Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

오프라인 회의 기록 지원시스템을 위한 화자 인식 기법

Speaker recognition technique for offline conference recording system

박한무, 손윤식, 정진우[1]

[1] 서울시 중구 동국대학교 컴퓨터공학과

E-mail: lilees00@dongguk.edu, sonbug@dongguk.edu, jwjung@dongguk.edu

요 약

최근 영상 처리 기술이 발달함에 따라 다양한 응용시스템에 영상 처리 기술을 접목하려는 시도가 나타나고 있다. 특히 영상 내의 얼굴을 객체로 다루는 인식 기술의 발전으로 얼굴 정보를 이용한 기술의 응용 분야는 게임 및 카메라 등 다양한 분야에서 사용되고 있다.

본 논문에서는 오프라인 회의 보조 시스템에서 화자를 구분하기 위한 기법을 제시한다. 제안된 기법은 얼굴 객체 정보에서 화자 구별을 위한 특징 값을 제시하고, 이를 이용하여 얻어진 입 주변 엣지(Edge)를 이루는 픽셀들의 분산 값으로 화자 여부를 판단한다.

Key Words : Conference recording, Speaker recognition

1. 서 론

최근 영상 기기 및 영상 처리 기술이 발달함에 따라 다양한 응용시스템에 영상처리 기술을 사용하려는 시도가 많은 곳에서 나타나고 있다.

응용 시스템 중에서 흔히 볼 수 있는 시스템의 예로 화상회의 시스템을 들 수 있다. 화상회의 시스템은 원거리의 회의 참석자들이 물리적인 거리를 넘어서 실시간으로 회의를 할 수 있게 해준다.

그러나 편리한 화상 회의 시스템이 보편화되었음에도 불구하고, 아직까지는 오프라인으로 이루어지는 회의가 대부분이다. 때문에 오프라인 회의를 지원하기 위한 시스템이 필요하며, 그 중에서도 오프라인 회의 기록 지원 시스템을 제안한다.

오프라인 회의에서는 참석자의 발언이나 회의 분위기에 따른 보다 정확한 정보의 기록 등을 위해서 손으로 회의록을 작성하는 것이 대부분이다. 오프라인 회의 기록 지원 시스템이란, 이러한 필기형 회의록에 영상처리를 접목해 회의록을 영상으로 기록하는 것을 말한다. 이를 위해 필수적으로 필요한 기반 기술로 회의 참석자의 신원을 인식하는 식별 기술과 함께 현재 화자가 누구인지를 판단하는 화자 인식 기술이 필요하다.

본 논문에서는 오프라인 회의 기록 지원 시스템을 위한 화자 인식 방법을 제안한다. 제안하는 방법은 입의 위치 및 모양을 특정 짓지 않고, 입 근처 엣지(Edge)를 이루는 픽셀들이 얼마나 퍼져 있는지를 가지고 인식 대상이 말을 하고 있는 중인지 아닌지를 판단한다.

2. 관련 연구

2.1 오프라인 회의 기록 지원 시스템

오프라인 회의 기록 지원 시스템은 얼굴 인식 및 화자 인식 기술을 이용하여 오프라인 회의에서 있었던 일을 영상 회의록의 형태로 기록하는 시스템을 말한다.

대부분의 오프라인 회의는 회의록을 기록할 때 필기를 하고 문서의 형태로 남긴다. 이런 종류의 회의록은 기록 시 작성자의 주관적인 판단이 들어갈 수 있고, 작성자 역시 사람이기 때문에 회의 과정 중 누락된 내용이 생길 수 있다.

이런 단점을 개선하기 위한 방안 중 하나로 음성을 녹음하여 그것으로 회의록을 대체할 수도 있지만, 이러한 음성 회의록은 화자가 누구인지를 구별하기 위한 정보로 목소리만을 이용해야 하기 때문에 구별이 힘들며, 사람이 의사

소통하는데 큰 부분을 차지하는 몸짓이나 표정 등을 알 수가 없어 정확한 정보의 기록이 되지 못 하는 단점이 있다.

필기형 회의록과 음성 회의록의 단점을 개선하고자 제안된 것이 영상 회의록이다. 회의에서 있었던 일을 기록할 때 음성과 영상을 함께 기록하면, 음성 회의록에 비해 정리 및 열람이 편리하고, 회의에서의 생생한 분위기를 느낄 수 있게 된다.

그림 1. 회의 지원 시스템의 수행 과정

2.2 화자 위치 분석

화자를 인식하기 위해서는 먼저 화자의 대략적인 위치를 찾을 필요가 있다. 이것은 화자 인식을 하려고 해도 현재 그 사람이 기록되고 있는 영상에 있지 않다면 불가능하기 때문이다.

화자의 위치를 찾아내는 방법은 여러 가지가 있을 수 있으며, 본 논문에서는 원탁회의를 가정해 총 360도를 커버할 수 있도록 3개의 마이크를 사용한다. 그러면 각 마이크에 전달되는 음성신호의 크기비로 화자의 위치를 추정할 수 있다.

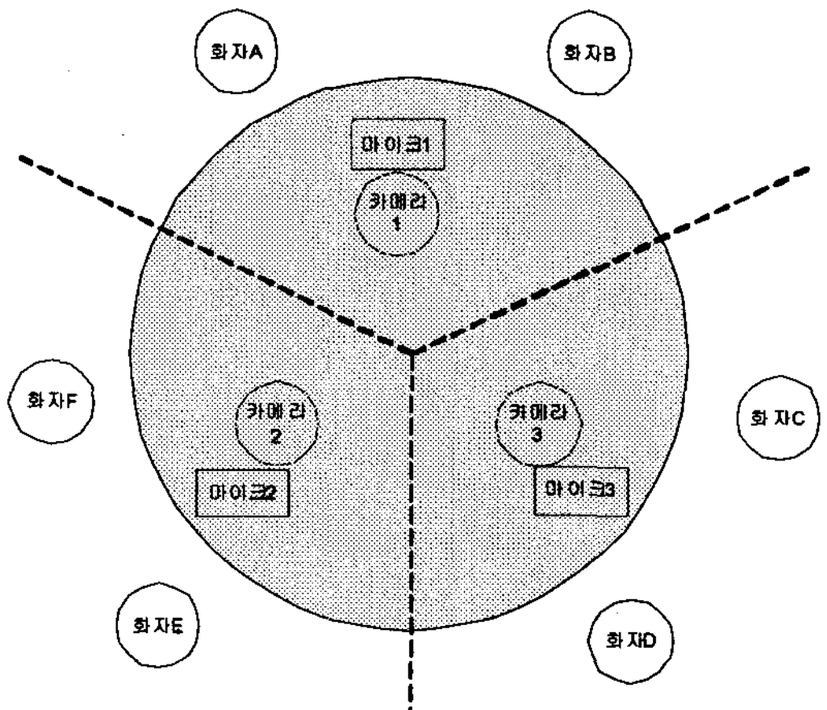

그림 2. 화자 위치 분석 시스템의 구성 및 사용 환경

2.3 화자 정보 분석 기법

회의 지원 시스템에 포함된 얼굴 인식 및 화자 인식 기능은 이후 회의록을 열람할 때 열람자의 집중도를 높여주거나 시스템의 품질을 높이기 위해서 사용되는 기능으로, 화자를 중심으로 기록을 할 경우 회의록을 열람하는 사용자는 현재 회의의 흐름을 보다 자세하게 알 수 있다.

영상 정보 내에서 화자를 검출하는 방법은 주로 입술의 움직임 정보를 이용한다. 입술의 움직임은 코, 턱선 같은 특징 값을 이용해서 먼저 입술의 영역을 찾고, 입술 영역으로 추정되는 부위의 변화도를 이용해 화자 여부를 판단한다. 여기서 입술 영역의 변화도를 알기 위해서는 여러 장의 연속적인 얼굴 영상이 필요하다.

3. 화자 인식 기법

3.1 가정

화자를 인식하기 위해서도 한 장의 사진만으로는 현재 말을 하고 있는 것인지 아닌지 알 수 없기 때문에, 일정시간 동안 연속적으로 찍힌 여러 장의 영상이 있어야 화자 여부를 판단할 수 있다. 따라서 제안하는 방법에서도 연속된 여러 장의 영상이 주어지는 것을 전제로 한다.

두 번째로 문제를 간단히 하기 위해 한 영상 내에서는 화자가 한명만 존재하고, 눈, 코, 입 등의 특징정보를 판단할 수 있는 상태에서만 화자를 검출하는 것으로 가정하였다. 마지막으로, 하품을 하거나 껌을 씹는 등의 발언 행위 이외의 입 움직임은 없다고 가정하였다.

3.2 화자 인식 기법

제안하는 방법은 입 부위에서 엣지를 이루는 점들의 산포도를 구하고, 그 변화도를 이용해 화자를 구분하는 방법이다. 이것을 수행하기 위한 과정은 다음과 같다.

1) 얼굴 영역의 엣지 추출
2) 얼굴 영역 내에서의 입 부위 설정
3) 입 부위 내부의 엣지 산포도 계산
4) 연속된 영상들의 산포도의 변화도 측정
5) 화자 검출

3.3 얼굴 영역의 엣지 추출 및 입 부위 설정

얼굴 영역에서 엣지를 추출하는 이유는 얼굴 영상을 단순화시켜 필요 없는 정보를 줄이고, 화자 검출을 위한 계산을 단순화시키기 위함이다.

엣지를 추출하는 방법은 여러 가지를 사용할 수 있으며, 본 논문에서는 Canny 알고리즘을 사용하여 엣지를 추출하였다.

엣지를 추출하고 나면 얼굴 영상에서 윤곽선만이 남게 되는데 이 영상에서 입 부위를 설정

하여야 한다. 입 부위를 설정하는 방법 역시 여러 가지가 있을 수 있으나, 본 논문에서는 코의 중심점을 찾고 그 주위의 일정 영역을 입 부위로 설정하였다. 코의 중심점은 화자 여부 판단 시 산포도를 따지기 때문에 중심점의 위치가 반드시 정확할 필요는 없으며, 대략의 비슷한 위치를 중심점으로 설정해도 된다.

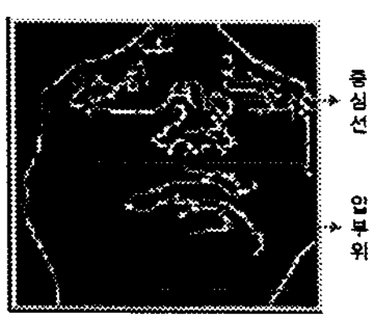

그림 3. 화자 인식을 위한 입 부위 영역 설정

3.4 입 부위의 엣지 산포도 계산

입 부위가 결정되면 그 안에 있는 엣지들의 산포도를 계산해야한다. 산포도를 계산하는 이유는 사람이 입을 벌리거나 닫을 때, 윤곽선의 분포가 달라지기 때문이다. 즉, 엣지들의 산포도를 계산하면 현재 영상이 입을 얼마나 벌리고 있는 상태인지를 수치적으로 계산할 수 있게 되는 것이다.

그림 4. 입을 벌린 상태

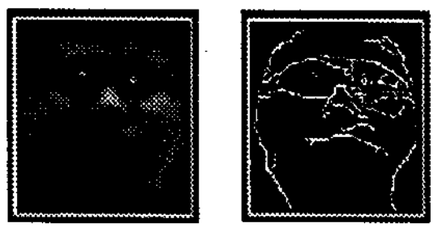

그림 5. 입을 닫은 상태

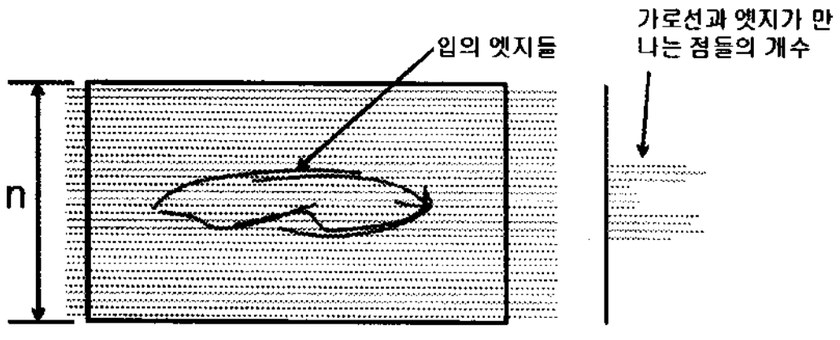

그림 6 엣지들의 산포도 계산

엣지들의 산포도는, 가로선을 기준으로 각 선과 엣지들이 만나는 점의 개수의 합으로 결정한다.

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n}(i-m)^2 f_i \qquad (1)$$

식 (1)은 통계에서 분산을 구하는 공식과 동일한 것으로 n은 입 영역의 길이, i는 각 가로선의 인덱스, f_i는 i번째 가로선이 엣지들과 만나는 점들의 개수, m은 평균값으로 식 (2)와 같다.

$$m = \frac{1}{n}\sum_{i=0}^{n} i f_i \qquad (2)$$

3.5 연속된 영상들의 산포도의 변화도 측정

산포도의 변화도를 측정하는 것은 연속된 여러 장의 영상들에서 입 움직임의 정도를 알기 위한 것이다. 산포도의 변화도를 구하면 현재 영상의 화자 여부를 구분할 수 있다.

각 영상별로 입 부위의 엣지 산포도를 구한 후 그것들의 변화도를 구하면, 연속된 영상들에서 입이 얼마나 움직였는지를 수치적으로 알 수 있다. 이 변화도가 클수록 입의 움직임이 자주 있었다는 것을 의미하므로 그 대상이 화자가 된다. 화자를 판단하는 구체적인 공식은 식 (3)과 같다.

$$\sum_{i=0}^{n}(\sigma_{i+1}^2 - \sigma_i^2) > T \qquad (3)$$

4. 실험 및 결과 분석

화자 인식 기법을 실험해보기 위해 카메라는 팬/틸트/줌이 가능한 Logitech QuickCam Sphere MP, 소프트웨어는 윈도우 환경에서 MFC와 OpenCV 라이브러리를 사용하였다.

화자 인식 기법은 $100\times100(pixel^2)$ 로 정규화된 얼굴 영상을 대상으로 진행했으며, 대상이 화자 상태인 경우와 말을 하지 않고 있는 비화자 상태인 경우를 따로 구분하여 확인했다.

테스트 영상은 4명의 대상자가 신문기사를 읽고 있는 영상을 촬영해 그 영상을 대상으로 하였으며, 이 영상에서 0.4초당 한 프레임씩을 추출해 총 10개의 프레임으로 한 회씩의 실험을 진행했다.

실험 결과 상태별 인식률은 표 1과 같이 나타났다. 화자 상태와 같이 입 부위의 변화가 큰 경우, 높은 확률로 인식해내는 것을 알 수 있다. 반면, 화자 상태에 비해 비화자 상태의 인식률은 비교적 낮게 나타났는데, 이것은 대상의 입이 무의식적으로 조금씩 움직이는 것에도 반응했기 때문으로 보인다.

표 1. 상태별 인식률.

실제 상태 \ 인식 상태	화자	비화자
화자 상태	90%	10%
비화자 상태	17%	83%

실험 후, 정보를 종합 분석해본 결과, 화자를 인식하는 데에는 충분한 성능을 보였으나, 화자가 아닌 사람을 화자로 인식하는 점에 대해서는 보완이 필요할 것으로 보인다.

5. 결론

본 논문에서는 오프라인 회의 기록을 위한 시스템에서 화자를 인식하는 방법으로, 입주위의 엣지 산포도를 조사해 그 변화도로 화자 여부를 판단하는 방법을 제안했다. 실험 결과 화자는 약 90%의 인식률로 찾아냈으며, 비화자는 약 83%의 인식률로 찾아냈다.

화자 상태에 비해 비화자 상태를 인식하는 것은 비교적 낮은 인식률을 보였는데, 이것은 조명 같은 주변 환경의 미묘한 변화와 대상의 의미 없는 입 움직임에 매우 민감하다는 것을 의미하며, 향후 연구에서는 이 부분에 대한 개선이 필요하다.

또한, 입 움직임을 좀 더 상세하게 분석하여 말을 하는 경우와 하품 같은 순간적인 입의 움직임을 구별할 수 있는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] 이병선, 고성원, and 권혁봉, "영상회의를 위한 화자 검출 시스템," 조명•전기설비학회 논문지, Vol. 17, No. 5, pp. 68-79, 2003.

[2] Canny, J., "A computational approach to edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, pp. 679-698, 1986.

[3] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," IEEE Proc. Computer Vision and Pattern Recognition, pp. 586-591, 1991.

[4] Delmas, P., P.Y. Coulon, and V. Fristot, "Automatic snakes for robust lip boundaries extraction," Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on, Vol. 6, No. pp. 3069-3072, 1999.

[5] Schneiderman, H. and T. Kanade, "Object Detection Using the Statistics of Parts," International Journal of Computer Vision, Vol. 56, No. 3, pp. 151-177, 2004.

[6] Hjelmas, E. and B.K. Low, "Face detection: A survey," Computer Vision and Image Understanding, Vol. 83, No. 3, pp. 236-274, 2001.