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

얼굴 인식 및 화자 정보를 이용한 오프라인 회의 기록 지원 시스템

Recording Support System for Off-Line Conference using Face and Speaker Recognition

손윤식, 정진우, 박한무, 계승철[1], 윤종혁, 정낙천, 오세만

동국대학교 컴퓨터공학과

E-mail: {sonbug, jwjung, lilees00, ronaldo, op1000, smoh}@dongguk.edu

[1] E-mail: ksch75@naver.com

요 약

최근 멀티미디어 서비스는 동영상 압축 기술 및 네트워크의 발달을 기반으로 하여 다양한 응용 서비스를 제공하고 있으며, 이 중 화상 회의 시스템은 이 두 가지 기술이 효과적으로 사용된 대표적인 예이다. 원격 사용자간의 원활한 의사전달을 위해 고려된 화상회의 시스템은 효과적인 응용 서비스로 분류되고 있지만, 이러한 서비스 제공을 위한 기술을 이용하여 빈도가 훨씬 많은 일반적인 회의를 지원하는 응용서비스는 드문 편이다.

본 논문에서는 얼굴 정보와 화자 정보를 기반으로 오프라인 회의를 보조하는 시스템을 제안한다. 제안된 시스템은 소규모의 마이크와 캠을 이용하여 화자의 위치를 파악하고 캠에서 얻어진 정보를 이용하여 얼굴 영역 정보를 분석하고 인식한 후 화자 정보를 추출하여 발언자들을 추적하여 기록하는 기능을 제공한다.

Key Words : Conference recording, Face recognition, Speaker recognition

1. 서 론

원격 사용자간의 효과적인 의사전달을 위해 고려된 화상회의 시스템은 동영상 압축 기술 및 네트워크의 발달을 통해 그 효과가 입증되고 있으며 사용 빈도가 더욱 늘어나고 있다. 특히 최근에는 이러한 화상 회의 시스템에 얼굴 인식과 화자 분석과 같은 다양한 기법이 결합되고 있는 추세이다. 반면 실제 사회에서 발생하는 대다수의 일반적인 회의를 위한 보조 시스템은 드문 편이기 때문에 이를 지원하는 시스템에 대한 고려가 필요하다.

본 논문에서 제안된 회의 기록 지원 시스템은 회의 중에 발생하는 영상과 대화를 기록하여, 향후 의사 결정이나 업무 진행시 자료로 사용 가능하도록 하는 것을 목표로 한다. 영상의 기록과정에는 얼굴 인식 정보와 화자 정보를 기반으로 회의 기록을 보다 사용자 중심으로 할 수 있도록 기능을 구현하였다. 특히 얼굴 인식에서 사용되는 특징 값을 얼굴의 회전에 강인하도록 추출하여 자유로운 상황의 회의에서 다양한 자세를 가지는 참석자를 인식하도록 하였다.

2. 관련 연구

2.1 얼굴 인식 기법

얼굴 인식 기술은 10여년에 걸쳐 많이 연구되고 있는 분야로 주로 얼굴 영역을 인식하는 과정을 다루고 있으며, 크게 피부색, 영역 분할, 얼굴 특징 값을 이용한 세 가지 방법론으로 분류할 수 있다.

피부색으로 얼굴을 인식하는 방법[1-3]은 4가지 과정을 통해 수행되는데, 1) 입력된 영상을 몇 개의 색 계층으로 나눈 후 2) 위치정보에 따라서 색 정보를 모으고, 3) 미리 정해진 얼굴 모델을 기준으로 얼굴의 위치를 찾고, 4) 검출된 위치를 토대로 얼굴 영역 검증을 한 번 더 수행한 후 결과를 출력하게 된다. 피부색을 이용한 이 방법론은 비교적 간단한 알고리즘으로 얼굴을 신속히 검출할 수 있는 장점이 있는 반면 얼굴 검출 성능이 다른 방법론에 비하여 떨어지고 색 정보만을 가지고 얼굴을 검출하므

로 복잡한 배경에서는 오류율이 많다.

영역 분할을 이용한 방식은 높은 얼굴 검출 성능을 자랑하는 방식이다[4-6]. 영역 분할, 후보 얼굴 영역 검출, 최종 얼굴 영역 검출의 세 가지 과정을 거치는데, 이 방법론에서 복잡도가 가장 높은 부분은 영역 분할 과정이다. 영역 분할 과정은 영상의 색 정보를 양자화한 후 그 결과를 이용해 클래스 맵과 j-영상을 생성하면서 그룹 정보를 세밀화하고, 그룹의 중심점을 찾는다. 그리고 각 영역이 서로 만날 때까지 영역을 확장한 후, 찾아낸 영역들을 다시 합쳐서 최종적으로 영역이 분할된 영상을 얻는다. 이 기법은 다른 기법들에 비해서 가장 높은 얼굴 검출률을 보인다. 그렇지만 영역 분할에서 많은 연산을 수행하기 때문에 실시간 얼굴 검출 기법으로는 부적절하다.

특징 값을 이용한 얼굴 영역 검출 방법론은 주로 학습 기법과 함께 사용되는데 얼굴뿐만 아니라 특정 물체를 계속 학습시켜 그 영역을 검출하는 방식으로 동영상에서 얼굴 검출 시 많이 사용되는 방식이다[7-9]. 이 방법론은 다음과 같은 실행 과정을 가진다. 1) Adaboost 알고리즘을 이용하여 Haar-like 특징 값들을 추출한 후, 2) 이 중 우수한 특징들을 사용하여 cascade 방식으로 분류기를 구성하고 학습하며, 3) 얼굴 영역과 배경 영역을 분류 후 검출로 끝난다. 학습 데이터로는 주로 얼굴 정면, 오른쪽 측면, 왼쪽 측면을 사용하여 얼굴 각도에 대해 강인한 검출이 가능하다. 이 방법론의 장점은 빠른 검출능력을 자랑하며, 색 정보를 이용한 방법보다 인식률이 높은 장점을 지닌다.

2.2 화자 정보 분석 기법

영상 회의 시스템에서 화자 정보는 사용자의 집중도를 높이거나 시스템의 품질을 향상시킬 수 있는 중요한 정보로 사용된다. 영상 정보를 기반으로 화자를 분석하는 기법은 주로 입술의 움직임 정보를 이용한다[10, 11]. 따라서 다수의 영상 프레임 정보가 필요하며, 각 프레임별로 영상에 존재하는 다수의 얼굴을 인식한 후 각 영상에서 눈이나 턱 선과 같은 입술의 위치를 추정하기 위한 특징 값을 추출하고, 이를 이용하여 여러 프레임 상에서 발생하는 추정 입술 영역의 변화도를 이용하여 화자를 결정한다.

3. 회의 지원 시스템

3.1 시스템 모델

제안하는 회의 지원 시스템은 오프라인 회의 과정을 기록하면서 회의의 참석자를 인식하고 화자를 분석하여 영상 기록시 화자 중심의 기록을 수행한다. 이와 같은 과정은 그림 1과 같이 회의 영상이 기록되는 과정 중에 4가지 분석 작업이 반복적으로 수행되면서 이루어진다.

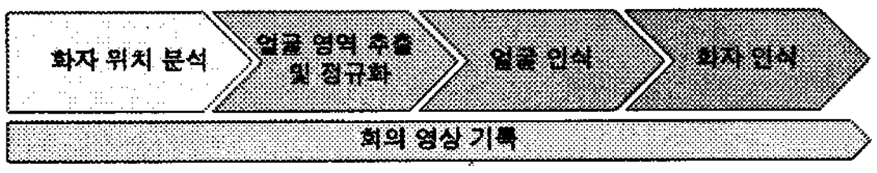

그림 1. 회의 지원 시스템의 수행 과정

캠과 마이크의 수를 최소화 하면서 화자의 위치를 파악하기 위하여 그림 2와 같은 시스템 구성과 사용 환경을 고려하였다.

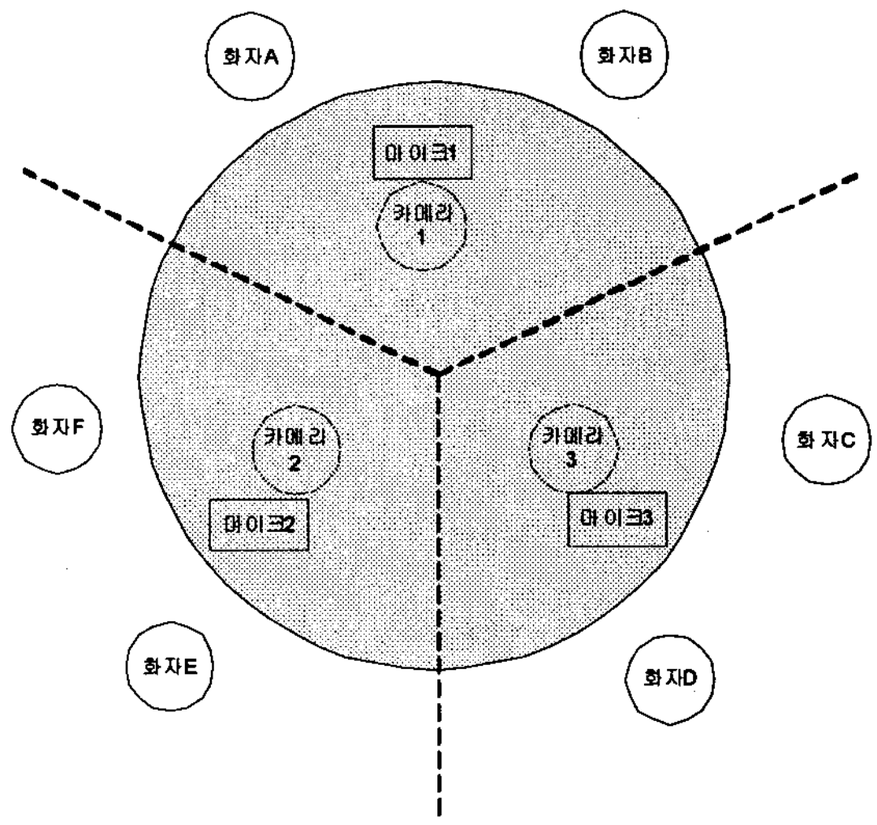

그림 2. 시스템 구성 및 사용 환경

원탁 테이블에 화자들이 위치하며 각각 3대의 카메라와 마이크가 원탁 테이블의 중앙에 위치하고 각 카메라는 팬 기능을 이용하여 전방 120도의 시야를 담당하여 동작한다.

3.2 화자 위치 분석

회의 중의 화자를 녹화하는 시스템에서 다수의 카메라를 사용할 때, 가장 먼저 고려해야 할 사항은 화자가 위치한 곳의 카메라를 찾아서 해당 카메라에서 얼굴 인식을 시작해야 한다는 것이다. 제안한 시스템은 마이크에서 입력되는 음성 신호를 기반으로 화자의 위치를 추정하고 해당 위치를 담당하는 캠은 추정된 화자의 위치로 회전시킨다.

마이크를 통한 위치 확인은 거리에 따라 음성 신호 크기가 줄어드는 원리를 이용한다. 일례로, 그림 2에서 화자D가 말을 하면 그림 3과 같이 음성 신호 크기가 가장 크게 측정되는 곳은 3번 마이크이고 두 번째로 큰 곳은 2번 마

이크이며 1번 마이크는 가장 낮은 신호 크기를 갖게 된다. 일정 시간 동안 연속적으로 신호크기의 평균치가 임계값 이상 되면 화자가 말을 하고 있다고 판단하며 이때 해당 시간 동안 평균 신호크기가 가장 큰 마이크 번호를 선택하여 카메라 동작 모듈로 알려주고 신호 크기의 차를 이용하여 화자의 위치를 추정한다.

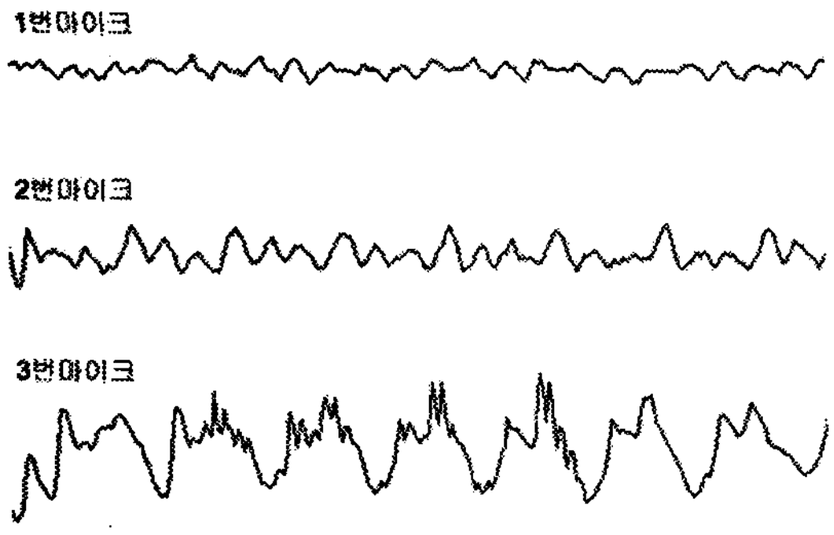

그림 3. 마이크로 입력된 음성신호

3.3 얼굴 영역 추출 및 정규화

추정된 위치의 캠으로부터 얻어진 영상에서 얼굴 영역을 추출하기 위해 먼저 영상 전체에 대한 히스토그램 평활화를 수행한 후, Harr-like 특징 값을 이용하여 영상에 존재하는 다수의 얼굴 영역을 검출한다. 얼굴 영역은 얼굴 색 정보에 기반하여 주변의 배경을 삭제한 후 얼굴 객체만 추출하여 정규화 작업을 수행한다.

정규화는 얼굴 객체를 타원이라고 가정한 후 무게 중심과 장축을 구하여 이를 기준으로 얼굴이 기울어지지 않도록 보정하며 양선형 보간법을 이용하여 다양한 크기의 얼굴 객체를 일정한 크기로 확대/축소한다. 그림4는 정규화에서 이루어지는 과정을 도시한 것이다.

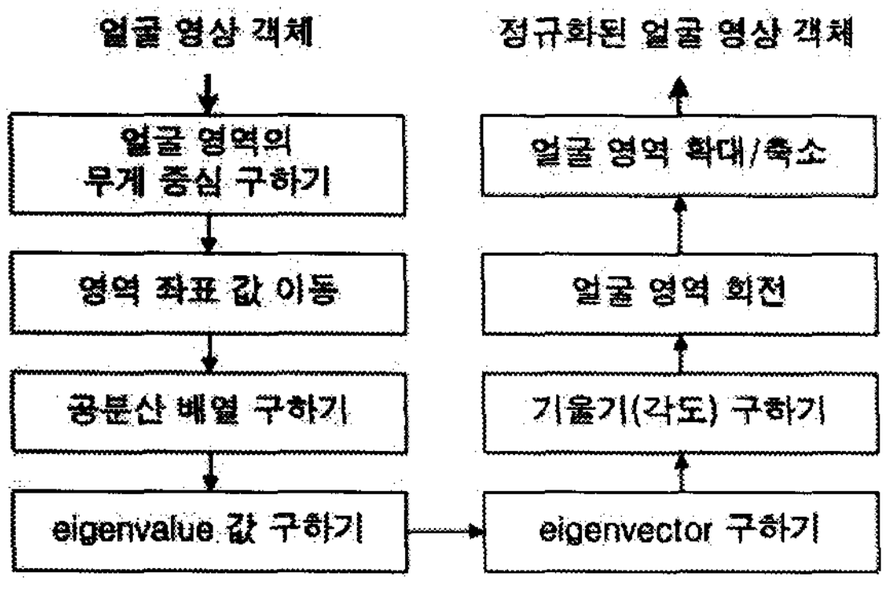

그림 4. 얼굴 객체 정규화 과정

3.4 얼굴 인식

얼굴 인식 과정은 정규화된 얼굴 객체의 특징 값을 이용하여 신원을 확인하는 작업이다. 이를 위해 양쪽 눈의 중점과 코의 중점, 윗입술의 위치를 특징 값으로 사용하며, 이미지의 크기나 얼굴의 상하, 좌우 회전을 고려하여 각 특징 값 사이의 비율을 비교 대상으로 하였다.

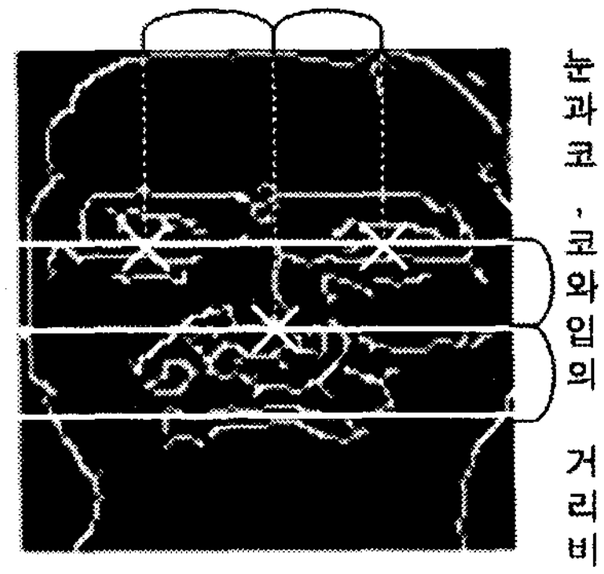

그림 5. 얼굴 인식을 위한 특징 값

사용되는 특징 값은 좌우 회전의 경우 상하 회전에 비해 변화가 크기 때문에 특징 값 추출 시 얼굴의 좌우 회전에 대한 고려가 필요하며, 그 과정은 다음과 같다. 먼저 정규화된 얼굴 영역 객체에 Canny 기법[13]을 사용하여 엣지를 추출하고 이를 기반으로 눈썹, 눈, 코, 입의 수평 특징 선을 찾아낸다. 이때 얼굴의 좌우 회전을 고려하여 얼굴 영역을 좌측, 우측의 두 부분으로 분리하여 엣지를 수집하고, 중간 영역에 대해 중복 수행하여 수집한 엣지를 통합하고 대표 값을 선택한다. 선택된 대표 값을 이용하여 해당 위치에 분포하는 픽셀의 빈도수를 이용하여 양 쪽 눈과 코의 대표 값을 찾아낸다.

얼굴의 좌우 회전에 대한 판단은 코의 대표 값을 이용한다. 정면을 향하고 있는 얼굴 영상의 경우 코의 위치가 전체 얼굴 영역 중 가로의 중점에 위치한다고 가정하고, 코 대표 값이 좌우로 치우침에 따라 얼굴이 회전하였다고 판단하여 이에 대한 보정 작업을 한다. 한 쪽으로 치우쳐진 코의 대표 값을 중앙으로 이동시키기 위한 각도를 계산하고, 이 값을 이용하여 양쪽 눈의 대표 값을 변환한다. 이는 얼굴상에서 코와 눈의 특징 점이 동일한 원주에 있다는 가정에 기반한 것이며 실제 캠과 사람 사이의 거리가 1m라고 고려했을 때 2~3%정도의 오차가 생기는 것으로 확인되었다.

얼굴 인식과정은 위의 과정을 미리 수행하여 얻어진 값과 실시간 영상에서 얻어지는 값의 비교를 통해 수행되며, 인식의 정확도를 높이기 위해 4프레임 이상의 영상에서 얻어진 특징 값의 평균을 사용한다.

3.5 화자 인식

얼굴 인식 과정 후 화자 인식을 수행하며, 얼굴 인식 과정에서 얻어진 특징값과 얼굴 객체를 사용한다. 화자 인식 문제의 특성상 시간적으로 연속적인 여러 개의 프레임이 필요하며 판단에 사용하는 초당 프레임의 개수가 일정량 이상 보장이 되어야 인식이 가능하고, 동시에 말하는 사람이 있을 경우 그 판단이 주화자가 누군지 판단하기 어렵기 때문에 한 시점에서는 화자가 한 명 뿐이라는 것을 가정하였다.

인식 과정은 먼저 코 아래 입술 부위의 엣지들의 분산을 구하고, 그것들의 변화율을 측정한다. 이 변화율이 일정 시간 동안 가장 큰 얼굴 객체를 화자라고 인식하며, 연속된 프레임 상에서 입주위의 변화도가 너무 적은 얼굴 객체는 화자가 아니라고 판단하여 화자인식 대상에서 제외시킨다.

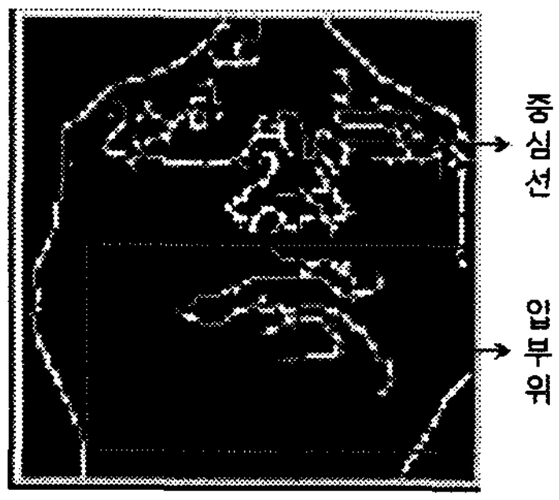

그림 6. 화자 인식을 위한 입 부위 영역 설정

화자가 인식되면 영상에서 화자를 정 중앙에 위치시키기 위해 캠의 팬/틸트/줌 모듈을 사용하여 위치를 조정한다.

4. 실험 및 결과 분석

시스템을 구현하기 최대 960x720 해상도와 팬/틸트/줌이 가능한 로지텍 QuickCam Sphere MP 3대와 USB 마이크 3개를 사용하였으며, 윈도우 환경에서 MFC와 OpenCV 라이브러리를 사용하여 구현하였다.

얼굴 인식을 위한 기본 정보로 15명의 얼굴 정보를 사전에 입력하였으며, 인원수에 따른 성능 평가를 위해 얼굴 인식 대상 데이터의 수에 변화를 주어 실험하였다. 실험상의 회의에는 5명이 참가하여 원탁 테이블 상에서 일반적인 형광등 조명하에 수행되었다. 회의의 녹화 과정은 MFC를 이용하여 이루어 졌으며, 본 시스템이 가지는 얼굴 인식과 화자 인식 과정을 확인하기 위해 각 과정에 대한 결과 분석을 수행하였다.

4.1 얼굴 인식 결과

얼굴 인식은 100x100 픽셀로 정규화 된 얼굴 객체 이미지를 대상으로 수행되었으며, 기본 데이터로 상하 20도, 40도, 좌우 25도 50도의 기울기를 가진 영상에서 얻어진 평균 특징값을 사용하여 회의 시 얼굴의 상하좌우 회전에 인식 과정이 얼마나 강인한지 확인 하였다. 또한 기본 데이터에 따라 인식률을 확인해보기 위해 기본데이터의 수에 변화를 주었다. 실험 결과, 표 1과 같은 인식률이 도출되었으며, 얼굴 인식에 사용된 특징 값은 좌우 회전 보다 상하 회전에 더욱 강인한 것으로 나타났다. 얼굴의 좌우 회전에 따른 인식률 저하는 그림 7과 같이 회전 보정에 의한 오차 때문인 것으로 판단되며, 회전 각도가 증가할수록 인식률이 떨어지는 것으로 확인되었다. 인식을 위한 대상 얼굴 객체 정보의 증가에 따라 인식률이 저하되는 것으로 나타났으며, 10명에서 15명으로 증가했을 때의 저하도가 낮은 것은 추가된 대상 데이터가 기존의 데이터와 차이가 크기 때문으로 확인되었다. 실험 후 전체 인식률을 종합해본 결과 소규모 회의에 참가하는 인원수의 범위에서는 충분한 인식률을 보인다고 판단된다.

표 1. 얼굴 회전 및 인식 대상 데이터 수에 따른 얼굴 인식률

회전 방향 / 데이터 수	정면	상하 회전		좌우 회전	
		20도	40도	25도	50도
5명	98%	96%	95%	95%	92%
10명	94%	92%	91%	90%	85%
15명	92%	91%	91%	89%	84%

그림 7. 좌우회전에 따른 특징 값의 변화

4.2 화자 인식 결과

얼굴 인식과 동일한 정규화된 얼굴 객체 이

미지를 입력으로 하였으며, 연속된 프레임 상에서 발생하는 입 영역의 움직임을 확인하기 위해 일정 시간동안의 프레임을 얻어서 실험을 진행하였다. 실험 결과 표 2와 같은 인식결과가 나타났다. 객체가 말을 하고 있지 않은 상태인 대기 상태와 말을 하고 있는 화자 상태가 확연히 구분이 되었으며, 기준 값을 각 상태의 사이 값으로 정할 경우 높은 인식률을 보였다.

표 2. 상태별 인식률

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

5. 결 론

본 논문에서는 얼굴 인식과 화자 인식 기법을 적용하여 일반적인 오프라인 회의 환경에서 사용 가능한 회의 기록 지원 시스템을 제안하였다. 사용된 얼굴 인식 기법은 얼굴의 특징 값을 이용하여 신원을 확인하는 방법으로 사용된 특징 값은 얼굴의 상하 좌우 회전을 고려하여 결정하였으며, 실험 결과 얼굴이 회전한 영상에서도 약 92%정도의 인식률을 보였다. 화자 인식 기법은 얼굴 객체에서 입 영역의 변화도를 기반으로 화자를 추출하는 방식이며 87%의 정확도를 보였다.

이러한 얼굴과 화자의 인식 정보를 토대로 회의 기록 지원 시스템에서 영상을 사용자가 보다 사용하고 집중하기 수월한 형태로 기록이 가능함을 보였으며, 향후 과제로는 얼굴 인식을 위한 특징 값과 추출과정을 개선하여 보다 정확하고 빠른 얼굴 인식에 대한 연구와 화자의 입 움직임에 대한 세분화를 통해 말을 하는 경우와 순간적인 입 움직임을 구별하여 더욱 자연스러운 회의 과정에서의 화자 인식에 대한 연구가 필요하다.

참 고 문 헌

[1] Chai, D. and K.N. Ngan, "Face segmentation using skin-color map in videophone applications," Circuits and Systems for Video Technology, IEEE Transactions on, Vol. 9, No. 4, pp. 551-564, 1999.

[2] Wang, Y. and B. Yuan, "A novel approach for human face detection from color images under complex background," Pattern Recognition, Vol. 34, No. 10, pp. 1983-1992, 2001.

[3] Hsieh, I.S., K.C. Fan, and C. Lin, "A statistic approach to the detection of human faces in color nature scene," Pattern Recognition, Vol. 35, No. 7, pp. 1583-1596, 2002.

[4] 이재원, et al., "복잡한 영상에서의 영역 분할을 이용한 얼굴 검출," 멀티미디어학회논문지, Vol. 9, No. 2, pp. 160-171, 2006.

[5] Deng, Y. and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 8, pp. 800-810, 2001.

[6] Deng, Y., et al., "Peer group filtering and perceptual color image quantization," Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on, Vol. 4, No. pp. 21-24, 1999.

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

[8] Lienhart, R. and J. Maydt, "An extended set of Haar-like features for rapid object detection," Image Processing. 2002. Proceedings. 2002 International Conference on, Vol. 1, No. pp. 900-903, 2002.

[9] Viola, P. and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proc. CVPR, Vol. 1, No. pp. 511-518, 2001.

[10] 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.

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

[12] 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.

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