바이모달 음성인식기의 시각 특징 추출을 위한 색상 분석과 SVM을 이용한 입술 위치 검출

(Lip Detection using Color Distribution and Support Vector Machine for Visual Feature Extraction of Bimodal Speech Recognition System)

정 지 년 † 양 현 승 ††
(Jinyun Chung) (Hyun Seung Yang)

요 약 바이모달 음성인식기는 잡음 환경하 음성인식 성능을 향상하기 위해 고안되었다. 바이모달 음성인식기에 있어 영상을 통한 시각 특징 추출은 매우 중요한 역할을 하며 이를 위한 입술 위치 검출은 시각 특징 추출을 위한 중요한 선결 과제이다. 본 논문은 색상분포와 SVM을 이용하여 시각 특징 추출을 위한 입술 위치 검출 방법을 제안하였다. 제안된 방법은 얼굴색/입술 색상 분포를 학습하여 이로부터 입술의 초기 위치를 빠르게 찾아내고 SVM을 이용하여 입술의 정확한 위치를 찾음으로써 정확하고 빠르게 입술의 위치를 찾도록 하였으며 실험을 통해 바이모달 인식기에 적용하기에 적합함을 알 수 있었다.

키워드 : 바이모달 음성인식, 입술 위치 검출, 입술 추적, 입술 인식, Support Vector Machine, 색상 분포

Abstract Bimodal speech recognition systems have been proposed for enhancing recognition rate of ASR under noisy environments. Visual feature extraction is very important to develop these systems. To extract visual features, it is necessary to detect exact lip position. This paper proposed the method that detects a lip position using color similarity model and SVM. Face/Lip color distribution is learned and the initial lip position is found by using that. The exact lip position is detected by scanning neighbor area with SVM. By experiments, it is shown that this method detects lip position exactly and fast.

Key words : bimodal speech recognition, lip detection, lip tracking, lip reading, Support Vector Machine, color distribution

1. 서 론

오늘날 음성인식 기술은 많은 발전을 이루었으며, 저잡음 환경에서 놀라운 인식 성능을 보이고 있다. 이러한 기술 축적을 바탕으로 이제 음성인식 연구자들은 잡음이 심한 환경에서의 음성인식을 연구하고 있다. 이 과정에서 연구자들은 시청각 신호를 함께 이용하여 음성인식을 수행하고자 하는 바이모달 음성인식 방법을 제안하였고, 음성과 시각 신호를 같이 사용할 경우 음성 인식 성능이 향상됨을 보여 왔다.

바이모달 음성인식을 수행함에 있어 기존의 음성 인식과 다른 점은 음성인식과 함께 융합할 시각 정보를 추출해야 한다는 점과 시각 정보와 음성 정보를 효과적으로 융합할 수 있는 방법을 개발해야 한다는 점이다. 이중 음성과 밀접한 시각 특징을 추출하는 과정은 음성인식 성능의 향상을 위해 매우 중요한 단계라고 할 수 있다. 음성과 밀접한 시각 특징은 대부분 입에 있기 때문에 입 모양에서 음성 정보를 추출하여야 한다.

시각 특징 추출은 크게 두 가지 방법으로 나뉠 수 있다. 하나는 기하학적 모델을 이용하여 정보를 추출하여 그 계수를 특징값으로 사용하는 모델 기반 접근 방법(Model-based approach)이고 다른 하나는 영상 기반 접근 방법(Image-based approach)이다. 전자는 입술 모양을 나타내는 곡선이나 Snakes, Deformable Template과 같은 기하 모형을 이미지에 적용하여 현재 이미지에 가장 잘 맞는 기하 모형의 모양으로 변화시킨 후,

† 비 회 원 : 한국과학기술원 전산학과
mage@paradise.kaist.ac.kr
†† 종신회원 : 한국과학기술원 전산학과 교수
hsyang@cs.kaist.ac.kr
논문접수 : 2003년 7월 15일
심사완료 : 2003년 12월 26일

이 기하 도형의 파라메터들을 시각 특징으로 사용하는 접근 방법이다. 이 방법은 입술 위치 추적과 특징 추출이 동시에 이루어지는 장점이 있으나 최적 모델을 발견하기가 어렵고 제공하는 정보량이 영상 기반 접근 방법에 비해 적은 단점이 있다. 후자는 영상 자체 혹은 영상의 코딩값이나 영상 처리 후의 영상 자체를 시각 특징으로 사용하는 접근 방법이다. KL Transform이나 PCA를 사용하여 입술 영상을 코딩하고 이를 영상 특징으로 사용하거나 영상의 Optical flow와 같은 저차원 특징을 분석하여 이를 시각 특징으로 사용하는 것을 예로 들 수 있다. 이와 같은 영상 기반 접근 방법은 모델 기반 접근 방법에 비해 정보가 풍부하므로 바이모달 음성인식에 더 도움이 되나[1] 모델 기반 접근 방법과는 달리 별도로 입술의 위치를 정확히 찾는 과정이 필요하다. 입술의 위치를 정확히 찾지 못할 경우 추출된 시각 특징의 편차가 커져 인식 성능의 저하를 야기하기 때문이다. 따라서 본 논문에서는 영상 기반 접근 방법이 바이모달 인식기에 더 적합하다고 보고, 이를 위한 선결 과제인 정확한 입술 위치 검출 방법을 고안하고자 하였다.

입술 위치 검출을 위해서 많이 사용하는 방법으로 입술 색상을 분석하여 색상으로 입술에 해당하는 픽셀을 분류하고 이 픽셀이 모여있는 곳을 찾는 접근 방법이 있다. 입술색상을 얼굴 색상과 구분하기 위해 히스토그램에 대해 Fisher's Linear discriminant analysis를 수행하여, 프로젝션시 얼굴색상과 입술색상이 가장 잘 분류 가능한 색상좌표축을 찾아 얼굴과 입술이 뚜렷이 구분되는 영상을 얻어 입술의 윤곽을 찾고자 하기도 하였으며[2], Expectation-Minimization 알고리즘을 사용하여 입술 내부의 색상 및 입술의 색상을 혼합 가우시안 모델로 모델링하여 입술과 입안에 해당하는 픽셀을 찾고자 하였다[2-3]. 또한 Red-Green 채널에서의 입술 색상의 평균과 표준편차를 계산하여 평균색상을 중심으로 일정 영역을 입술색상으로 판단하고 영상을 입술픽셀과 입술외 픽셀로 이진화하여 입술을 찾기도 하였다[4]. RGB에서 Red 채널의 값을 Green 채널의 값으로 나누어 이 값을 각 픽셀의 Fuzzy membership value로 하여 이를 Threshold함으로써 입술 영역을 구하기도 하였다[5]. 제안된 방법들은 입술 색상분포를 모델링하고자 시도하였으며 이를통해 영상에서 입술색상을 가진 픽셀을 구분하고자 하였다. 그러나 영상의 색상은 조명, 사람에 따라 편차가 심하며 경우에 따라서는 입술색과 피부색의 차이가 미비하기 때문에 색상을 아무리 잘 모델링한다고 하더라도 이로부터 정확한 입술을 찾기에는 무리가 있으므로, 제시된 방법들로부터 구한 입술 영역은 정확성이 떨어진다. 따라서 픽셀의 색상만을 의지하여 입술 영역을 정확히 찾기는 무리라고 할 수 있다. 이러한 한계 때문에 B-spline 곡선을 이용하여 입술 윤곽을 계산하거나[6], 특징점을 찾아 이를 Snake로 이은 후에 곡선식으로 근사하는 방법을 사용하거나[7], 엣지맵에 Snake를 적용하여 윤곽을 계산하는 등[8] 입술 모델을 만들어 영상과 입술 모델의 정합도를 계산하여 입술의 위치를 찾고자 하는 시도들이 있었다. 그러나 이런 접근 방법들은 혀나 이빨이 등장하거나 사라지는 등의 이유로 인해 최적의 입술 모델을 만드는 일이 어려우며, 입술의 모양 변화에 따른 입술 모델의 파라메터 변화를 조정하는 일도 쉽지 않아 정확한 입술 검출에 어려움이 많은 것이 사실이다.

따라서 이러한 어려움을 해결하는 방법으로 별도의 모델을 구축하지 않고 입술 영상 자체를 우수한 패턴 분류기에 학습시킴으로써 입술의 위치를 찾는 것이 보다 정확할 것으로 판단되었으며, 본 논문에서는 이를 위해 우수한 패턴분류기로 알려진 SVM을 사용하고자 하였다. 즉 먼저 얼굴색 분포를 이용하여 얼굴을 찾은 후 얼굴을 스캔해 가며 SVM을 적용하여 입술에 해당하는 위치를 찾는다. 그러나 SVM은 계산량이 많아 검색 범위가 크면 실시간 시스템 구현에 있어 불리하기 때문에, 검색 범위를 줄이기 위해 입술 색상 분포를 구하여 이를 바탕으로 입술의 대략적인 위치를 추정하고 이로부터 입술의 위치를 찾도록 하였으며 실험 결과 다양한 화자와 수많은 영상에 대해 입술의 위치를 정확하게 찾을 수 있었다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문에서 사용된 얼굴색 및 입술색 분포 학습 방법과 SVM 학습 방법, 그리고 얼굴 위치 검출 과정을 포함한 입술 위치 검출 과정에 대해 기술한다. 3장에서는 실험을 통해 입술 위치 검출 성능을 보이고 4장에서 결론을 내리며 논문을 마무리 짓는다.

2. 입술 위치 검출 방법

제안된 입술 위치 검출 과정은 얼굴 영역 세그먼테이션, 색상을 이용한 입술 위치 추정, SVM을 이용한 입술 위치 조정의 세 단계로 이루어진다. 화면에서 입술을 검출하기 위해서는 먼저 얼굴의 위치를 찾는 것이 유리하다. 입술을 찾기 위한 검색 범위를 얼굴로 한정함으로써 정확성을 높이고 보다 더 빨리 입술의 위치를 찾을 수 있기 때문이다. 한정된 얼굴 영역에서 입술은 얼굴의 하반부에 있으므로 그 영역에 대해 입술 색상을 한 픽셀을 찾아 입술의 대략적인 위치를 정한다. 그러나 찾은 입술의 위치는 정확하지 않기 때문에 입술 영상을 학습한 SVM을 이용하여 주변 영역을 제한적으로 스캔하여 정확한 위치를 찾는다.

각 단계에 대해 자세히 기술하면 다음과 같다.

2.1. 얼굴 영역 분할

얼굴을 찾는 방법에는 여러 가지가 있지만, 여기서는 화면에 한 사람의 얼굴만 있다고 가정을 하고, 간단하게 색상을 이용하여 얼굴 영역을 분할하는 방법을 사용한다.

2.1.1 얼굴 색상 분포 학습

색상을 이용하여 얼굴 영역을 분할하기 위해서는 어떤 색상이 얼굴 영역을 구성하는지에 대한 지식이 필요하다. 경험을 통해 대략적인 얼굴 색상을 정할 수 있으나, 통계적인 관찰 결과를 통해 얼굴색을 학습하는 것이 가장 정확하다. 특정 색상이 얼굴색인지는 얼굴색 유사도로 표시할 수 있으며, 통계적 관찰 결과를 통해 특정 색의 얼굴색 유사도를 결정할 수 있다. 특정 색상 c_1의 영상에서의 관찰 빈도수를 $F_B(c_1)$이라고 하고, c_1의 얼굴 색상으로의 분류 빈도수를 $F_F(c_1)$라고 할 때 c_1의 얼굴 색상 유사도(얼굴 색상 확률) $P_F(c_1)$는 다음과 같이 정할 수 있다.

$$P_F(c_1) = \frac{F_F(c_1)}{F_B(c_1)} \quad (1)$$

즉, 영상 내의 어떤 픽셀의 색상이 c_1일 경우 이 색상이 얼굴색일 확률(얼굴 색상 유사도)은 $P_F(c_1)$이 되는 것이다. 이때, 얼굴색상 유사도는 색상값 c_1를 색인으로 하는 테이블로 이해할 수 있으며, 얼굴 색상 분포의 학습은 이 테이블의 값을 채우는 과정으로 이해할 수 있다. 이때, 얼굴색상 유사도 테이블은 2차원 배열로 구현하였으며, 색인은 HSV(Hue-Saturation-Value)색상 공간에서 Hue/Saturation을 이용하였다. HSV 색상 공간의 Hue/Saturation을 선택한 것은 이 색상 공간이 다른 색상 공간에 비해 입술색을 비교적 잘 분류해 주어 색상을 이용한 입술의 초기 위치 검출에 유용하기 때문이다. 이러한 판단은 영상에서 색상만으로 얼굴 픽셀과 입술 픽셀을 분류하였을 때 잘못 분류되는 픽셀의 비율을 계산해 봄으로써 할 수 있다. 잘못 분류되는 픽셀의 비율은 실험 대상인 얼굴 영상에 대해 구한 채널 c_1과 c_2의 히스토그램 $H_f(c_1,c_2)$에 대해 입술 픽셀의 갯수 $H_{lip}(c_1,c_2)$와 비입술 픽셀의 개수($H_f(c_1,c_2)-H_l(c_1,c_2)$) 중 작은 값의 비율로 계산되며 그 식은 다음과 같다.

$$\frac{\sum \min(H_l(c_1,c_2), H_f(c_1,c_2) - H_l(c_1,c_2))}{\sum H_f(c_1,c_2)} \quad (2)$$

즉, 이는 색상만으로 최선의 선택을 하였을 경우 잘못 분류하는 픽셀의 비율을 계산한 것이다. 표 1은 화장하지 않은 남자 20명에 대해 RGB, YCbCr, HSV, CIE 및 RGB값을 Intensity값으로 나누어 얻은 색상영역({R-G-B}/Intensity), RGB를 Green으로 나누어주어 얻은 색상영역(R/G-G-G/G)의 대해 두개의 채널에 대해(RGB의 경우 R-G, R-B, G-B 세가지 경우) 잘못 분류되는 픽셀의 비율을 계산한 후 그 평균 값을 계산한 것이다. 실험 결과 HSV 색상 영역에서 Hue/Saturation이 잘못 분류하는 픽셀의 비율이 가장 낮음을 알 수 있었으며, 따라서 본 논문에서는 HSV 색상계의 Hue/Saturation을 사용하여 색상분포를 학습하였다.

표 1 각 색상 공간에서 2개 채널의 색상만을 사용하여 얼굴 픽셀과 입술 픽셀을 정하였을 때 영상의 전체 픽셀에서 잘못 분류되는 픽셀의 비율. 1-2는 첫번째 채널과 두번째 채널(RGB의 경우 RG), 2-3은 두번째와 세번째 채널(RGB의 경우 GB), 1-3은 첫번째와 세번째 채널(RGB의 경우 RB)을 사용하여 얼굴과 입술 픽셀을 분류한 경우를 의미한다.

색상 영역 채널 조합	R-G-B	Y-Cb -Cr	H-S-V	{R-G-B} /Intensity	R/G-G -B/G	C I-E (L-a-b)
1-2	0.6452	0.7271	0.5104	0.8801	0.8214	0.6071
2-3	0.6312	0.5524	0.7745	0.7761	0.9425	0.7311
1-3	0.9108	0.7347	0.5545	0.8023	0.9752	0.7643

학습된 얼굴색 분포는 Discrete하며 관찰 경험에 제약을 받기 때문에, 미처 관찰하지 못한 경우에 대해서나, 관찰하였으나 편향된 관찰 경험에 대해서 일반화의 과정이 필요하다. 이러한 일반화는 얼굴색 분포 테이블 $P_F(c_1)$에 대한 Gaussian convolution을 통해 달성될 수 있다. Gaussian convolution을 수행하면 평균화 과정을 통해 관찰하지 못한 얼굴색 유사도 값의 일부를 예측할 수 있으며, 관찰된 결과의 편향성을 완화하게 된다(그림 1). 이상 기술한 얼굴색상 분포 학습 과정은 다음과 같이 요약 기술할 수 있다.

1. 준비된 예제 영상들에 대해 다음을 수행한다.
1.1. 영상에서 얼굴영역을 지정한다.
1.2. 영상전체의 히스토그램 계산($F_B(H,S)$) 및 이전의 $F_B(H,S)$와 누적합산
1.3. 얼굴영역의 히스토그램 계산($F_F(H,S)$) 및 이전의 $F_F(H,S)$와 누적합산
1.4. 준비된 모든 예제 영상에 대해 1.1-1.3. 수행
2. $P_F(H,S)$ 계산

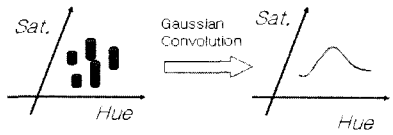


그림 1 Gaussian convolution을 통한 색상 분포 일반화

3. $P_F(H,S)$에 대해 Gaussian convolution 수행
4. 학습종료

사용된 색상 분포 Table의 크기는 128×128이며, Gaussian convolution kernel의 크기는 5×5이었다. 실험을 통해 계산된 얼굴색 분포 테이블은 그림 2에 나타난 바와 같았다. Hue의 값이 70인 부근에 분포가 얼굴색 유사도가 높게 나타나고 있으며, 이 값이 피부색임을 알 수 있다. Saturation은 비교적 편차가 큰 편인데, 이는 조명이나 개인차로 인한 영향으로 해석된다.

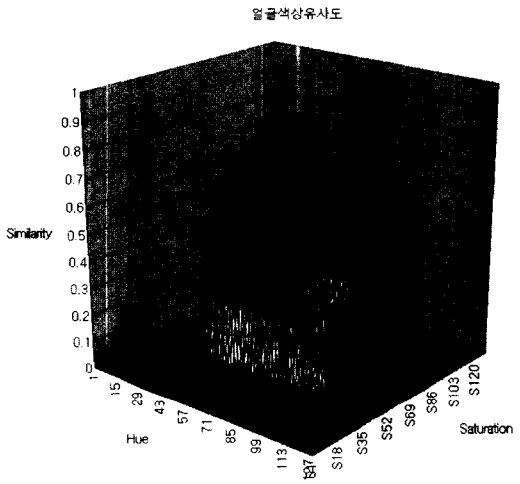


그림 2 얼굴색 유사도 테이블. X축과 Y축은 Hue와 Saturation 인덱스를 나타내며 Z값은 해당 Hue, Saturation에서의 얼굴색 유사도(P_F)를 의미한다.

2.1.2 얼굴색 분포에 의한 얼굴색 유사도 계산

앞에서 얻은 색상 유사도 분포 테이블로부터 영상의 각 픽셀의 Hue/Saturation 값에 해당하는 얼굴색 유사도값을 얻고 이를 이용하여 영상을 재구성한다. 재구성된 영상은 다음 그림 3과 같이 된다.

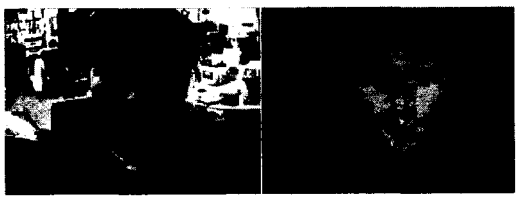


그림 3 얼굴색 유사도 영상

2.1.3 이진화 및 얼굴 영역 분할

재구성한 얼굴 유사도 영상을 이진화 한 후, 이에 대해 모멘트 계산을 통해 얼굴 영역을 타원 모양으로 추정한다[9]. 얼굴 영역을 둘러싸는 타원의 방정식은 다음과 같이 계산된다.

영상의 y방향으로의 이차 모멘트를 μ_{yy}, 2차 혼합 모멘트를 μ_{xy}, x방향으로의 2차 모멘트를 μ_{xx}라고 할 때, 그 값은 다음과 같이 구한다.

$$\mu_{yy} = \frac{1}{A}\sum_{(x,y)\in R}(y-\bar{y})^2$$

$$\mu_{xy} = \frac{1}{A}\sum_{(x,y)\in R}(x-\bar{x})(y-\bar{y})$$

$$\mu_{xx} = \frac{1}{A}\sum_{(x,y)\in R}(x-\bar{x})^2$$

$$\bar{x} = \frac{1}{A}\sum_{(x,y)\in R}x \quad \bar{y} = \frac{1}{A}\sum_{(x,y)\in R}y \tag{3}$$

A는 얼굴 영역으로 되어 있는 픽셀의 개수(면적)를 의미한다. 타원의 영역을 나타내는 부등식을

$$R = \{(x,y) \mid dx^2 + 2exy + fy^2 \le 1\} \tag{4}$$

라고 할 때, 타원의 식을 구성하는 계수 d, e, f는 다음과 같이 계산된다.

$$\begin{pmatrix} d & e \\ e & f \end{pmatrix} = \frac{1}{4(\mu_{xx}\mu_{yy} - \mu_{xy}^{\ 2})}\begin{pmatrix} \mu_{yy} & -\mu_{xy} \\ -\mu_{xy} & \mu_{xx} \end{pmatrix} \tag{5}$$

타원의 식을 구한 후, 타원의 내부의 점을 얼굴 영역으로 삼고 이를 분할하여 입술 위치 검색을 수행하게 된다. 그림 4에서 청록색 선이 얼굴 유사도 영상을 이진화 한 후 분할된 얼굴 영역을 나타내고 있다.

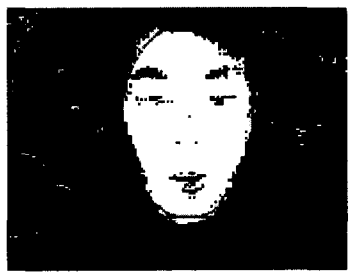


그림 4 타원으로 분할된 얼굴 영역

2.2 입술 색상 분포를 이용한 입술 위치 추정

입술 색상 분포를 이용한 입술 위치 추정 방법은 입술 유사도 영상을 얻는 것까지는 얼굴 영역 분할에서 사용한 방법과 같다. 단 입술 색상 분포를 학습할 때 다소 차이가 있을 뿐이다.

2.2.1 입술 색상분포 학습

얼굴 하반부 색상 c_1의 빈도수가 $F_{lf}(c_1)$이고, 입술 색상의 빈도수가 $F_{lip}(c_1)$일 때 입술 색상 유사도 $P_{lip}(c_1)$는 다음과 같이 계산된다.

$$P_{lip}(c_1) = \frac{F_{lip}(c_1)}{F_{lf}(c_1)} \quad (6)$$

계산된 입술 색상 유사도를 얼굴 색상 유사도와 마찬가지로 테이블로 구성한 후 Gaussian Convolution을 통해 일반화를 시키면 입술 색상 유사도 학습을 마치게 된다. 입술색상 분포 학습 과정은 다음과 같이 요약 기술할 수 있다.

1. 준비된 예제 영상들에 대해 다음을 수행한다.
1.1. 영상에서 입술영역을 지정한다.
1.2. 타원 추정을 통해 얻은 얼굴의 하반부의 히스토그램 계산($F_{lf}(H,S)$) 및 이전의 $F_{lf}(H,S)$와 누적합산
1.3. 입술영역의 히스토그램 계산($F_{lip}(H,S)$) 및 이전의 $F_{lip}(H,S)$와 누적합산
1.4. 준비된 모든 예제 영상에 대해 1.1-1.3. 수행
2. $P_{lip}(H,S)$ 계산
3. $P_{lip}(H,S)$에 대해 Gaussian convolution 수행
4. 학습종료

다음 그림 5는 실험을 통해 계산된 입술색 유사도 테이블을 나타내고 있다. 얼굴색 유사도 테이블에 Hue가 85를 중심으로 얼굴색 유사도 테이블에 비해 좁은 범위로 한정됨을 볼 수 있으며, Saturation은 100 이하의 값을 가짐을 알 수 있다. 이는 입술 색에 있어 Hue는 개인차가 적은 편이나 Saturation의 경우는 사람에 따라 상당한 편차를 보고 있음을 보여주고 있다. 또 얼굴색에 비해 유사도 값이 낮은 편인데, 이는 얼굴색과 입술색의 유사성이 큼을 보여주고 있다.

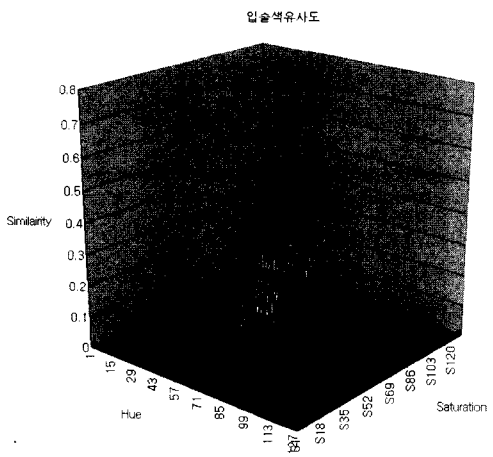


그림 5 입술색 유사도 테이블. X축과 Y축은 Hue와 Saturation 인덱스를 나타내며 Z값은 해당 Hue, Saturation에서의 얼굴색 유사도(P_{lip})를 의미한다.

2.2.2 입술 초기 위치 추정

앞에서 찾은 얼굴 영역에 대해 카메라와의 거리에 의한 영상의 크기 변화를 보상해주기 위해 얼굴의 폭에 대해 표준화를 한다. 그런 후 얼굴 하반부 영상에 대해 학습된 입술 색상 분포를 사용하여 입술 유사도 영상을 구한다. 이 입술 유사도 영상의 질량 중심을 계산하여 이를 입술의 초기 위치로 정한다(그림 6).

유사도 영상의 각점의 입술 유사도 값을 I(x,y)라고 할 때 입술의 초기 위치(C_x, C_y)는 다음과 같이 계산된다.

$$R = \sum_{(x,y)\in I} I(x,y)$$
$$C_x = \frac{\sum_{(x,y)\in I} xI(x,y)}{R} \quad C_y = \frac{\sum_{(x,y)\in I} yI(x,y)}{R} \quad (7)$$

계산된 입술 중심 위치는 일반적으로 정확하지 않으므로 이후 SVM을 이용한 입술 위치 미세 조정 과정을 통해 정확한 입술의 위치를 찾게 된다.

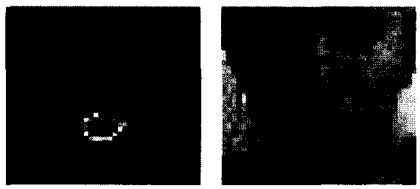


그림 6 입술 색상 분포를 이용한 입술색 유사도 영상 및 질량 중심을 계산하여 입술의 위치를 추정한 결과

2.3 SVM을 이용한 입술 위치 미세 조정

2.3.1 SVM의 학습

SVM은 예제와 반례의 패턴 공간을 구분하는 최적의 hyperplane을 찾고 이를 이용하여 패턴 공간을 구획짓는 패턴 분류기이다[10]. 따라서 학습을 위해서는 입술 영상과 입술 영상이 아닌 것의 예제가 필요한데, 이 예제는 16×16 gray 이미지를 사용하였다. 영상의 차원이 크므로 정보를 가급적 잃지 않고 차원을 줄이기 위해 영상을 2차원 DCT를 이용하여 변환한 후 얻은 계수 중 저주파 계수를 학습 대상으로 삼았다. 저주파 계수는 영상 압축시 사용하는 지그재그 스캔방식으로 선정하였으며, 실험을 통해 잡영 제거 효과와 영상의 주요 정보 보존을 고려하여 46개를 선택하였다. SVM에 사용하는 Kernel은 gaussian kernel을 사용하였으며 그 식은 다음과 같다. γ 값은 실험적으로 0.01로 하였다.

$$K(x,y) = e^{-\gamma|x-y|} \quad (8)$$

2.3.2 SVM을 이용한 입술 위치 미세 조정

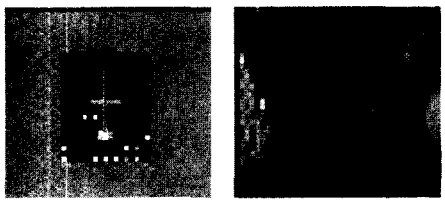
입술 색상 분포를 이용하여 찾은 입술의 위치는 실제 입술의 위치와 비슷하지만 그 정확성이 상당히 떨어진다. 이는 사람에 따라 입술의 색상이 얼굴의 색상과 크

계 다르지 않은 경우가 많고, 조명의 변화에 따른 변화와 영상 자체의 색상의 부정확성(노이즈)에 그 원인을 들 수 있다. 따라서 이러한 이를 보완하기 위해 패턴 분류의 정확성이 높은 SVM 패턴 분류기를 이용하여 입술의 위치를 정확히 찾고자 하였다.

물론 SVM 패턴 분류기만으로 입술의 위치를 찾을 수도 있다. 그러나 그럴 경우 그 계산량으로 인해 실시간 시스템의 구현이 어려워지므로 색상을 이용하여 대략적인 위치를 추정하고 작은 범위를 검색하여 정확한 입술의 위치를 찾는 것이다.

검색은 색상을 이용하여 찾은 입술의 초기 위치에서 일정 범위 내에서 16×16 입술 영상 블럭의 중심을 옮겨 가며 입술 영상을 획득하고 이를 SVM 패턴 분류기에 입력으로 주어 입력으로 넣은 영상이 입술인지 아닌지를 판별하게 된다. 이때 검색 범위는 색상으로 추정한 입술 위치의 오차에 따라 탄력적으로 정할 수 있는데, 얼굴 폭을 40픽셀로 정규화하였을 때, 오차는 평균적으로 X로 3~4픽셀 정도, Y로 5~6픽셀 정도였다. 검색 범위를 이를 포괄할 정도로 X로 ±6, Y로 ±10정도로 설정하였다. 검색 범위의 모든 점을 이동하며 SVM 패턴 분류기의 출력인 입술 정합도 값을 계산하고 그 중 가장 큰 값을 입술의 위치로 정한다.

그림 7의 (a)는 초기 입술의 위치로부터 일정 범위의 검색을 통해 계산된 SVM 입술 정합도 값을 보여주고 있으며, (b)는 이를 통해 조정된 입술의 위치를 나타내고 있다.



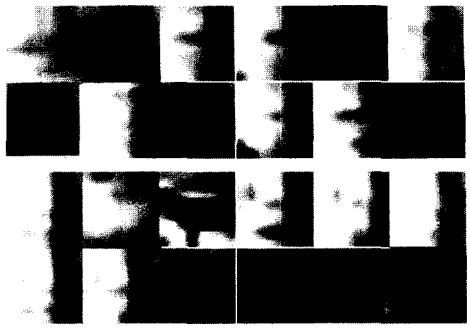
(a) (b)

그림 7 SVM으로 검색한 영역과 각 점에서의 입술 정합도를 밝기값으로 표현한 영상과 최고의 정합도 값을 가지는 지점을 조정된 입술 위치로 정한 결과

3. 실험 결과

3.1 입술 검출의 견실성

제안한 방법을 통해 입술 검출을 수행해보았다. 얼굴색/입술색 학습은 18장의 영상에서 직접 얼굴 영역과 입술 영역을 선택하여 수행하였고, SVM 학습은 예제와 반례 입술 영상 153장을 통해 수행하였다. 그림 8은



(a) (b)

그림 8 SVM 학습에 사용한 예제(a)와 반례(b). 본 영상은 DCT 저주파 계수를 복원하여 표현된 것이다.

그림 9 입술 검출 실험 결과의 예

SVM 학습에 사용한 예제와 반례의 일부이다.

실험은 별도의 화장을 하지 않은 20대~30대 연령의 39명의 화자가 555개의 단어를 발음한 모습을 담은 동영상 21627개에 대해 제안된 방법을 사용하여 입술 위치 검출을 수행하였다. 제안된 방법은 PC PentiumIV 1.6GHz의 컴퓨터에서 Visual C++로 구현되었으며, 동영상 획득에 사용된 입력장치는 QuickCam Pro3000 USB 카메라였다. 조명은 형광등 조명을 사용한 일반 실내 조명 환경이었으며, 실험 결과 118개의 동영상에서 입술 검출 오류가 발생하여 0.5%의 추적 오류를 나타내었다. 다음 그림 9는 실험에 사용된 영상과 입술 검출 모습을 나타내고 있다.

추적 오류가 발생한 경우는 머리카락 색상이 피부색과 유사하여 얼굴이 크게 잡혀 얼굴 크기 표준화시 스케일링 문제가 생긴 경우였다. 이 경우는 학습한 입술크

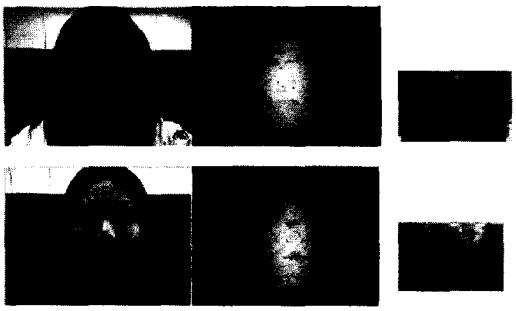


그림 10 입술을 잘못 찾은 예. 머리칼 색이 피부색과 비슷하여 얼굴을 크게 잡아 입술 모양을 제대로 찾지 못하였다.

기 영상내의 입술크기가 다르기 때문에 SVM 패턴분류기에서 입술의 위치를 잘못 찾을 가능성이 높았다(그림 10).

3.2. 입술 검출 속도

PentiumⅣ 1.6GHz의 컴퓨터에서 제안된 알고리즘을 구현하여 실험을 해본 결과 초당 11.5 프레임의 영상에 대해 입술 검출을 수행함으로써 실시간 바이모달 음성 인식 시스템 구현에 사용할 수 있음을 알 수 있었다. 그러나 색상을 이용한 입술의 초기 위치 추정 없이 SVM만을 이용하여 입술의 위치를 찾을 경우 초당 8 프레임의 영상에 대해 입술 검출을 수행하였다. 이로부터 색상을 이용한 입술의 위치 추정을 통해 입술 위치 검색의 범위를 줄임으로써 수행 시간을 단축할 수 있었음을 알 수 있다.

4. 결 론

본 논문은 바이모달 음성 인식기에 사용할 시각 특징의 추출을 위해 필요한 입술 위치 검출 방법을 제안하고 있다. 기존의 방법들의 경우 입술의 위치 검출을 위해 조명이나 개인차의 영향을 받는 색상만을 이용하기 때문에 입술의 위치 를 정확히 얻을 수 없었다. 이를 극복하기 위해 입술 윤곽을 Active contour나 Deformable template 등 가변 모델을 사용하여 색상의 부정확성을 어느정도 극복하여 입술 윤곽을 찾고자 하였으나 이 방법 역시 모델구축의 어려움과 파라메터 조절의 어려움으로 인해 실제 적용에 있어 한계점을 지니고 있다. 따라서 본 논문에서는 기존의 방법들이 지닌 이러한 한계를 극복하기 위해 SVM 패턴 분류기를 이용하여 보다 정확한 입술 위치 검출을 하고자 하였으며, 이로 인한 속도 감소를 극복하기 위해 입술 색상 분포를 학습하여 이로부터 대략적인 입술의 위치를 찾아 입술 위치 검출을 위한 검색 범위를 한정하였다. 제안된 방법은 다양한 화자와 발음에 대해 낮은 에러율로 입술의 위치를 찾음으로써 바이모달 음성 인식기에 적합한 입술 특징을 제공할 수 있도록 기여하고 있다.

또한 본 논문에서는 기존의 논문들이 얼굴의 위치가 고정되어 있다고 제한한 가정을 탈피, 얼굴 검출 단계부터 입술 위치 검출 과정까지 기술함으로써 실제 바이모달 음성인식 시스템에 적용할 수 있는 실용성 있는 기술을 제시하고 있다. 비록 얼굴 영역 검출 방법은 색상만을 이용하여 단순하나 이는 실시간 시스템 구축을 위해 적합한 선택으로 보이며, 적용 분야의 특성상 단일 화자의 얼굴만이 영상입력장치에 들어오기 때문에 얼굴 영역 검출 성능에 있어 크게 불리하지 않다고 판단된다.

그러나 입술색 분포를 이용함에 있어 개선해야 할 점은 남아있다. 사람의 입술색은 얼굴색과 밀접한 관련을 가지는데, 입술색 분포를 사용할 때 이 연관성을 반영하여 입술색 분포를 재가공하여 반영한다면 더욱 입술의 초기 위치를 정확하게 추정할 수 있을 것이며, 따라서 입술의 위치를 미세 조정하기 위해 검색해야 할 공간도 줄어들 것이다. 더불어 Edge와 같이 입술에 두드러지게 나타나는 특징도 함께 사용할 경우 입술의 위치 설정에 더욱 도움이 될 것이다. 추후 연구는 이와 같이 초기 위치 선정 방법 개선에 초점을 맞추어 진행할 것이다.

참 고 논 문

[1] G. Potamianos, H. P. Graf, and E. Cosatto, An Image Transform Approach for HMM Based Automatic Lipreading, Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, vol.3, Page(s): 173-177, 4-7 Oct 1998.

[2] Kaucic R. and Blake A., Accurate, real-time, unadorned lip tracking, Sixth International Conference on Computer Vision, Page(s): 370-375, 4-7 Jan 1998.

[3] Sadeghi M., Kittler J. and Messer K., "Modelling and segmentation of lip area in face images," IEEE Proceedings on Vision, Image and Signal Processing, Volume: 149 Issue: 3, Page(s): 179 -184 , Jun 2002.

[4] Zhang Jian, Kaynak M.N., Cheok A.D., Ko Chi Chung, Real-time lip tracking for virtual lip implementation in virtual environments and computer games, The 10th IEEE International Conference on Fuzzy Systems, Volume: 3, Page(s): 1359-1362, 2001.

[5] Lucey S., Sridharan. S. and Chandran. W., Chromatic lip tracking using a connectivity based fuzzy thresholding technique, ISSPA '99. Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, Volume: 2, Page(s): 669-672 vol.2, 1999.

[6] Chan M.T., Zhang, Y. and Huang T.S., Real-time lip tracking and bimodal continuous speech recognition, IEEE Second Workshop on Multimedia Signal Processing, Page(s): 65-70, 7-9 Dec 1998.
[7] Delmas P., Eveno. N. and Lievin. M., Towards robust lip tracking, Proceedings of 16th International Conference on Pattern Recognition, Volume: 2, Page(s): 528-531 vol.2, 2002.
[8] Zhilin Wu, Aleksic P.S. and Katsaggelos A.K. Lip tracking for MPEG-4 facial animation, Proceedings of Fourth IEEE International Conference on Multimodal Interfaces, Page(s): 293-298, 2002.
[9] Robert M. Haralick and Linda G. Shapiro, Computer and Robot Vision, Vol1. pp. 73-74, Addison-Wesley publishing company., 1992.
[10] Christopher J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2, pp. 121-167, 1998.

정 지 년
한국과학 기술원 전산학과 학사(1999년) 석사(2000년). 현재 한국과학기술원 전산학과 박사과정재학중(2000년~현재). 관심분야는 컴퓨터 시각, 패턴 인식

양 현 승
서울대 학사(1976). Purdue University 전자과 석사(1983). Purdue University 전자과 박사 학위 취득(1986). University of IOWA 전자전산과 조교수(1986년~1988년). 인공지능 연구센터 시각 연구실장(1988년~1999년). 한국과학기술원 전산학과 정교수(1988년~현재). 관심분야는 컴퓨터 시각, 로보틱스, 인공지능, 멀티미디어