견고한 입술 영역 추출을 이용한 립리딩 시스템 설계 및 구현

이은숙, 이호근, 이지근, 김봉완, 이상설, 이용주, 정성태
원광대학교 컴퓨터공학과

Design & Implementation of Lipreading System using Robust Lip Area Extraction

Eun-Suk Lee, Ho-Geun Lee, Chi-Geun Lee, Bong-Wan Kim, Sang-Seol Lee, Yong-Joo Lee, Sung-Tae Jung
Dept. of Computer Engineering, Wonkwang University

요 약

최근 들어 립리딩은 멀티모달 인터페이스 기술의 응용분야에서 많은 관심을 모으고 있다. 동적 영상을 이용한 립리딩 시스템에서 해결해야 할 주된 문제점은 상황 변화에 독립적인 얼굴 영역과 입술 영역을 추출하는 것이다. 본 논문에서는 움직임이 있는 영상에서 화자의 얼굴영역과 입술영역을 컬러, 조명등의 변화에 독립적으로 추출하기 위해 HSI 모델과 블록 매칭을 이용하였고 특징 점 추출에는 이미지 기반 방법인 PCA 기법을 이용하였다. 추출된 입술 파라미터와 음성 데이터에 각각 HMM 기반 패턴 인식 방법을 개별적으로 적용하여 단어를 인식하였고 각각의 인식 결과를 가중치를 주어 합병하였다. 실험 결과에 의하면 잡음으로 음성 인식률이 낮아지는 경우에 음성인식과 립리딩을 함께 사용함으로써 전체적인 인식 결과를 향상시킬 수 있었다.

1. 서론

립리딩은 잡음 환경에서 음성인식 저하를 보완하는 방법의 하나로 음성 청취가 어려운 조건에서 음성인식을 위한 보조수단으로 활용될 수 있기 때문에 최근에 들어 이에 대한 활발한 연구가 진행되고 있다. 립리딩 시스템 구현을 위한 주요 기술로는 입술 분할 기술, 특징 추출 기술, 음성-입술 정보 통합 기술이 있다[1]. 립리딩의 성능을 높이기 위해서는 견고한 입술 분할이 필수적이다. 많은 기존의 시스템들은 견고한 입술 분할을 위해 실험 영상에 여러 가지 제한 조건을 전제로 함으로써 특정한 환경[2]에서만 사용될 수 있는 문제점을 가지고 있다. 따라서 본 논문에서는 화자의 움직임이 허용되고 컬러, 조명등의 환경 변화에 독립적으로 입술 영역을 추출함으로써 입력 영상에 아무런 제한을 두지 않는 립리딩 시스템을 설계하고 구현한다. 본 논문의 립리딩 시스템의 구조는 그림 1에 나타나있다.

본 연구는 한국과학재단 목적기초연구(과제번호: R05-2003-000-10770-0(2003))지원으로 수행되었음

본 논문에서는 얼굴 영역의 추출을 위해서는 픽셀의 색도와 블록 매칭을 이용하였고, 검출된 얼굴 영역 내에서 입술 영역 검출을 위해서는 색도와 명도를 이용하였다. 입술영역의 파라미터를 추출하기 위해 이미지 기반 방법인 주성분 분석(Principle Component Analysis)을 사용하였고, 입술정보와 음성정보의 학습 및 인식은 HMM(Hidden Markov Model)을 이용하였다. 본 논문에서는 자동차에서 사용될 수 있는 10개 단어를 15명이 발성한 장면으로 학습을 수행하였고, 2명의 발성 장면을 사용하여 인식 실험을 수행하였다. 실험결과 음성 인식과 립리딩을 합병함으로써 잡음이 있는 경우에 인식률을 향상시킬 수 있었다.

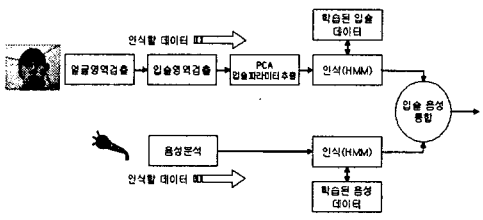


그림 1. 본 논문의 립리딩 시스템 구조

2. 입술 분할 및 특징 추출

2.1 얼굴영역 검출

본 논문에서는 픽셀의 색도와 함께 블록 매칭 기법을 이용하여 얼굴 영역을 검출한다. 영상 캡쳐 장치에서 컬러 영상은 RGB 컬러인데, 이는 조명 변화에 민감하게 반응한다. 그러므로 본 논문에서는 RGB 컬러를 HSI 컬러로 변환하여 원래 객체가 가지고 있는 색을 표현하는 색도 성분 이용으로 영상이 조명에 영향을 받지 않게 하였다. 이와 같은 픽셀의 색도 성분은 얼굴이 분포한 위치를 찾는데 사용되고 여기에 더 견고한 얼굴 영역을 검출하기 위해 사람의 동적인 특성을 나타낼 수 있는 블록 매칭 기법으로 프레임간의 모션 벡터를 이용하여 얼굴 영역을 검출한다.

2.1.1. 색도를 이용한 얼굴 영역 분포

RGB 컬러를 HSI 컬러로 변환하는 방법으로 여러 가지가 있는데, 본 논문에서는 색도 값을 구하는 방법으로 참고문헌[3]에서 제안된 식(1)을 사용하였다. 얼굴 영역 분포를 구하는데 있어서 먼저 각 행과 열의 픽셀의 색도 누적 값을 구하고 각 행과 열의 누적된 색도 값의 평균을 구한 다음 이 평균의 1/2을 임계값으로 정하였다. 그림 2에서 영상내의 사각형은 임계값을 적용하여 추출한 얼굴 영역을 보여준다.

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

그림 2. 색도를 이용한 얼굴 영역 추출

그러나 색도 분포로 얼굴영역을 구하는 것은 배경에 얼굴과 비슷한 색도를 갖는 객체가 있을 경우 오류를 범하게 된다. 이를 보완하기 위해 얼굴의 색도 분포와 블록 매칭의 모션벡터가 모두 있는 곳을 얼굴 영역으로 추출한다.

2.1.2. 블록 매칭 기법을 이용한 얼굴 영역 검출

카메라는 고정적이고 사람은 동적인 실제 상황에 적용하기 위해 연속적 프레임간의 모션 벡터를 구하는 블록 매칭 기법을 이용한다. 블록 매칭 방법에는 전역탐색(full-search) 방법과 K-단계 탐색 방법 등이를 응용한 다양한 방법들[4]이 있는데 본 논문에서는 3-단계 블록 매칭 기법을 이용하였다. 형태가 변형적이지 않는 한 물체가 움직일 때 블록 매칭의 모션 벡터는 한 방향으로 일정하다. 그림 3과 같이 모션 벡터가 있는 곳을 사람의 움직임으로 파악하여 더 견고한 얼굴 영역을 추출할 수 있도록 하였다.

그림 3. 얼굴 영역의 블록 매칭 모션 벡터

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 마스크와 블록 매칭을 이용하여 추출하였다. 그림 4(a)와 같이 색도 평균을 구하는 마스크로 얼굴 영역 내에서 회선 검색하여 입술이 분포한 영역을 찾는다. 입술은 얼굴의 안쪽 영역에 분포하므로 검색영역을 안쪽에서부터 시작한다. 또한 입술은 얼굴의 크기에 비례적으로 있으므로 마스크의 크기를 얼굴 영역의 x축의 1/4 로 한다. 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 부분인 된다. 이와 같이 검출된 입술 영역이 그림 4(b)에 나타나있다.

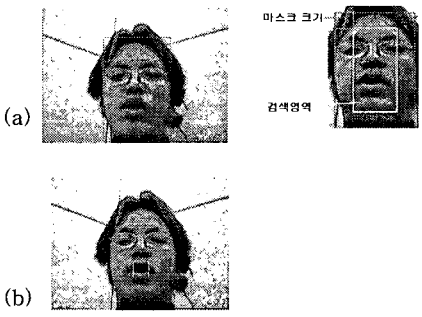


그림 4.(a)입술검색 마스크선정 (b)입술영역 검출결과

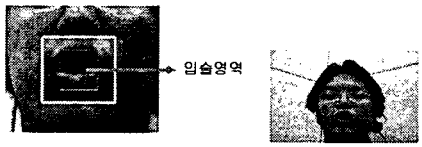
다음으로 색도를 이용한 입술 영역 추출의 오류를 보완하고 더욱 정확한 입술 영역을 추출하기 위해 블록 매칭 기법을 추가한다. 입술은 그림 5와 같이 모션 벡

터가 상하 반대방향을 가지는 특성이 있다. 상하 반대 방향의 모션 벡터가 나타나는 부분을 입술 영역 부분으로 간주한다.

그림 5. 입술 영역의 블록 매칭 모션 벡터

위 과정으로 개략적인 입술 영역을 구한 다음 PCA를 통한 입술 파라미터를 추출하기 위해 일정한 입술 위치와 크기를 갖는 입술영역을 추출한다. 본 논문에서는 그림 6에 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(2)과 같이 구하였다.

$$\text{명암도} = \frac{(R+G+B)}{3} \quad (2)$$



(a) (b)

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

입술의 양 끝점을 추출하기 위해 검출된 영역에서 식(2)로 각 픽셀의 명암도를 구하고 영역 내에서 각 열의 명암도 평균을 구한 다음 색도를 이용한 얼굴영역 검출에서 임계값을 구한 방법과 같은 방법을 이용하여 입술의 양 끝점을 검출한다. 검출한 입술의 양 끝점의 두 좌표는 입의 중앙 좌표를 구하는데 이용되고 구한 중앙 좌표를 기준으로 70*50 크기의 입술영역을 저장한다. PCA을 이용하여 입술의 파라미터를 구하기 위해 그레이 레벨의 PGM(Portable GrayMap) 파일로 저장한다. 그림 7은 "다음"라는 발음을 하였을 경우에 입력영상에서 추출되는 24개 프레임의 입술 영역이다.

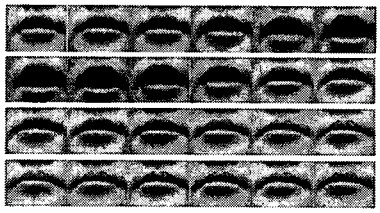


그림 7. 입술 영역 저장 프레임

2.3 주성분 분석(PCA)을 이용한 입술 파라미터 추출

주성분 분석(Principle Component Analysis)은 고차원 입력 벡터를 저차원의 벡터로 표현하여 몇 개의 주성분 값으로 나타내어 주는 방식이다. 서로 연관 있는 n개의 변수의 전치행렬은 식(3)과 같고, 이를 식(4)와 (5)에 적용하여 평균벡터와 벡터 x들의 집합에 대한 공분산 행렬을 구한다.

$$x = [x_1, x_2, \cdots, x_n]^T \quad (3)$$

$$m_x = \frac{1}{M}\sum_{k=1}^{M} x_k \quad (4)$$

$$C_x = \frac{1}{M}\sum_{k=1}^{M} x_k x_k^T - m_k m_k^T \quad (5)$$

평균벡터와 공분산 행렬을 통해 고유 벡터를 구한 뒤 고유벡터들을 대응되는 고유 값의 크기에 따라 고유벡터들을 정렬하여 새로운 행렬A를 만들어 이를 변환행렬로 사용한다. 이 변환행렬은 식(6)과 같이 사용하면 차원이 큰 벡터 x를 차원이 작은 벡터 y로 변환할 수 있게 된다. 벡터 y와 A의 역행렬을 이용하면 x와 유사한 벡터를 복원할 수 있는 특성이 있기 때문에 벡터 y를 벡터 x의 특징 계수로 사용할 수 있게 된다. 즉, 입술 영상 한 프레임을 행렬 A와 연산하여 특징 계수를 구하게 된다.

$$y = A(x - m_x) \quad (6)$$

3. 립리딩 시스템

3.1 입술, 음성정보 학습 및 인식

입술, 음성 정보의 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 자동차 실내 환경에서 오디오 및 CD 플레이어를 작동시키기 위해 필요한 단어들(재생, 정지, 종료, 앞으로, 뒤로, 목차, 다음, 이전, 선택, 취소)을 사용하였다. 입력되는 영상은 30 Frames/sec이고 저장되는 입술 정보는 대략 20~30 프레임의 이미지를 얻었다. 인식 실험에서 입술인식 및 음성인식은 화자독립으로 수행하였으며 학습데이터는 15명, 실험 데이터 2명으로 하였다.

입술정보 인식과정으로 먼저 PCA로 입술의 학습 데이터와 실험 데이터 특징 파라미터를 추출하고 이를 HTK에 적용하여 학습 및 인식 결과를 구하였다. 립리딩 결과 40%의 인식률을 보였다. 음성 인식의 경우에는 표 1과 같이 잡음의 비율에 따라 인식률은 변화하였다.

잡음	30	20	10	5
음성 인식률	20%	30%	35%	80%

표 1, 잡음 인식률

이처럼 잡음에 따라 낮아지는 음성 인식률의 보완을 위해 입술 정보를 합병하여 실험하였다.

3.2 음성-입술 정보 합병에 의한 인식

본 실험의 립리딩 시스템은 입술정보와 음성정보를 각각 인식한 후 인식 결과를 합병하는 방법[5]을 사용하였다. 입술인식률(Mlip)과 음성인식률(Mspeech)결과에 가중치(α)를 주어 합병하여 식(7)과 같이 합병 인식결과(M)를 구했다.

$$M = \alpha Mlip + (1-\alpha) Mspeech \quad (7)$$

표 2는 입술인식, 음성인식 결과에 가중치를 달리하여 합병한 인식한 결과이다.

가중치(α)		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
음성 입술 합병 인식률	잡음30	20	20	25	25	25	30	35	35	**40**
	잡음20	30	30	30	30	30	25	25	35	**40**
	잡음10	35	40	45	40	50	**50**	45	45	45
	잡음5	80	85	**85**	75	75	75	55	55	45

표 2. 가중치에 따른 합병 인식률

4. 실험 결과

표 3은 본 실험의 입술, 음성, 음성-입술 합병 인식 결과를 보여준다. 음성-입술 합병 인식이 음성 인식의 저하를 보완해주고 전체적인 인식률을 향상시키는 것을 볼 수 있다.

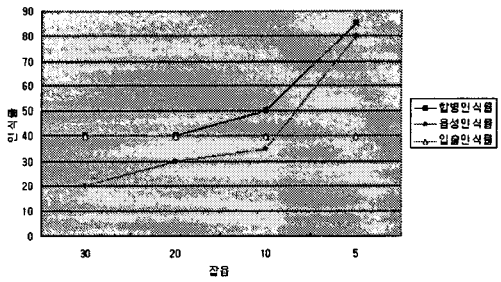


표 3. 입술, 음성, 음성-입술 합병 인식결과

5. 결론 및 향후 과제

본 논문에서는 상황에 독립적으로 얼굴 및 입술 영역을 추출하였고 추출한 입술정보로 립리딩 시스템을 구현하였다. 실험 결과 잡음으로 인한 음성 인식률의 저하를 입술 정보와 합병함으로써 보완할 수 있었다. 현재 더 많은 입술 및 음성정보 데이터베이스를 구성하고 있다. 향후 입술정보 인식률을 보다 더 높일 수 있는 방법에 대한 연구가 필요하고 음성 인식 결과와 립리딩 결과를 효과적으로 합병하는 방법에 대한 연구가 필요하다.

[참고문헌]

[1] Tsuhan Chen, "Audiovisual Speech Processing", IEEE Signal Processing Magazine, Volume: 18 Issue: 1, pp.9-21, 2001

[2] 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

[3] M. Lievin, F. Luthon, "Lip features automatic extraction", Image Processing, ICIP 98. Proceedings. 1998 International Conference On, Vol.3, pp.168-172, 1998

[4]"http://www.image,cityu.edu.hk/~ckcheung/thesis/node19.html"

[5] Tsuhan Chen, R.R. Rao, "Audio-visual integration in multimodal communication", Proceedings of the IEEE, Volume: 86 Issue: 5, pp.837-852, 1998