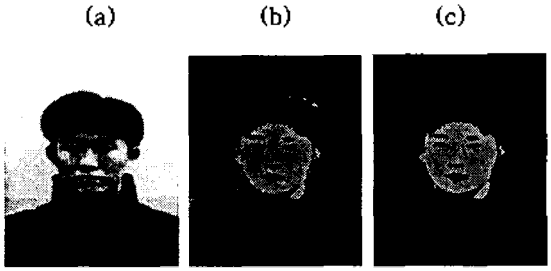

그림 1 (a) 원영상 (b) 피부영역의 이진영상
(c) 최종적으로 얻어진 얼굴 영역

그림 1 은 얼굴영역을 추출하는 과정을 나타낸다. 그림 1-(b)는 점들을 분할한 결과로부터 구성한 이진영상이다. dilation, erosion 과 같은 후처리 기법들을 통해서 최종적으로 원하지 않는 영역을 제거한 그림 1-(c)와 같은 결과를 얻는다.

3. 화자 검출

얼굴 영역이 성공적으로 추출된 다음에는 입의 위치를 찾고 입의 움직임을 검출해내야 한다. 그림 1 에서도 볼 수 있듯이 얼굴의 특징점들과 피부 영역은 휘도신호에 있어서 두드러지는 차이를 가지고 있다. 따라서, 어렵지 않게 얼굴 영역으로부터 입을 포함한 특징점들을 추출해낼 수 있다.

3.1 입의 위치 검출

분할된 얼굴영역을 나타내는 영상으로부터 눈, 코, 입과 같은 특징점들을 추출하는 다양한 알고리즘이 소개되어 있다. 그러나 여기에서는 정지 영상이 아닌 비디오 데이터를 다루기 때문에 템플릿에 기반한 방법이나 신경망 이론을 이용한 방법은 계산상의 복잡도를 증가시킨다는 측면에서 적절하지 않다. 따라서 여기에서는 입의 위치를 실시간으로 찾는 간단하고 효과적인 알고리즘을 제안하고자 한다.

그림 1 에서 보았듯이 피부를 나타내는 얼굴영역으로 둘러싸인 몇몇의 영역이 얼굴의 특징점들을 나타낸다. 다시말하면, 특징점들은 얼굴영역에 포함되어 있으면서, 피부 영역에는 속하지 않는 특성을 가진다.

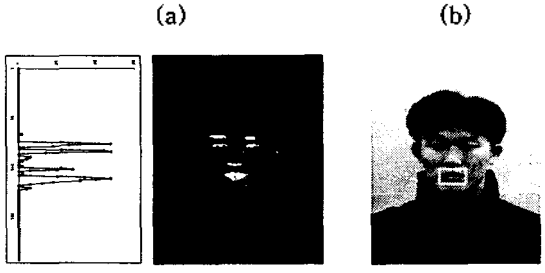

그림 2 (a) 얼굴의 특징점 영역 (b) 관심 영역

영역 채우기와 같은 후처리 과정을 거침으로써, 우리는 그림 2-(a)와 같이 특징점들을 얻어낼 수 있다.

그런 다음 수평방향과 수직방향 각각으로 투영된 히스토그램을 통해서 눈, 코, 입과 같은 특징점들의 위치를 얻어낸다[4]. 보다 자세히 설명하자면, 우선 그림 2-(a)의 영상을 수직선을 따라 투영시켜서 특징점들의 수직위치를 알아낸다. 투영 히스토그램에서 피크값들은 눈, 코, 입과 같은 특징점들에서 나타나게 되기 때문이다. 일단 이 과정을 통해서 입의 수직위치를 검출하고 나면, 다시 수평방향의 투영을 통한 히스토그램으로부터 입의 수평위치를 검출해내게 된다. 임의의 실험영상에 대한 수직, 수평방향의 히스토그램은 그림 3 에서 나타내었다.

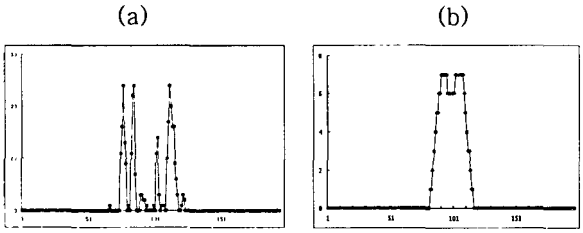

그림 3 (a) 수직 히스토그램 (b) 수평 히스토그램

3.2 입의 움직임을 검출하는 알고리즘

입의 위치를 찾고 나면, 같은 과정을 주어진 입력 영상의 매 프레임에 대해 수행하게 된다. 만약 입의 위치는 고정되어 있고 입을 나타내는 영역은 프레임에 따라 변화하게 된다면, 이는 화자가 그 시점에서 말하고 있다는 사실을 나타낸다. 이 경우, 인접한 프레임간의 차이가 특정 임계치를 넘어서게 되면 최종적으로 화자가 말하는 것으로 판단하게 된다. 그러나 초당 30 프레임으로 구성된 영상에서 인접 프레임간의 차이는 매우 작다는 점을 감안하여, 매 10 개 프레임간의 차이를 합한 결과로 입의 움직임을 판단하도록 하였다.

$$Diff(l)=\sum_{i=-5}^{4}\sum_{(x,y)\in B}\left|I(x,y,l)-I(x,y,l+i)\right| \tag{2}$$

이 식에서 B 는 입을 포함하고 있는 블록 영역을 나타낸다. 이 등식은 입의 움직임 검출 함수로서, 만약 이 함수값이 임계치를 넘어서게 되면, l 번째 프레임에서 화자가 말하고 있는 것으로 판단하게 된다.

4. 실험 결과

10 개의 영상 시퀀스를 사용하여 제안된 방법에 대한 실험을 수행하였다. 영상 시퀀스는 10 명의 화자를 포함하고 있으며, 각각의 영상 시퀀스에 대해서는 적어도 두 명 이상의 화자가 무작위 순서로 교대로 말을 하도록 하였다. 그림 4 는 임의의 실험 영상 시퀀스 중에서 한 프레임을 나타내고 있다.

그림 4 테스트 영상의 프레임

그림 5 는 제안된 방법을 실험영상에 적용한 결과를 나타내고 있다. 두 화자는 번갈아가며 말하게 되고 직접 눈으로 프레임을 보아가며 판단한 말하는 구간은 그림 5 의 상단에 표시되어 있다. 또한 입의 움직임 검출 함수를 이용해서 구해진 함수값들을 임계치와 함께 그래프로 도시하였다.

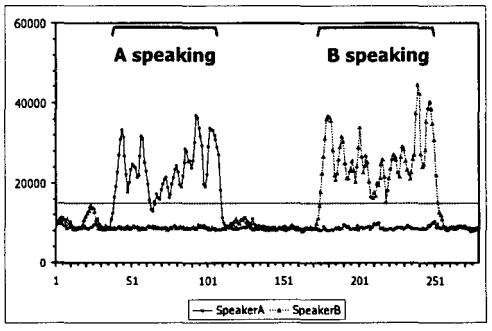

그림 5 입의 움직임 검출결과

함수값들이 임계치를 넘게 되면 말을 하는 것으로 판단한다고 하였으므로, 결과적으로 그래프에 의하면, A, B 순서로 말을 하였다고 결론지을 수 있다. 또한

제안된 방법에 의해 검출된 말하는 영역과 직접 눈으로 프레임을 보고 판단한 구간이 거의 일치하는 것도 확인할 수 있다.

표 1 Speak detection accuracies.

Index	Speaking Frames	None Speaking Frames	Total Frames
Data1	0.9251	0.8490	0.8932
Data2	0.9534	0.8640	0.9137
Data3	0.9276	0.8	0.8713
Data4	1.0	0.8350	0.9409
Data5	0.8993	0.9120	0.904
Data6	0.8862	0.9056	0.8937
Data7	0.9727	0.84	0.9190
Data8	0.9722	0.8958	0.9340
Data9	0.9652	0.875	0.9291
Data10	0.8910	0.8521	0.8745
Avg.	0.9393	0.8628	0.9073

10 개의 실험 영상에 대하여 이 과정을 반복해서 수행하였고, 표 1 은 그 결과를 나타낸다. 우선 눈으로 판단한 말하는 영역과 그렇지 않은 영역(이후로는 speaking region 과 non-speaking region 으로 표기)을 나누었다. 그렇다면 이 알고리즘을 적용한 결과에서는 두 가지 에러, 즉 speaking region 에서의 에러와 Non-speaking region 에서의 에러가 존재하게 된다. 표 1 에서 볼 수 있듯이, 제안된 알고리즘은 94%의 정확도로 말하는 구간을 성공적으로 검출해 내었다. 말하지 않는 구간을 포함한 전체 정확도는 86%로 다소 낮지만 많은 데이터를 사용한 통계적 분석을 통해 적절한 임계치 설정을 한다면, 보다 나은 결과를 얻을 수 있을 것으로 본다.

5. 결론

본 논문에서는 영상으로부터 화자의 위치를 찾는 방법에 대하여 소개하였다. 실험결과를 통해 얼굴영역 추출이 성공적으로 이루어진 경우, 제안된 방법이 화자를 검출하는 데 비교적 양호한 성능을 보임을 알 수 있었다. 그러나 실제적으로 말하는 과정에서 발생하는 화자의 다양한 움직임이나 얼굴표정은 에러를 유발할 수 있으며, 특히 화자의 입이 완전히 가려지는 경우에는 제안된 방법을 통해 올바른 결과를 얻기 힘들다. 이러한 경우에는 음향 기반의 화자검출 이론과의 상호 보완을 통하여 성능을 향상시킬 수 있을 것이다.

참고문헌

[1] Mingkun Li, Dongge Li, Nevenka Dimitrova, Ishwar Sethi : Audio-Visual Talking Face Detection. IEEE International Conference on Multimedia and Expo, Baltimore, US (2003)

[2] Shinjiro Kawato, Jun Ohya : Automatic Skin-color Distribution Extraction for Face Detection and Tracking. The 5th International Conference on Signal Processing, Beijing, China (2000)

[3] Gi-Jeong Jang, In-so Kweon : Robust Real-Time Face Tracking Using an Adaptive Color Model. International Symposium on Mechatronics and Intelligent Mechanical System for 21 Century, Changwon, Korea (2000)

[4] Raffaella Lanzarotti, Paola campadelli, N. Alberto Borghese : Automatic features detection for overlapping face images on their 3D range models. International Conference on Image Analysis and Processing, Palermo, Italy (2001)

[5] Douglas Chai and King N. Ngan : Locating facial Region of a Head-and-Shoulders Color Image. Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan (1998)

[6] Michael Siracusa, Louis-Philippe Morency, Kevin Wilson, John Fisher, Trevor Darrell : A multi-modal approach for determining speaker location and focus. 5th international conference on Multimodal interfaces, Vancouver, Canada (2003)