

PC카메라를 이용한 실시간 립리딩 시스템 설계 및 구현

이은숙, 이지근, 이상설, 정성태
원광대학교 전기 전자 및 정보공학부

Design & Implementation of Real-Time Lipreading System using PC Camera

Eun-Suk Lee, Chi-Geun Lee, Sang-Seol Lee, Sung-Tae Jung
Dept. of Electrical Electronic and Information Engineering, WonKwang University

요약

최근 들어 립리딩은 멀티모달 인터페이스 기술의 융용분야에서 많은 관심을 모으고 있다. 동적 영상을 이용한 립리딩 시스템에서 해결해야 할 주된 문제점은 상황 변화에 독립적으로 얼굴 영역과 입술 영역을 추출하고 오프라인이 아닌 실시간으로 입력된 입술 영상의 인식을 처리하여 립리딩의 사용도를 높이는 것이다. 본 논문에서는 사용자가 쉽게 사용할 수 있는 PC카메라를 사용하여 영상을 입력받아 학습과 인식을 실시간으로 처리하는 립리딩 시스템을 구현하였다. 본 논문에서는 움직임이 있는 화자의 얼굴영역과 입술영역을 컬러, 조명등의 변화에 독립적으로 추출하기 위해 HSI모델을 이용하였다. 입력 영상에서 일정한 크기의 영역에 대한 색도 히스토그램 모델을 만들어 색도 영상에 적용함으로써 얼굴영역의 확률 분포를 구하였고, Mean-Shift Algorithm을 이용하여 얼굴영역의 검출과 추적을 하였다. 특징 점 추출에는 이미지 기반 방법인 PCA 기법을 이용하였고, HMM 기반 패턴 인식을 사용하여 실시간으로 실험영상데이터에 대한 학습과 인식을 수행할 수 있었다.

1. 서론

립리딩은 잡음 환경에서 음성인식 저하를 보완하는 방법의 하나로 음성 청취가 어려운 조건에서 음성인식을 위한 보조수단으로 활용되어 최근에 이에 대한 활발한 연구가 진행되고 있다. 기존의 립리딩 시스템의 실험 영상은 여러 가지 제한 조건을 전제로 하여 특정한 환경[1]에서만 사용될 수 있고 미리 준비한 입술 영상 데이터베이스로 오프라인 인식 실험하는 문제점을 가지고 있다. 따라서 본 논문에서는 화자의 움직임이 허용되고 컬러, 조명등의 환경 변화에 독립적으로 입술 영역을 추출함으로써 입력 영상에 아무런 제한을 두지 않고 사용자가 쉽게 사용할 수 있는 PC 카메라를 이용하여 실시간 립리딩 시스템을 구현한다. 본 논문의 실시간 립리딩 시스템의 구조는 그림 1에 나타나 있다.

본 연구는 한국과학재단 목적기초연구과제(과제번호:R05-2003-000-10770-0)연구비에 의해 연구되었음

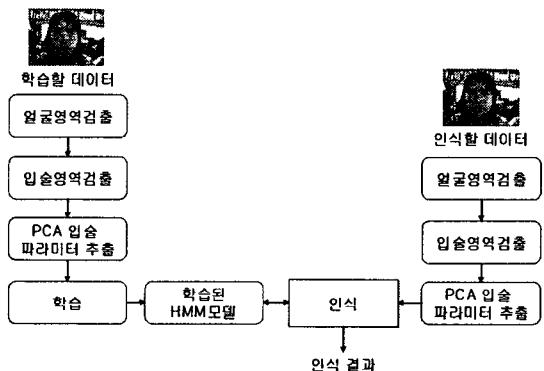


그림 1. 본 논문의 립리딩 시스템 구조

본 논문에서는 얼굴 분포 영역을 구하기 위해 입력 영상에 일정한 크기의 영역에서 픽셀들의 색도를 이용하여 색도 히스토그램 모델을 만들고 이를 색도 영상에 적용하여 얼굴의 확률 분포 영역을 추출한다. 얼굴 영역의 검출과 움직이는 화자의 얼굴을 추적하기

위한 방법으로는 Mean-Shift Algorithm을 사용하였다. 검출된 얼굴 영역 내에서 입술 영역 검출을 위해 서는 색도와 명도를 이용하였다. 입술영역의 파라미터를 추출하기 위해 이미지 기반 방법인 주성분 분석(Principal Component Analysis)을 사용하였고 입술 정보의 학습과 인식은 HMM(Hidden Markov Model)을 이용하였다. 본 논문에서는 화자가 PC카메라에 직접 학습데이터를 입력시켜 학습을 하고 인식하고자 하는 데이터를 입력시켜 실시간으로 인식 결과를 보는 립리딩 시스템을 구현하였다.

2. 입술 분할 및 특징 추출

2.1 얼굴 영역 검출

영상 캡처 장비에서 컬러 영상은 RGB 컬러인데, 이는 조명 변화에 민감하게 반응한다. 그러므로 본 논문에서는 RGB컬러를 HSI 컬러로 변환하여 원래 객체가 가지고 있는 색을 표현하는 색도 성분 이용으로 영상이 조명에 영향을 받지 않게 하였다. 다양한 화자의 얼굴 색도를 고려하여 얼굴 영역 분포를 구하기 위해 그림 2와 같이 색도 히스토그램 모델을 만든다. 입력영상의 첫 프레임에서 영상 중앙에 위치하는 64×64 크기의 윈도우를 지정하고 촬영 시작 전에 얼굴을 윈도우 크기 보다 더 크게 위치시킨다. 영상 입력 시에는 영상의 첫 프레임에서 윈도우 영역 안의 색도 값을 추출하여 색도 히스토그램 모델을 생성한다. 색도 값을 구하는 방법으로 참고논문[2]에서 제안된 식(1)을 사용하였다.

$$Hue = 256 \times \left(\frac{G}{R} \right) \quad (1)$$

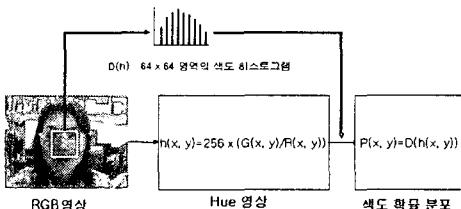


그림 2. 얼굴의 색도 확률 분포 추출 과정

색도 히스토그램 모델과 Hue 영상은 입력되는 이미지에서 R성분이 G성분보다 클 경우는 식(1)을 이용하여 색도 값을 구하고, R성분이 G성분보다 작을 경우는 색도 값을 0으로 하였다. Hue 영상의 각 픽셀을 색도 히스토그램 모델($D(h)$)에 적용하여 색도의 확률 분포 영역($P(x,y)$)을 구하게 된다. 그림 3.(a)는 색도

히스토그램 모델을 적용하여 얼굴의 색도 확률 분포를 구한 것이다. 그림 3.(a)에서 보는바와 같이 얼굴의 색도 분포와 비슷한 색을 가진 영역이 배경에 있더라도, Mean-shift Algorithm을 적용하면 그림 3.(b)와 같이 얼굴 객체만을 추출할 수 있다.



그림 3. (a) 얼굴의 색도 확률 분포 (b) 얼굴영역검출

또한 얼굴의 색도 확률 분포 영상에서 Mean-Shift Algorithm을 사용하여 그림 4와 같이 임의의 방향으로 이동하거나 앞뒤로 움직여 얼굴의 크기가 변하더라도 얼굴 영역을 추적한다.



그림 4. Mean-Shift Algorithm 이용하여 얼굴영역추적

Mean-Shift Algorithm의 절차는 다음과 같다.[3]

1. 탐색 윈도우 크기를 선택한다.
2. 탐색 윈도우의 초기 위치를 선택한다.
3. 탐색 윈도우에서 평균 위치를 계산한다.
4. 과정 3에서 계산된 평균 위치에서 탐색 윈도우 중심을 구한다.
5. 고밀도 영역으로 수렴할 때까지 3-4과정을 반복한다.

3-4 과정은 참고논문[3][4]에서 제시한 식(2)을 사용하여 탐색 윈도우의 중심과 방위 θ 를 구한다.

만약, $P(x,y)$ 가 x, y 의 위치에 이미지 색도 값을이라 할 때 이미지 모멘트들과 두 번째 모멘트는 다음과 같다.

$$\begin{aligned} M_{00} &= \sum_x \sum_y P(x,y) & M_{11} &= \sum_x \sum_y xyP(x,y) \\ M_{10} &= \sum_x \sum_y xP(x,y) & M_{01} &= \sum_x \sum_y yP(x,y) \quad (2) \\ M_{20} &= \sum_x \sum_y x^2 P(x,y) & M_{02} &= \sum_x \sum_y y^2 P(x,y) \end{aligned}$$

식(2)을 사용하여 식(3)과 같이 검색 윈도우의 중심을 구한다.

$$x_c = \frac{M_{10}}{M_{00}}, \quad y_c = \frac{M_{01}}{M_{00}} \quad (3)$$

식(3)을 이용하여 중간 값 a, b, c 를 식(4),(5),(6)과 같이 구하고 이들을 이용하여 방위 θ 를 구한다.

$$a = \frac{M_{20}}{M_{00}} - x_c^2 \quad (4) \quad b = 2\left(-\frac{M_{11}}{M_{00}} - x_c, y_c\right) \quad (5)$$

$$c = \frac{M_{02}}{M_{00}} - y_c^2 \quad (6) \quad \theta = \frac{\arctan(b, (a-c))}{2} \quad (7)$$

영상에서 분포된 얼굴 영역의 길이 (l)와 너비(w)는 다음 식(8),(9)를 이용하여 구한다.

$$l = \sqrt{\frac{(a+c) + \sqrt{b^2 + (a-c)^2}}{2}} \quad (8)$$

$$w = \sqrt{\frac{(a+c) - \sqrt{b^2 + (a-c)^2}}{2}} \quad (9)$$

위 식에서 구해진 얼굴영역의 길이 (l)와 너비(w)의 값으로 사각형의 얼굴 영역을 검출한다

2.2 입술 영역 검출

입술은 주위 피부와 색상 대비나 명암대비가 크지 않고 말을 하는 동안 입술의 형태가 계속적으로 변하기 때문에 입술 영역만을 두드러지게 구분하는 것은 단순하지 않은 문제로 인식되고 있다.

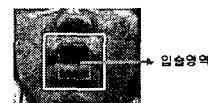
본 논문에서는 입술영역을 색도 평균 마스크를 이용하여 추출하였다. 입술은 얼굴의 크기에 비례적으로 있으므로 마스크의 크기를 얼굴 영역의 x축의 1/4로 한다. 색도 평균을 구하는 마스크로 얼굴 영역 내에서 회선 검색하여 입술이 분포한 영역을 찾는다. 입술은 얼굴의 안쪽 영역에 분포하므로 검색영역을 안쪽에서부터 시작한다. 입술의 R성분은 얼굴 영역의 다른 픽셀의 R성분 값보다 크다. 그러므로 입술의 색도 값은 식(1)을 통해 작은 값의 분포를 갖는다. 즉, 입술 분포 영역 검출은 마스크의 색도 평균이 최저인 부분이 된다. 이와 같이 검출된 입술 영역이 그림 5에 나타난다.



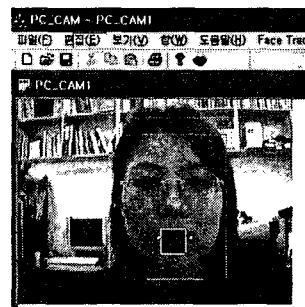
그림 5. 입술 영역 검출

위 과정으로 개략적인 입술 영역을 구한 다음 PCA를 통한 입술 파라미터를 추출하기 위해 일정한 입술 위치와 크기를 갖는 입술영역을 추출한다. 본 논문에서는 그림 6.(b)에 나타나 있는바와 같이 정확한 입술의 분할과 일정한 크기의 입술 영역을 추출하기 위해 입술의 양 끝점을 검출한다. 앞 단계에서 검출한 개략적인 입술 영역에서 입술의 양 끝점을 검출하기 위해 그림 6.(a)처럼 개략적인 입술 영역보다 더 큰 영역을 잡는다. 양 끝점의 검출은 입안의 명암도가 낮은 특성을 이용하는데 명암도는 식(10)와 같이 구하였다.

$$\text{명암도} = \frac{(R+G+B)}{3} \quad (10)$$



(a)



(b)

그림 6. (a) 입술 양 끝점 검출 영역 (b) 입술 양 끝점 검출

입술의 양 끝점을 추출하기 위해 검출된 영역에서 식(10)으로 각 픽셀의 명암도를 구하고 영역 내에서 각 열의 명암도 평균을 구한 다음 이 평균의 1/2 을 임계값으로 정하여 입술의 양 끝점을 검출한다. 검출한 입술의 양 끝점의 두 좌표는 입의 중앙 좌표를 구하는데 이용되고 구한 중앙 좌표를 기준으로 70×50 크기의 입술영역을 저장한다. PCA를 이용하여 입술의 파라미터를 구하기 위해 그레이 레벨의 PGM(Portable GrayMap) 파일로 저장한다. 그림 7은 “정지”라는 발음을 하였을 경우에 입력영상에서 추출되는 17 프레임의 입술영역이다.



그림 7. 입술 영역 저장 프레임

2.3 주성분 분석을 이용한 입술 파라미터추출

주성분 분석(Principal Component Analysis)은 고차원 입력 벡터를 차원의 벡터로 표현하여 몇 개의 주성분 값으로 나타내어 주는 방식이다. 서로 연관 있는 n 개의 변수의 전치행렬은 식(11)과 같고, 이를 식(12)과 (13)에 적용하여 평균벡터와 벡터 x 들의 집합에 대한 공분산 행렬을 구한다.

$$x = [x_1, x_2, \dots, x_n]^T \quad (11)$$

$$\bar{m}_x = \frac{1}{M} \sum_{k=1}^M x_k \quad (12)$$

$$C_x = \frac{1}{M} \sum_{k=1}^M (x_k - \bar{m}_x)(x_k - \bar{m}_x)^T \quad (13)$$

평균벡터와 공분산 행렬을 통해 고유 벡터를 구한 뒤 고유벡터들을 대응되는 고유 값의 크기에 따라 고유 벡터들을 정렬하여 새로운 행렬 A 를 만들어 이를 변환 행렬로 사용한다. 이 변환행렬은 식(14)과 같이 사용하면 차원이 큰 벡터 x 를 차원이 작은 벡터 y 로 변환 할 수 있게 된다. 벡터 y 와 A 의 역행렬을 이용하면 x 와 유사한 벡터를 복원할 수 있는 특성이 있기 때문에 벡터 y 를 벡터 x 의 특정 계수로 사용할 수 있게 된다. 즉, 입술 영상 한 프레임을 행렬 A 와 연산하여 특징 계수를 구하게 된다.

$$y = A(x - \bar{m}_x) \quad (14)$$

본 실험에서는 각 프레임에서 12개의 입술 특징 계수를 추출하였고, 이 계수를 HMM 기법에 적용하여 학습과 인식 실험을 하였다.

4. 실시간 립리딩 시스템

본 논문에서 학습데이터는 PC 카메라를 이용하여 영상을 입력받아 얼굴영역을 검출하고 검출한 얼굴 영역 내에서 입술영역을 검출하여 실시간으로 학습한다. 실험데이터는 학습데이터 준비과정과 같은 과정을 거쳐 데이터를 준비한다. 입술 정보의 학습데이터와 실험데이터의 학습 및 인식은 HTK(Hidden Markov Model ToolKit)을 사용하였다. HTK는 영국 캠브리지 대학에서 공개한 것으로 음성 인식 성능 면에서 아주

우수한 것으로 평가되고 있다. 실험 단어로는 오디오 및 CD플레이어를 작동시키기 위해 필요한 단어들(재생, 정지, 종료, 앞으로 뒤로, 목차, 다음, 이전, 선택, 취소)을 사용하였다. 입력되는 영상은 PC 카메라 사용으로 15 Frame/sec 이고, 저장되는 입술정보는 대략 10~20 프레임의 이미지를 얻었다. 학습데이터와 실험데이터는 동일한 화자의 입술영상인 화자종속으로 하였다.

5. 결론 및 향후과제

본 논문에서는 상황에 독립적으로 얼굴 및 입술 영역을 추출하였고 추출한 입술정보로 실시간 화자종속 립리딩 시스템을 구현하였다. 기존의 립리딩 실험에서는 입술 영상만을 이용한 제한된 실험데이터를 사용하였으나 본 실험에서는 입술 영상에 제한을 하지 않은 영상으로 입술 실험데이터를 추출하였다. 또한 오프라인 상으로 실험되었던 립리딩 시스템을 실시간으로 작동하도록 하여 일반 사용자가 립리딩 시스템의 학습과 인식을 직접 할 수 있게 하였다. 향후 보다 더 견고한 입술 영역 추출에 대한 연구가 필요하고 다른 종류의 카메라에서 실시간 추적 및 인식에 대한 연구가 필요하다.

[참고문헌]

- [1] J.S.D. Mason, J. Brand, R. Auckenthaler, F. Deravi, C. Chibelushi, "Lip Signatures for automatic person recognition", multimedia Signal Processing, 1999. IEEE 3rd Workshop on, pp.457-462, 1999
- [2] M. Lievin, F. Luthon, "Lip features automatic extraction", Image Processing, ICP 98. Proceedings. 1998 International Conference On, vol.3, pp.168-172, 1998
- [3] Gary R. Bradski, "Real Time Face and Object Tracking as a Component of a Perceptual User Interface", Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on, 19-21, pp.214 -219, Oct. 1998
- [4] W.T. Freeman, K. Tanaka, J. Ohta, K. Kyuma, "Computer vision for computer games", Automatic Face and Gesture Recognition, 1996, Proceedings of the Second International Conference on, 14-16, pp.100 -105, Oct. 1996