얼굴과 음성 정보를 이용한 바이모달 시스템 설계 및 구현

김명훈*, 이지근*, 정성태*
원광대학교 컴퓨터공학과
e-mail: {bbongi09, lcg74, stjung}@wonkwang.ac.kr

Design and Implementation of Bimodal System using Face and Audio

Myung-Hun Kim, Chi-Geun Lee, Sung-Tae Jung
Dept. of Computer Engineering, Wonkwang University

요 약

최근 들어 바이모달 인식에 관한 연구가 활발히 진행되고 있다. 본 논문에서는 음성과 얼굴을 이용하여 바이모달 시스템을 구현하였다. 얼굴인식은 객체 분류 기법인 SVM을 이용하여 얼굴을 검출 및 인식하였으며, 음성인식은 HMM을 이용하여 음성인식을 하였다. 각기 인식된 결과에 대해 합성을 통하여 잡음에 의해 낮아지는 음성 인식률을 얼굴 인식과 같이 사용함으로서, 전체적인 인식률 향상을 볼 수 있다.

1. 서론

최근 인체의 인식에 대한 연구가 활발히 진행되고 있다. 인식 기술에는 지문, 홍채, 얼굴, 음성 등 각각 독립적으로 수행되어 왔다. 이러한 방식은 각 생체 정보들이 가지는 문제점으로 인해 인식률의 저하를 가져온다. 따라서 이러한 단점을 보완하기 위해 복수개의 생체정보를 이용하는 바이모달 연구가 진행되고 있다. [1,2]

본 논문에서는 이 중에서 거부감이 없는 인식방법인 얼굴인식과 음성인식의 합성 방법을 사용하였다. 얼굴인식 방법으로는 얼굴 영역의 파라미터를 구하기 위해 주성분분석을 사용하였고, 학습과 인식 방법으로는 SVM(Support Vector Machine)을 사용하였다. 또한, 음성인식을 하기 위해 HMM(Hidden Markov Model)을 사용하였다. 합성 방법으로는 각각 독립적으로 얼굴과 음성을 인식 한 다음 인식 결과를 합성하는 방법을 사용하였다.[3]

이 논문은 2005년도 교육인적자원부 지방연구중심대학육성사업 헬스케어기술개발사업단의 지원에 의하여 연구되었음

실험 결과 잡음에 의한 음성의 인식률 저하를 얼굴 인식과 합성함으로서, 전체적인 인식률 향상을 볼 수 있다.

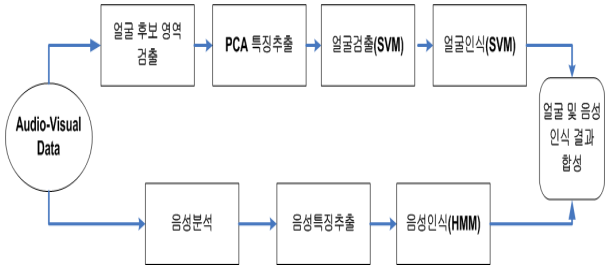

그림 2 바이모달 시스템 구조

2. 얼굴 검출 및 인식

본 논문에서는 얼굴을 검출하기 위해 이미지 내, 간단한 특징정보를 추출하고, Cascade 구조의 Adaboost 방법을 이용하여 얼굴 후보 영역을 검출하였다. 또한 얼굴을 검증하기 위하여 주성분 분석을 통하여 얼굴 특징을 축소하여, SVM을 통하여 얼굴을 검증하였다.

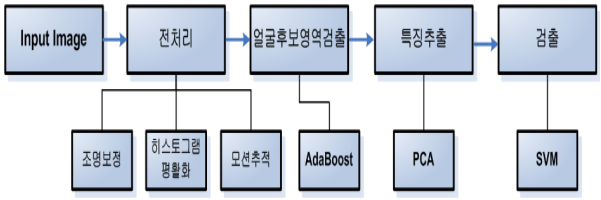

그림 3 얼굴검출 시스템

2.1 얼굴 검출 및 검증

2.1.1 얼굴 후보 영역 검출

A. 전처리 단계

본 논문에서는 전처리 단계로 조명 보정, 히스토그램 평활화, 모션 추적 과정을 적용하였다.

B. Harr-like 특징과 인티그럴 이미지

얼굴 후보 영역은 간단한 특징의 집합으로부터 검출 될 수 있다. 그림 3은 얼굴 영역 내 특징을 보여주고 있다.

본 논문에서는 얼굴 검출을 위한 특징으로 Papageoriou et al 에 의해 제안된 간단하면서도 연산이 빠른 Harr-like 특징을 사용하였다.[4] Harr-like 특징은 인티그럴 이미지를 이용하여 빠르게 연산할 수 있다. 입력 영상에 윈도우를 이동시켜 가면서 특징을 추출하며, 각 모형에서 얻어지는 특징의 수는 윈도우의 크기 및 그 수에 따라 달라진다.

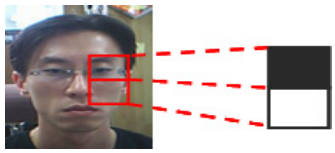

그림 4 얼굴 영역 내 특징

C. 얼굴 후보 영역 검출

본 논문에는 얼굴 후보 영역을 검출하기 위해 Adaptive Boosting(AdaBoost) 기법과 Cascade 구조를 이용하였다. AdaBoost 알고리즘은 Freund와 Schapire에 의해 소개되었다.[5] AdaBoost 알고리즘은 약한 분류기를 결합하여 강한 분류기를 생성하는 방법이다.[6] 약한 학습 알고리즘은 학습 영상 중 얼굴 영상을 가장 잘 구분할 수 있는 특징을 학습하며, 각각의 특징에 대해 정의된 약한 분류기가 가장 낮은 오류 비율을 달성하도록 한다.

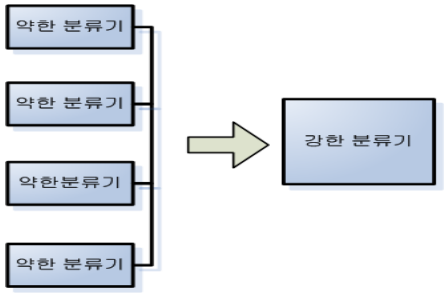

그림 5 강한 분류기 생성

Cascade 구조는 계산 시간을 줄이고 얼굴 검출결과는 향상되게 한다. 처음 스테이지에서 비얼굴 영역을 제거하고, 순차적으로 뒤로 갈수록 얼굴을 검출하게 된다.

2.1.2 얼굴 검증

본 논문에서는 얼굴을 검증하기 위하여 SVM 분류기를 이용하였다. SVM 분류를 하기위해 얼굴 후보 영역의 특징 계수를 줄이는 방법으로 주성분 분석(Principal Component Analysis, PCA)을 사용했다. PCA를 통해 얼굴 후보 영역의 특징 벡터 계수를 줄였다.[7]

SVM은 이진 분류기로서, 얼굴과 비얼굴을 분류한다.[8] SVM 학습에서는 19×19 크기의 MIT 데이터베이스[9]를 PCA로 학습된 이미지를 사용하였고, 32× 32 크기의 정규화 된 BioID 얼굴 데이터를 이용하여 검증하였다.[10]

A. 얼굴 특징 추출

본 논문에서 얼굴 검증을 실험하기 위해, PCA와 SVM을 이용하였다. 즉, 훈련 데이터를 19X19=361의 크기로 정규화 하였고, 마스크를 적용하였다. 이중, 훈련 데이터에서 비얼굴 영역을 제거 시키면 벡터의 크기는 301개가 된다. 또한, PCA를 통해 특징 벡터의 크기를 줄이게 된다.

B. SVM(Support Vector Machine)

SVM은 이진 분류를 하기위해 최적의 하이퍼분리면을 찾게 된다. 그림 5에서 A범주와 B범주를 구분하기 위한 하이퍼분리면은 무수히 많다. 그러나 두 범주간의 점들의 거리를 최대화되도록 학습을 시키게 되면, OSH(Optimum Separating Hyperplane)은 유일한 해로 존재하게 된다.

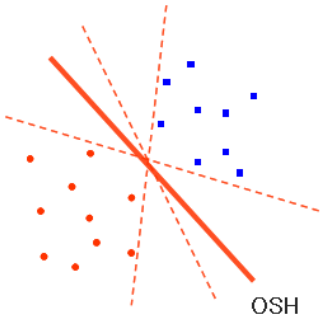

그림 6 하이퍼분리면

2.2. 얼굴 인식

본 논문에서는 검출된 얼굴영역을 인식하기 위해 멀티 SVM을 사용하였다. 즉, 모든 클래스에 대해 가능한 모든 쌍을 구성하여, 각각의 이진 SVM을 만든다. 각각의 이진 SVM 결과로 최대의 값을 선택하게 되고, 해당 클래스로 분류하게 된다. 그림 6은 SVM을 이용한 멀티클래스 분류를 보여주고 있다.[11]

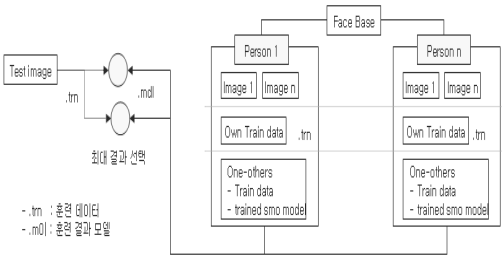

그림 7 SVM을 이용한 멀티클래스 분류

3. 음성인식

3.1 음성 특징 추출

음성의 특징을 추출하는 방법은 화자의 음성에 대한 주파수 특성을 이용한다. 본 논문에서는 MFCC(Mel Frequency Cepstrum Coefficient)을 이용하여 특징을 추출했다. 이 방식은 인간의 청력이 일정 주파수 이상이 되면, 로그 형태로 되는 주파수 대역별 에너지 측정방법이다.

3.2 음성 인식

음성 인식을 하기 위해 Hidden Markov Model(HMM)을 사용하였다. HMM은 다중 확률구조를 갖는 프로세스 모델에 적합하다. HMM은 특정 사건의 의존 관계를 확률적으로 모델링 할 수 있다. HMM은 Markov 체인에 은닉 상태를 추가하여 구성하게 된다. 본 논문에서는 그림 7과 같이 left-to-right 모델을 사용했다.

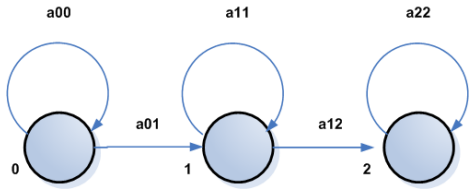

그림 8 left-to-right 모델

4. 바이모달 시스템

본 논문에서는 얼굴 인식률과 음성 인식률을 합성하는 늦은 통합(late integration) 방법을 사용하였다. 각각의 얼굴 인식과 음성 인식의 결과 값을 합성하여, 잡음에 따른 음성 인식률 저하를 보상하였다. 얼굴인식은 각각의 영상에 대해 인식된 결과에서 가장 많이 투표된 값을 선택하였다.

5. 실험결과

5.1 얼굴 검출 및 인식

얼굴 검출 및 인식 실험은 총 20명의 사람을 개인당 10장씩 영상을 추출하여 그 중 7장을 가지고 학습시켰다. 테스트는 개인당 3장씩 추출하여 테스트하였다.

본 논문에서 제안하는 얼굴 검출 방법은 표 1의 실험 결과와 같이 AdaBoost+PCA(60개)+SVM을 사용하여 얼굴 영역 검증을 수행하였다.

구분	AdaBoost + PCA(60) + SVM
검출률(%)	96.7
오검출률(%)	0.32
검출시간(ms)	15.1

표 1 얼굴 검출 실험 결과

또한 검출된 얼굴영역에서 PCA와 SVM을 통해 얼굴인식 실험을 하였다.

구분	커널	SVM + PCA
	특징계수	60
	이미지 크기	32×32
RR	96.0	
FAR	1.1	
FRR	2.0	
평균 인식시간(ms)		27.5

표 2 얼굴인식 실험결과

5.2 음성 인식

본 논문에서는 음성인식 실험은 HTK(Hidden Markov Model Toolkit)을 이용하였다. 음성인식을

하기 위해 20명씩 10번 반복 발음하여 7번은 학습 데이터로 3번은 테스트 데이터로 이용하였다. 음성 데이터는 기본적으로 공간적 제약을 두지 않는 자연잡음이 들어 있는 상태에서 실험을 하였으며, 인위적인 잡음(Brown noise)을 점차 추가하면서 실험을 하였다.

잡음(Brown noise)	인식률(%)
noise-0.0	98.33
noise-0.5	95.00
noise-1.0	95.00
noise-1.5	85.00
noise-2.0	75.00
noise-2.5	43.33
noise-3.0	20.00

표 3 음성인식 실험결과

5.3 바이모달 시스템

본 실험에서는 바이모달 시스템은 음성과 얼굴 인식 결과를 합성하는 방법을 사용하였다. 표 4는 얼굴과 음성 합성에 의한 방법은 단일 인식보다 인식률의 향상을 보여주고 있다.

잡음(Brown noise)	바이모달 인식률(%)
noise-0.0	100%
noise-0.5	100%
noise-1.0	100%
noise-1.5	97.2%
noise-2.0	96.00%
noise-2.5	96.00%
noise-3.0	96.00%

표 4 바이모달 실험결과

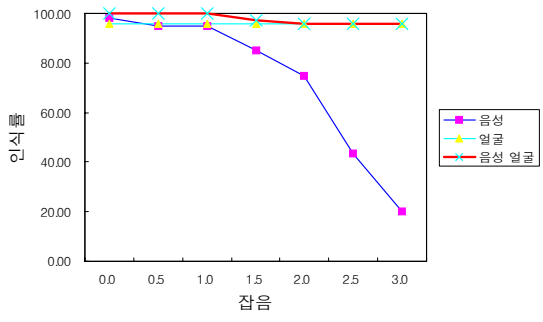

그림 9 인식결과

6. 결론 및 향후 과제

본 논문에서는 얼굴 인식 및 음성 인식을 각기 따로 추출 하였고, 추출한 정보로부터 바이모달 시스템을 구축하였다. 잡음으로 인해 음성 인식률이 저하됨을 볼 수 있는데, 얼굴 인식과 합성함으로서, 인식률을 보완할 수 있었다. 향후 각각의 인식률에 대한 보완이 더 필요하고, 단순한 합성이 아닌 효과적으로 합성하는 방법에 대한 연구가 필요하다.

참고문헌

[1] M. N. Kannak, Q. Zhi, A. D. Cheok, and K. C. Chung "Audio-Visual Modeling For Bimodal Speech Recognition", *2001 IEEE International Conference*, Vol. 1, 7-10 Oct. 2001 pp.181 - 186

[2] S. Meng, and Y. Zhang "A Method Visual Speech Feature Area Localization", *IEEE Neural networks and Sigral Processing*, Dec. 2003

[3] N. A. Fox, R. Gross, P. D. Chalzal, J. F. Cohn, and R. B. Reilly, "Person Identification Using Automatic Integrated of Speech, and Face Experts", *WBMA'03*, Nov. 2003

[4] R. Lienhart, and J. Maydt, "An extended set of Harr-like features for rapid object detection," *Image Proceedings. 2002 International Conference on*, vol. 1, 22-25, pp.I-900 - I-903, Sept. 2002

[5]Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[6] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection," *DAGM'03, 25th Pattern Recognition Symposium*, Madgeburg, Germany, pp.297-304, Sep. 2003.

[7] Jian Yang, and Jing-ju Yang, "Why can LDA be perfomed In PCA transformed space?", *Patter Recognition 36,* pp.563-566, 2003.

[8] V. Vapnik, "The Nature of Statistical Learning Theory,"h Springer-verlag, New York, 1995.

[9] Center for Biological and Computational Learning at MIT, "MIT CBCL DATASETS," hhttp://cbcl.mit.edu/cbcl/software-datasets, 2003.

[10] HumanScan AG, "BioID Face Database", http://www.humanscan.de/support/downloads/facedb.php, 2003

[11] 이호근, 김명훈, 이지근, 정성태 "SVM-SMO와 Pan-tilt 웹카메라를 이용한 실시간 얼굴 추적과 얼굴인식", 한국정보과학회, 제 31권 제 2호, 2004