생체인식 시스템을 위한 통계적 식별 방법

이관용[0] 박혜영*
한국싸이버대학교 컴퓨터정보통신학부
kylee[0]@mail.kcu.or.kr

*일본 이화학연구소 뇌과학연구센터 뇌수리과학연구팀
hypark@brain.riken.go.jp

A Statistical Verification Method for Biometrics Systems

Kwanyong Lee[0] Hyeyoung Park*
Div. of Computer, Information and Communication, Korea Cyber University
*Lab. for Mathematical Neuroscience, RIKEN BSI, Japan

요 약

생체인식 시스템은 개인의 물리적/행동적 특성을 측정하여 신원을 확인하기 위한 시스템이다. 이러한 시스템에서 사용되는 특징들은 잡음 등에 의해서 쉽게 영향을 받기 때문에 매우 많은 변형들이 존재하고, 따라서 변형된 특징들을 효과적으로 다루기 위해 다양한 기계학습 방법들이 사용되고 있다. 그런데, 기존의 자료주도적인 방법들을 특정 생체인식 시스템에 적용하기 위해서는 시스템에 등록할 각 사람들로부터 충분히 많은 데이터를 획득해야하는 어려움을 겪게 된다. 또한 시스템에 미등록된 사람의 데이터가 제시될 가능성 등, 무한한 수의 변형이 존재하는 문제점을 갖고 있다. 이러한 문제점들로 인해 데이터의 분포 특성을 분석하고 예측하는 것이 어렵다. 생체인식 시스템의 이러한 고유의 문제점을 극복하기 위해서는 새로운 효율적인 식별 및 검증 방법을 필요하다. 따라서, 본 논문에서는 통계적 가설 검증 이론에 기초한 간단한 방법을 제안하고, 실세계 데이터에 대한 실험을 통해 제안한 방법의 가능성을 확인한다.

1. 서 론

컴퓨터 비전과 패턴 인식 기술에 기초한 개인 신원 확인을 위한 새로운 보안 응용 기술로서 각광을 받고 있는 것이 바로 생체인식 기술이다. 생체인식 기술에서 사용되는 개인 특성으로는, 물리적 특성으로 지문, 장문, 홍채, 망막, 얼굴, 정맥 등이 있고, 행동적 특성으로는 서명, 음성, 타이핑 등이 있다. 개인의 생체 특성은 절도나 누출에 의해 전달될 수 없으며 변경되거나 분실할 위험성이 없으므로 패스워드나 ID 카드와 같은 기존의 방법보다 신뢰성이 높은 신원 확인은 물론이고 보안 침해를 누가 행했는지 추적이 가능해지는 등 감사 기능이 완벽하게 구축될 수 있다는 장점이 있다.

실제적으로 신뢰성이 높은 생체인식 시스템을 구축하기 위해서 기계학습, 인공지능, 신호처리 등과 관련된 다양한 정교한 방법들이 사용되어 왔다[1-5]. 하지만, 대부분의 연구들은 주어진 특정한 생체 데이터의 특성에 적합하도록 시스템을 최적화하는 것을 목적으로, 주어진 입력 데이터에 대해 정교한 특징을 추출하는 문제와 같은 제한적인 부분에 중점을 두고 진행되어 왔다.

새로운 데이터가 학습 시스템에 제시되면 등록된 데이터와의 유사도를 측정하여 승인/기각 여부를 결정하게 된다. 이때 유클리디안 거리가 보편적으로 사용되는데, 이것은 너무 단순하고 데이터 분포 특성을 고려하지 않는 단점을 갖고 있다. 이것의 개선된 방법으로 정규화된 유클리디안 거리나 마하라노비스 거리 등이 사용된다[5]. 하지만, 생체인식 시스템에서 각 사람이 나타나는 클래스에 속하는 데이터의 수는 매우 작기 때문에 각 클래스에서 데이터 분포에 관한 정확한 정보를 얻기가 어렵다. 결국 이러한 문제는 유사도 측정에 있어서 정확도를 저해시키는 심각한 요인이 될 수 있다. 이러한 문제를 극복하기 위해서 좀 더 정교한 분류방법, 예를 들어 신경회로망 또는 커널 머쉰 등이 적용되고 있다[3]. 이런 종류의 기계학습 시스템에서도 좋은 성능을 얻기 위해서는 충분히 많은 수의 데이터가 필요하게 된다. 하지만 생체인식 시스템의 경우에는 충분한 데이터를 얻기가 결코 쉽지 않게 된다. 더욱이 일반적인 분류 문제와 생체인식에서의 분류 문제와의 큰 차이점은 생체인식 시스템에서는 어떤 학습 클래스에도 속하지 않는 새로운 데이터의 출현 가능성이 매우 높다는 것이다. 또한 이러한 데이터들은 거의 무한한 수의 변형을 가질 수 있기 때문에, 보통의 분류 방법으로 이들 데이터에 대한 성능 을 보장하는 것은 매우 힘들다. 이러한 생체인식 문제의 고유한 특성을 고려하여, 생체인식에서의 검증과 식별에 좀 더 적절한 방법의 개발이 필요하게 되었다.

본 논문에서는, 생체인식 문제에 있어서 핵심적이며 잡음에 강인한 데이터분포정보를 추출하기 위한 전략을 제안하고, 그것을 바탕으로 유사도 측정과 검증 임계치를 결정하는 간단한 방법을 제시한다. 본 방법에 의해 추출되는 데이터분포정보는 각 사람에 해당하는 클래스의 분포에 종속적이 아니라 전체 데이터 집합에 종속적이기 때문에, 개인별 데이터 수가 많지 않은 경우에도 신뢰성 높은 시스템 설계를 기대할 수 있다.

유사도 측정과 더불어 승인 여부를 결정하는 적절한 임계치의 설정은 성능에 영향을 미치는 주요한 요소가

본 연구는 학술진흥재단(2001-003-E00234)에 의해 지원받았음

된다. 기본적인 방법은 학습데이터를 이용해서 오인식율(FAR)과 오거부율(FRR)을 동시에 최소화시킬 수 있는 경계를 결정하는 것이다. 이 경우에 임계치는 학습 데이터 집합의 분포에 절대적으로 의존하게 된다. 따라서, 학습데이터가 많지 않은 경우에는 잡음과 변형을 포함한 새로운 입력 데이터에 대해서 좋은 성능을 보장할 수가 없게 된다. 본 방법에서 제안하는 데이터분포정보의 추출 방법은 생체인식 문제의 기본 특성을 바탕으로 검증에 기본이 되는 확률분포 함수의 설정이 가능하며, 이를 통해 임계치 결정이 가능하다. 이것에 대해서는 3장에서 다루도록 하겠다.

2. 유사도 측정 방법

데이터의 통계적 특성에 기초한 유사도 평가 방법을 다루기 위해서, 우선 데이터를 D차원의 확률변수 $\boldsymbol{x}=(x_1,\cdots,x_D)$로 표현한다. 전체 데이터의 집합, $X=\{\boldsymbol{x}_n|n=1,\cdots,N\}$는 부분집합 $X_k=\{\boldsymbol{x}_{n_k}\}_{n_k=1,\cdots,N_k}$, $(k=1,\cdots,K)$로 분해될 수 있다. 여기서 각 부분집합 X_k는 k라는 한 사람으로부터 획득한 데이터가 나타내는 하나의 클래스를 의미한다. 정규화된 유클리디안 거리나 단순화된 마하라노비스 거리와 같은 기존의 방법에서는 각 클래스의 평균과 분산이 유사도 측정에 사용된다. 하지만, 생체인식 시스템의 각 클래스의 데이터 수가 적기 때문에 데이터로부터 의미 있는 통계 특성을 얻기에는 불충분하고, 시스템의 성능 저하의 원인이 된다.

각 클래스 당 데이터 개수의 불충분으로 야기될 수 있는 문제를 극복하기 위해서, 본 방법에서는 식(1)과 같이 정의되는 새로운 확률변수 $\boldsymbol{y}$를 도입한다.

$$\boldsymbol{y}=\boldsymbol{x}-\boldsymbol{x}' \tag{1}$$

여기서 $\boldsymbol{x}$와 $\boldsymbol{x}'$가 같은 사람으로부터 얻은 것이라면, 두 데이터간의 차이는 측정시의 잡음에 기인하는 것으로 볼 수 있다. 따라서 그 잡음의 확률분포가 $p(\boldsymbol{y})$가 되고, 이를 생체인식 시스템 설계에 사용할 수 있다. $\boldsymbol{x}$의 확률분포 대신 $\boldsymbol{y}$를 사용함으로써 얻는 중요한 장점은 $\boldsymbol{y}$가 클래스 X_k의 분포에 의존하지 않는다는 것이다. 왜냐하면 잡음은 각 사람의 특성이 아니라 사용된 생체특징을 측정하는 시스템 자체의 특성이기 때문이다. 또한 생성되는 $\boldsymbol{y}$의 샘플 수는 $\boldsymbol{x}$의 샘플 수보다 많으므로 보다 정확한 분포정보를 얻을 수 있다.

본 논문에서는 간단하면서 일반적인 경우로, $p(\boldsymbol{y})$가 가우시안 분포를 따른다고 가정한다. 따라서 $\boldsymbol{y}$의 평균 $\boldsymbol{\mu}$와 공분산 $\boldsymbol{\Sigma}$를 추정함으로써 $p(\boldsymbol{y})$가 정해진다. 본 논문에서는 문제를 좀더 간단히 하여 $\boldsymbol{x}$의 각 원소인 $x_d\,(d=1,\cdots,D)$가 서로 독립적이라고 가정하여, 그 표준편차 $\sigma_d\ (d=1,\cdots,D)$만을 추정한다.

$\boldsymbol{y}$의 표준편차는 식(2)와 같이 추정될 수 있다.

$$\sigma_d=\sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_d^m-\mu_d)^2} \tag{2}$$

이때 평균 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_D)$는 다음 식으로 얻어진다.

$$\mu_d=\frac{1}{M}\sum_{m=1}^{M}y_d^m \tag{3}$$

y_d^m은 $\boldsymbol{y}$의 집합 Y에서 m번째 원소 $\boldsymbol{y}^m$의 d번째 요소를 의미한다.

유사도 측정 함수 $s(\boldsymbol{x},\boldsymbol{x}')$는 정규화된 유클리디안 거리와 유사한 형식으로 식(4)와 같이 정의될 수 있다.

$$s(\boldsymbol{x},\boldsymbol{x}')=\sum_{d=1}^{M}\frac{(x_d-x'_d-\mu_d)^2}{\sigma_d^2} \tag{4}$$

그러나, 여기서 사용된 $\boldsymbol{\mu}$와 $\boldsymbol{\sigma}$는 정규화된 유클리디안 거리 또는 단순화된 마하라노비스 거리와 같은 기존의 방법들에서 사용되는 값들과는 다르다. 집합 Y의 데이터의 개수는 각 부분집합 X_k의 개수보다 훨씬 많기 때문에 여기서 이용되는 추정치는 보다 정확하고 잡음에 강한 특성을 갖게 된다.

3. 검증 임계치

새로운 입력 데이터에 대한 검증을 수행할 때 필요한 임계치는 유사도 값의 분포를 고려하여 결정된다. 식(4)를 살펴보면, 두 데이터 $\boldsymbol{x}$, $\boldsymbol{x}'$가 같은 부분집합 X_k에 속하면 $\frac{(x_d-x'_d-\mu_d)}{\sigma_d}$항은 $p(\boldsymbol{y})$에 대한 가정으로부터 표준정규분포를 따르게 된다. 따라서 유사도 측정 함수 $s(\boldsymbol{x},\boldsymbol{x}')$는 자유도가 D인 χ^2-분포 $p_{\chi^2}(s;D)$를 따르는 확률변수로 생각할 수 있다.

이러한 사실로부터 검증과정에 유도비검증(likelihood ratio test)를 쉽게 적용할 수 있음을 알 수 있다. 검증의 기본 과정은 다음과 같다 (그림 1 참조).

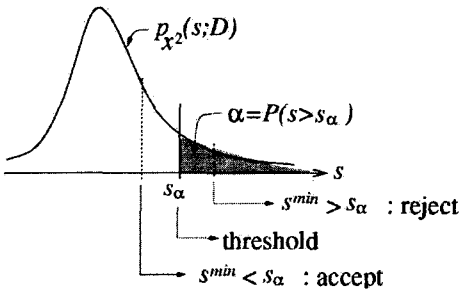

그림 1. 유도비 검증의 개념도

① 새로운 데이터 $\boldsymbol{x}^{new}$에 대해서, 집합 X에 있는 모든 데이터 $\boldsymbol{x}^n$에 대한 $s(\boldsymbol{x}^{new},\boldsymbol{x}^n)$를 계산하고, 식(5)와 식(6)에 의해서 최소값 s^{min}과 그에 대응되는 가장 유사한 데이터 $\boldsymbol{x}^{min}$를 찾는다.

$$\boldsymbol{x}^{min}=\mathrm{argmin}\{s(\boldsymbol{x}^{new},\boldsymbol{x}^n)|\,\boldsymbol{x}^n\in X\} \tag{5}$$

$$s^{min} = s(\boldsymbol{x}^{new}, \boldsymbol{x}^{min}) \quad (6)$$

② 확률 $P(s > s^{min})$를 계산한다.

③ 계산된 확률 $P(s > s^{min})$이 미리 정한 유의수준 α보다 크면 해당 데이터를 기각하고, 아니면 승인한다.

여기서 임계치 역할을 하는 α는 검증시 허용가능한 오차율을 의미하는 것으로, 사용자에 의해 결정된다. 또한 $P(s > s^{min})$는 $s(\boldsymbol{x}^{new}, \boldsymbol{x}^{n})$의 분포가 정해지면 계산될 수 있다. 특히 여기서는 χ^2-분포를 따르므로, 테이블을 이용해서 쉽게 얻을 수 있다. 그러나 실제의 경우에는, α가 정해지면, $P(s > s_\alpha) = \alpha$가 되는 s_α를 계산해 두고 그것과 s^{min}을 비교하여 검증을 수행한다. 본 논문에서 제안하는 전체 검증 과정을 정리하면 다음과 같다.

① 원하는 오기각율 α를 정하고, 식(7)과 χ^2-분포 테이블을 이용하여 대응되는 임계치 s_α를 찾는다.

$$P_{\chi^2}(s < s_\alpha) = 1 - \exp\{\frac{1}{n}\log\alpha\} \quad (7)$$

② 새로운 $\boldsymbol{x}^{new}$에 대해 s^{min}과 $\boldsymbol{x}^{min}$를 찾는다.

③ $s^{min} < s_\alpha$이면 $\boldsymbol{x}^{new}$가 등록된 것으로 판단, $\boldsymbol{x}^{min}$에 해당하는 사람으로 승인한다.

④ $s^{min} \geq s_\alpha$이면, $\boldsymbol{x}^{new}$를 기각한다.

4. 실험결과

제안한 방법의 가능성을 확인하기 위해, 홍채영상에 대한 실험을 수행하였다. 21명으로부터 375개의 홍채영상 데이터를 획득하여, 이 중 등록된 사람으로 14명을 선정하고 각 사람으로부터 임의로 5개씩, 총 70개의 데이터를 이용하여 시스템을 구축하였다. 등록된 14명에 해당하는 나머지 190개의 데이터는 오거부율(FRR) 평가를 위해, 등록되지 않는 나머지 7명의 115개의 데이터는 시스템의 오인식율(FAR) 평가를 위해 사용하였다.

320×240 크기의 홍채영상 데이터를 홍채영역 추출 및 좌표계 변환 등의 필요한 전처리 과정을 거쳐 얻은 그림2와 같은 225×32 크기의 영상을 사용하였다. 먼저 주성분분석 기법[6]을 적용하여 70차원을 특징벡터를 구성하고, 이를 이용해서 집합 Y를 구성하였다.

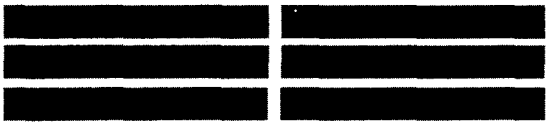

그림 3. 직교좌표로 변환된 홍채영역 영상

제안한 유사도 측정 방법의 성능을 평가하기 위해, 테스트 데이터 집합의 등록된 그룹의 데이터를 이용하여 등록된 사람에 대한 식별(분류) 실험을 수행하였다. 각 테스트 데이터에 대해 모든 등록된 사람과의 유사도를 계산하여 $\boldsymbol{x}^{min}$을 찾아 테스트 데이터를 $\boldsymbol{x}^{min}$이 속한 클래스로 지정하게 된다. 제안한 방법과 단순화된 마하라노비스 거리와 비교한 결과가 표1에 제시되어 있다.

표 1. 등록된 데이터에 대한 클래스 식별율

유사도 평가 방법	클래스 분류율
단순화된 마하라노비스 거리	83.68 %
제안한 방법	98.95 %

검증 평가를 위해서는 두 개의 임계치를 사용하였다. 즉, $\alpha = 0.05$에 해당하는 47.89와 $\alpha = 0.1$에 해당하는 46.76를 식(7)을 통해 얻어서 사용하였다. 검증과정에 대한 성능평가 실험은 190개의 등록된 데이터와 115개의 비등록된 데이터 모두에 대해서 수행하였으며, 그 결과가 표2와 같다.

표 2. 테스트 데이터에 대한 성능

	임계치 = 47.89	임계치 = 46.76
오거부율 (FRR)	1.05 %	1.58 %
오인식율 (FAR)	8.70 %	6.96 %

5. 결론

본 논문에서는 생체인식에서 사용되는 데이터 집합의 특성을 고려하여 생체인식 시스템을 위한 식별과 검증 과정과 관련된 통계적 프레임워크를 제안하였다. 제안한 방법은 식별 및 검증을 위해 데이터를 처리하는 하나의 표준적인 통계적 방법으로서 의미를 가질 것으로 기대된다. 본 논문에서 새롭게 제안된 확률변수 y의 확률 분포를 추정하기 위해 신경망 등 정교한 학습시스템을 사용함으로써 보다 향상된 성능을 기대할 수 있을 것이다. 또한 이 경우에도 제안한 임계치 결정 방법은 추정된 밀도함수 $p(y)$에 대해 같은 방법으로 적용될 수 있다.

참고문헌

[1] Belhumeur, P., Hespanha, J., and Kriegman, D., "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE trans. on Pattern Recogntion and Machine Intelligence, 19(7), 711-720, 1997.

[2] Boles, W.W., Boashash, B., "A Human Identification Technique Using Images of the Iris and Wavelet Transform", IEEE Trans. on Signal Processing, 46(4), 1185-1188, 1998.

[3] Campbell, W., "A Sequence Kernel and its Applications to Speaker Recognition", Advances in Neural Information Processing Systems, In Press, 2001.

[4] John G. Daugman, "High Confidence Visual Recognition of Persons by a Test of Statistical Independence", IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(11), 1148-1161, 1993.

[5] Kee, G., Byun, Y., Lee, K., and Lee, Y., "Improved Techniques for an Iris Recognition System with High Performance", AI 2001:Advances in Artificial Intelligence, LNAI 2256, 177-188, 2001.

[6] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.