

AI 생체 테스트와 표준화 현황

손영수
어니컴
ysson@onyc.com

최영재
STA테스팅컨설팅
yjchoi@sta.co.kr

스튜어트 리드
STA테스팅컨설팅



안면 인식 시장은 2015년 기준 1,521백만 달러에서 2020년 2,836백만 달러로 증가하여 매년 13.3% 성장하고 있다. 현재 중국(1:1 Verification 1위-여권사진과 실제 인물이 동일인인지 비교), 일본(1:N Identification 1위-A라는 사람을 수많은 데이터를 이용해 식별해 내는 작업), 미국 기업(API 서비스-누구나 사용할 수 있는 서비스, SDK로 제공)들이 시장을 리딩하고 있다.

대한민국의 경우 선두 그룹과 기술력에서 많은 격차가 존재한다(NIST 미국 국립 기술 표준 연구소의 FRVR 안면인식 벤더 테스트-2020년 12월 기준 169개의 기업 중 28위). 이에 이러한 격차를 줄이기 위해, 과기정통부, 법

무부가 협력하여 인천공항의 안면인식 데이터를 학습 데이터로 국내 업체에 제공을 하고 학습, 평가를 돕고 있다.

해외 동향

먼저 해외 동향을 다루도록 하겠다. 얼굴 인식 테스트에 사실상 표준이라고 불리는 테스트로 FRVT(Face Recognition Vendor Test-얼굴인식 벤더 테스트)가 있다. 매년 미국 국립표준기술연구소(NIST) 주최로 진행되는 안면인식 벤더 테스트이며, CCTV와 같이 임베디드 환

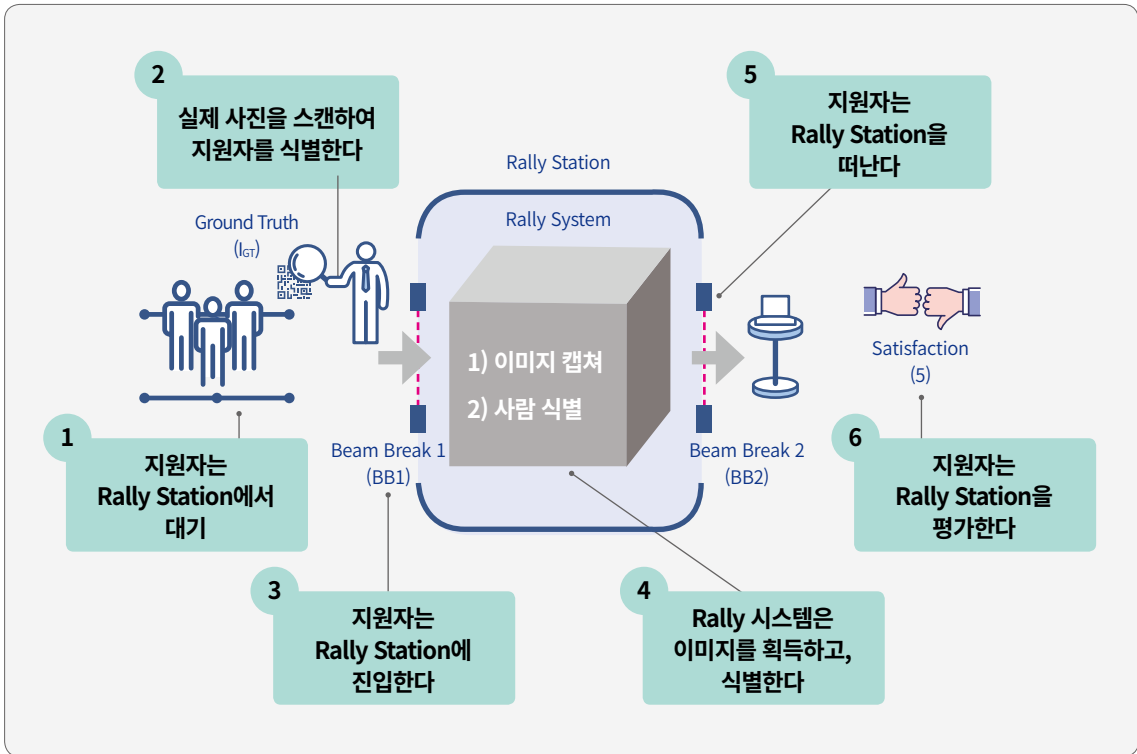
평가 유형	설명
FRVT 1:1	• 얼굴 인식 알고리즘 평가 • 몇 가지 다른 데이터 세트에서 FNMR(Fal-se Non-Match Rate)로 측정된 최고 1:1 성능의 알고리즘 평가(여권심사에 사용됨)
FRVT 1:N	• FRVT 1:1에서 비자 사진, 범죄자 사진 인식율이 높은 업체만 심사 참여 가능 • 다수의 아이덴티티가 등록된 갤러리를 검색하는 일 대 다 얼굴 인식 알고리즘의 정확도와 속도 향상을 측정
FRVT MoRPH	• 얼굴 탐지 알고리즘의 지속적인 평가 • 프로토 타입 얼굴 탐지 기술에 대한 지속적인 독립적 테스트를 제공 1. 스틸 사진의(두 명의 얼굴을 섞은) 모핑을 감지하는 알고리즘 개발 2. 모핑에 대한 얼굴 인식 알고리즘 저항
FRVT Quality Assesment	• 얼굴 이미지 품질 평가 • 단일 이미지에서 품질 스칼라의 알고리즘 출력을 평가
Face In Video Evaluation	• 실시간 스트리밍이 아닌 1분 내외의 녹화한 비디오를 입력으로 주고 1분 동안 다 읽은 후 결과를 반환 • DB에 등록된 사람 중에 유사도가 높은 순으로 여러 후보와 해당하는 유사도를 출력
FRVT Face Mask Effect	전 세계 COVID-19 사태 이후, 2020년 신설된 테스트로 마스크를 쓴 대상에 대해서 1:1 테스트를 진행

경에서 성능을 측정하는 테스트다. 테스트 플랫폼으로 리눅스와 C 언어로만 테스트가 가능하다.

또 다른 대표적인 테스트로는 미국 국무부에서 제공하는 MdTF(Maryland Test Facility) 테스트가 있다. 미국무부는 MdTF(Maryland Test Facility)을 통해 2018년부터 생체인식 기술 랠리(Biometric Technology Rally)를 개최하고 있다. 다양한 업체 참여를 위해 테스

트 인터페이스를 Restful API(언어 제약이 없는 표준화된 웹 인터페이스)로 제공하고 있다. 안면인식, 지문 인식, 홍채 인식 3분야를 다 측정한다(2019년 텐센트가 1위). FRVT와 달리 이미지가 아닌 실제 사람이 참여하여 테스트가 진행되며, 정확도, 응답시간, 참여한 사람의 만족도를 측정한다. 크게 유효성, 효율성, 만족도를 지표로 삼고 있다.

[그림 1] 미국 국토안보부의 MdTF 테스트 진행 방법



평가 유형	설명
이미지 획득 시스템 평가	• RESTful API(HTTP)로 통신 • 평가 항목 - Failure to Acquire Rate - 이미지 획득 시간: 평균 10초 이내 충족해야함 - True Identification Rate : 보유한 여러 매칭 알고리즘 수행 후 95% 이상 식별해 내는가? - 획득한 이미지가 여러 매칭 알고리즘에서 일관성 있게 동작 하는가 - 사용자들의 만족도를 평가에 반영함
매칭 알고리즘 평가	• RESTful API HTTP 서버 기능이 있는 docker container를 제공 • 이미지를 입력으로 받아 이미지 성질을 분석해 저장한 템플릿 출력 • 평가 항목 - 제대로 사람을 찾아낸 비율(True Identification Rate)과 잘못 매칭한 비율(False Matching Rate) 1:10,000, 1:100,000, 1:1,000,000 - 충족 조건: FMR 1:10,000에서 정확도 95% 이상 - 각각 이미지 획득 시스템으로부터 얻어진 이미지를 사용하여 측정 - 이미지 획득 시스템에 민감하지 않게 잘 작동 하는가 - 인식을 성공하든, 실패하든 최대 20초, 평균 5초 이내 수행

국내 동향

국내의 대표적인 생체인식 테스트 사례로는 2019년, 2020년 과학 기술 정보통신부와 법무부가 범부처 협력 사업으로 진행한 AI 식별추적 시스템 실증 사업이 있다. 어니컴과 STA테스팅컨설팅이 법무부 공항 내 실증을 위한 학습 및 검증환경을 상암동 누리꿈스퀘어에 구축하여 제공하고 있으며 다음과 같은 테스트를 진행해 왔다.

한국 인터넷 진흥원이 FRVT 테스트를 준용한 얼굴인식 알고리즘 성능시험을 제공하고 있다. 동영상 테스트도 제공하고 있으며, 1분 파일 영상을 다 읽은 후 결과를 전송하는 형태로 되어 있다.

국내의 생체 시스템 개발 또는 상용화 동향

인천국제공항에서는 여권과 탑승권 없이 간단한 얼굴 인식만으로 출국 가능한 시스템을 구축했다. 개인의 인식

정보로 탑승권, 여권 등 종이 서류를 대체하는 ‘스마트패스’ 서비스가 시범 도입되었으며, 지문과 얼굴 등의 인식 정보를 활용하고 있다. 이로 인해 출입국 심사가 평균 1인당 15초로 줄고, 92~94%의 정확도를 가지고 있다.

미국 샌프란시스코 대학교는 기숙사 보안을 강화하기 위해 얼굴인식 인공지능 시스템을 도입했다. 얼굴인식 인공지능 출입 시스템으로 허가된 출입자인지 판단하고 기숙사 출입을 허가하기 때문에 외부인 출입으로 인한 부정적인 사건이 줄어들게 되었다.

자동차 렌트회사인 허츠(Hertz)사는 사전 등록된 고객들을 위해 허츠 패스트 레인이라는 생체인식 서비스를 제공하고 있으며, 이는 렌탈 프로세스 시간을 75% 단축시켰다. 허츠는 이 서비스를 2019년 말까지 40개 이상의 주요 공항에 출시하였다.

평가 유형	설명
1:1	두 장의 사진을 입력으로 받아 사진 속의 두 사람이 같은 사람인지 다른 사람인지를 판단하는 능력을 평가
1:N	한 장의 사진을 입력으로 받아 사전에 등록된 사람 중 누구인지 혹은 등록되지 않은 사람인지를 판단하는 능력을 평가
실시간 동영상 1:N	자유로이 움직이는 사람들의 얼굴을 인식하고 미리 등록된 사람 중 누구인지 혹은 등록되지 않은 사람인지를 판단하는 능력을 평가
실시간 동영상 이상행동 파악	○ 큰 영역에서는 4대의 카메라로 돌진, 역주행, 물건방치, 2인 감지 ○ 작은 영역에서는 2대의 카메라로 돌진, 역주행, 물건방치, 2인 감지
시스템 안정성 평가	○ 장기간 시스템을 테스트 하여 안정성 테스트를 진행함 ○ 앱 비정상 종료/ 크래시 발생 시 원인 파악이 가능한 코드 레벨의 리포트 전달 ○ Netflix사에서 사용하는 USE 메소드를 활용한 자원 기반 성능 분석 리포트 전달

과기부, 법무부에 적용된

생체인식 시나리오 및 검증 방법

여권사진과 출입국자가 동일인인지 판단하는 1:1 매칭, 출입국자가 DB에 기 등록된 사람들 중 누구인지 판단하기 위한 1:N 매칭기술을 개발하며, 이미지와 동영상 정보를 처리하도록 개발하였다. 동영상 테스트는 실시간 스트리밍 되는 데이터를 기반으로 테스트 하였다.

검증 방법은 이미지 기반의 테스트(동일인 감지, 등록인/비등록인 식별 테스트)와 동영상 기반의 테스트(동일인 감지, 등록인/비등록인 식별, 실시간 이상행동 감지, 실시간 이상행동 감지 및 행위자 식별)로 나뉘며, 다음과 같이 시스템을 구축하였다.

사진 검증은 실증서버가 평가 서버에게 사진을 요청한다. 평가 서버는 요청된 사진을 실증서버가 동시에 다 운받을 수 있도록 제공한다.

동영상 검증은 실증서버가 평가 서버에게 동영상을 요청한다. 평가서버는 미디어 서버가 동영상을 송출할 수 있도록 동영상 재생을 요청한다. 미디어 서버는 동영상

재생 요청을 받으면 RTMP 프로토콜을 이용해 정해진 URL로 동영상 송출을 시작한다. 많은 스트리밍 프로토콜 중에 RTMP 프로토콜을 선택한 이유는 실시간 스트리밍 프로토콜 중에 네트워크 속도에 따라 품질이 변경되지 않는 고정 비트레이트를 지원하기 때문이다.

ISO/IEC 생체인식 테스트 표준

안면인식을 포함한 생체인식 시스템은 여러 구성 요소로 이루어진 시스템이기 때문에 그것을 올바르게 검증하기 위해서는 다양한 측면에서 고려할 필요가 있다. ISO와 IEC에서도 관련 분야의 여러 표준을 개발 및 배포하고 있다.

ISO와 IEC의 JTC1(Joint Technical Committee 1) 내에는 여러 개의 전문 위원회(SC, sub-committee)가 존재하며, 각 전문 위원회는 특정 분야 표준을 담당하며 각 전문 위원회 안에는 여러 워킹그룹(WG, Working Group)이 존재한다. [그림 6]은 생체인식 시스템과 관련된 JTC1의 주요 전문 위원회를 보여주고 있다. 좌측의 청

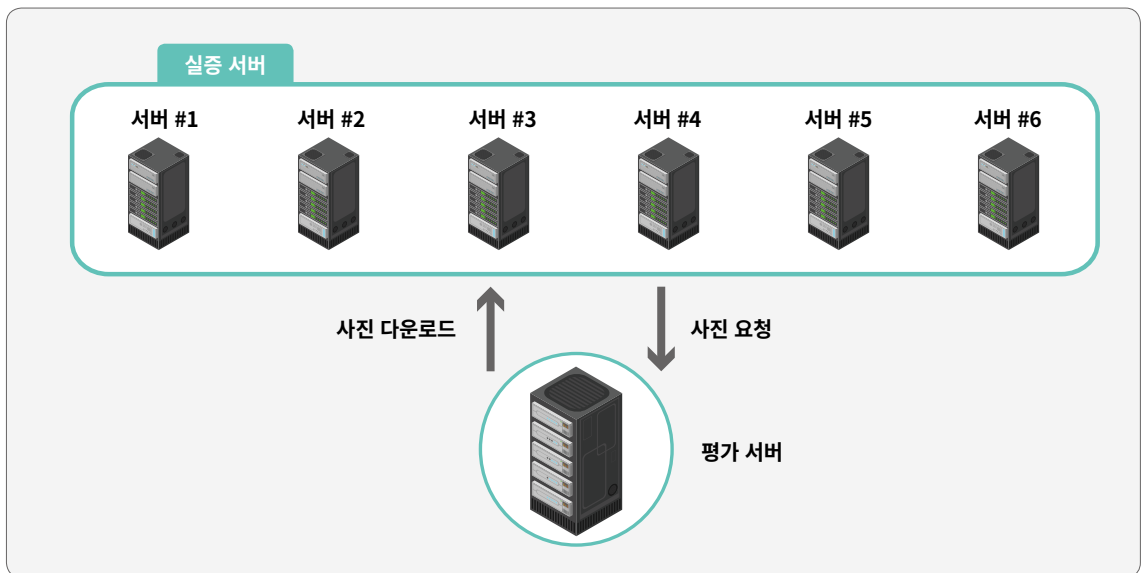
[그림 2] 식별추적 시스템 안면인식 시나리오



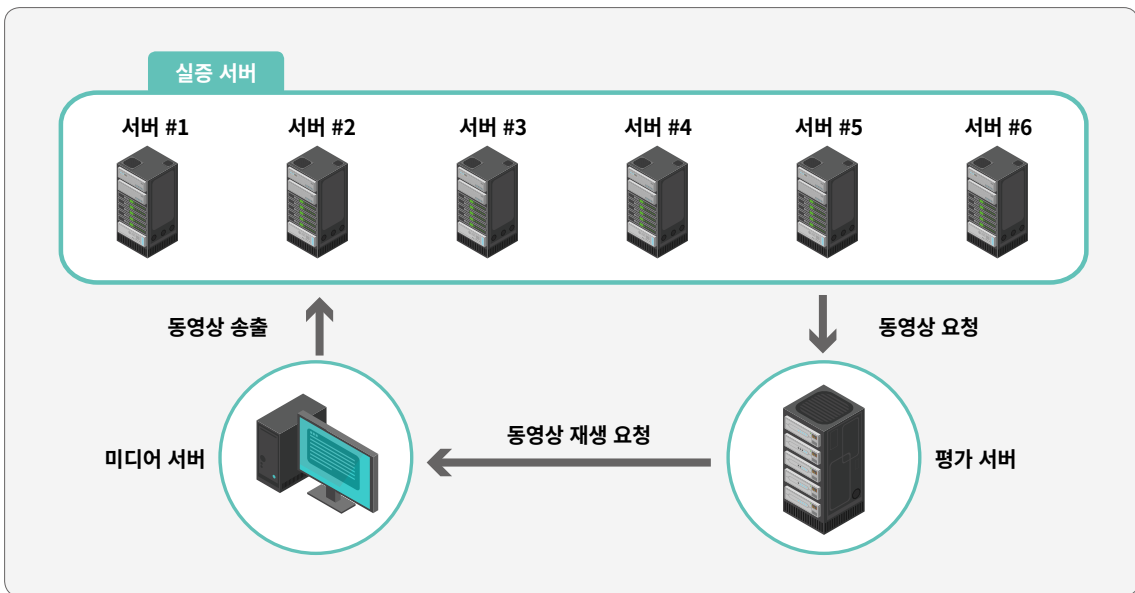
[그림 3] 식별추적시스템 이상행동패턴 시나리오



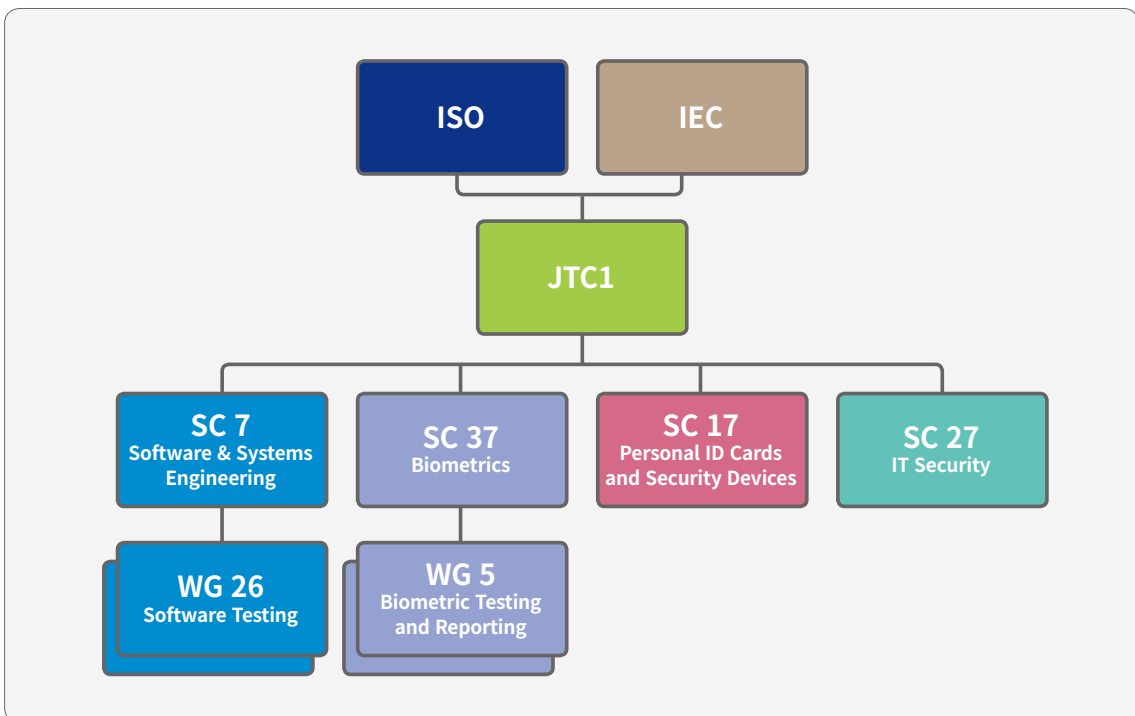
[그림 4] 사진 검증



[그림 5] 동영상 검증



[그림 6] 생체인식 테스트 관련 전문 위원회



색으로 표시된 SC7 산하 WG26은 소프트웨어 테스트를 담당하고 있다. SC37은 생체인식을 담당하고 있는 전문 위원회이며 WG5는 생체인식 테스트를 맡고 있다. 생체인식 시스템과 관련된 보안 테스트이나 생체인식 정보 저장 매체에 관한 표준을 담당하는 다른 2개의 전문 위원회도 그림에 표시돼 있다.

2002년에 출범한 SC37은 생체인식 표준을 현재까지 131개 배포했으며 지금도 29개의 새로운 표준을 개발하고 있다. SC37에는 29개국의 표준 기구가 참여하고 있으며 우리나라의 KATS(국가기술표준원)도 참여하고 있다. 참관 자격으로 참여하는 국가 기구도 19개가 있다.

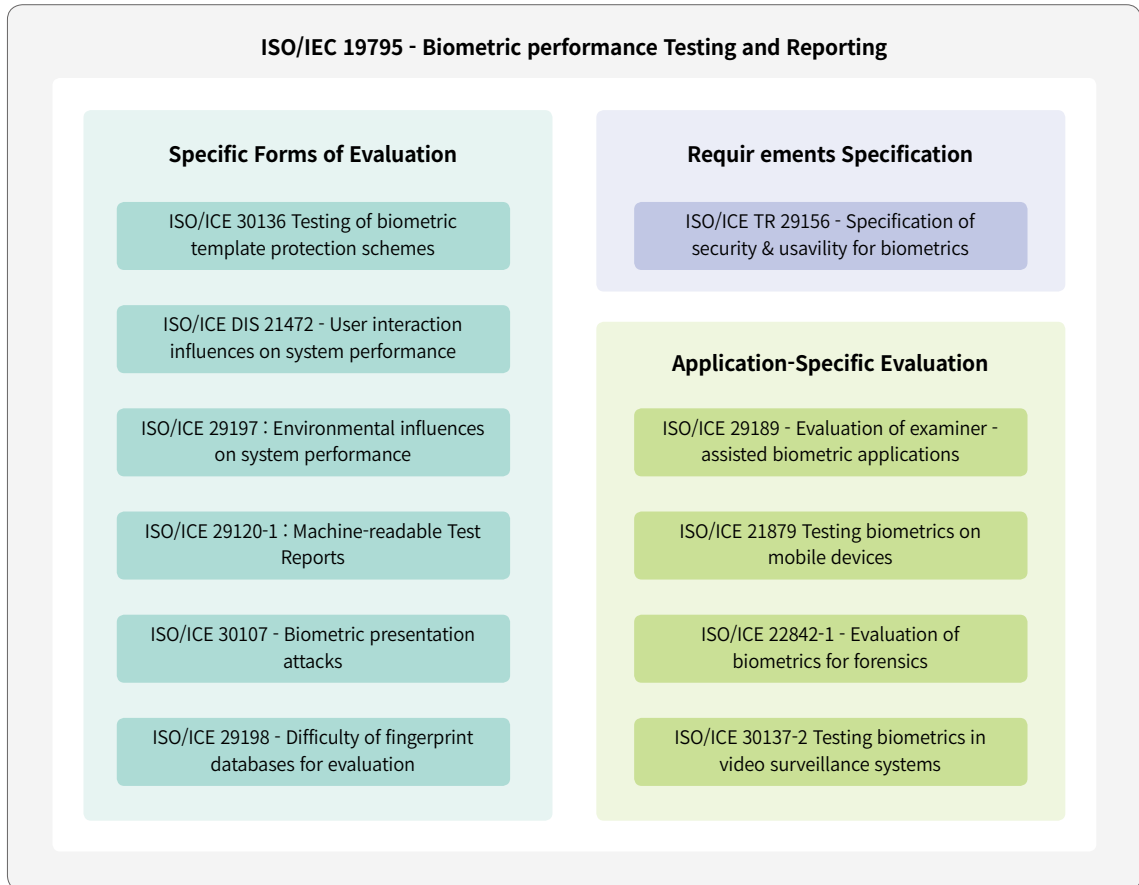
생체인식 테스트 표준 분류

생체인식 테스트 관련 표준은 여러 가지 방법으로 분류할 수 있다. 한 가지 방법은 [그림 7]과 같이 4가지 범주로 나눠보는 것이다.

좌측 녹색으로 특정 유형의 평가에 사용할 수 있는 표준이 표시돼 있다. 예를 들어, 시스템 성능을 측정할 때 고려해야 할 환경 요소나 사용자 상호작용에 관한 표준, 또 생체인식 성능을 평가하는 데 사용하는 지문 데이터베이스 난이도 측정에 관한 표준이 여기에 해당한다.

우측 상단의 ISO/IEC TR(Technical Report, 기술보

[그림 7] 생체인식 테스트 표준 분류



고서) 29156은 생체인식 요구사항을 정의하는 데 활용할 수 있는 표준이다. 우측 하단의 노란색은 법의학, 모바일 등 특정 분야에서 생체인식 테스트를 수행하는 데 활용할 수 있는 표준을 표시하고 있다.

마지막으로, 전체를 감싸는 ISO/IEC 19795 표준 시리즈가 표시돼 있다. ISO/IEC 19795는 다른 생체인식 테스트 및 보고 표준의 근간이 되는 프레임워크를 제공한다.

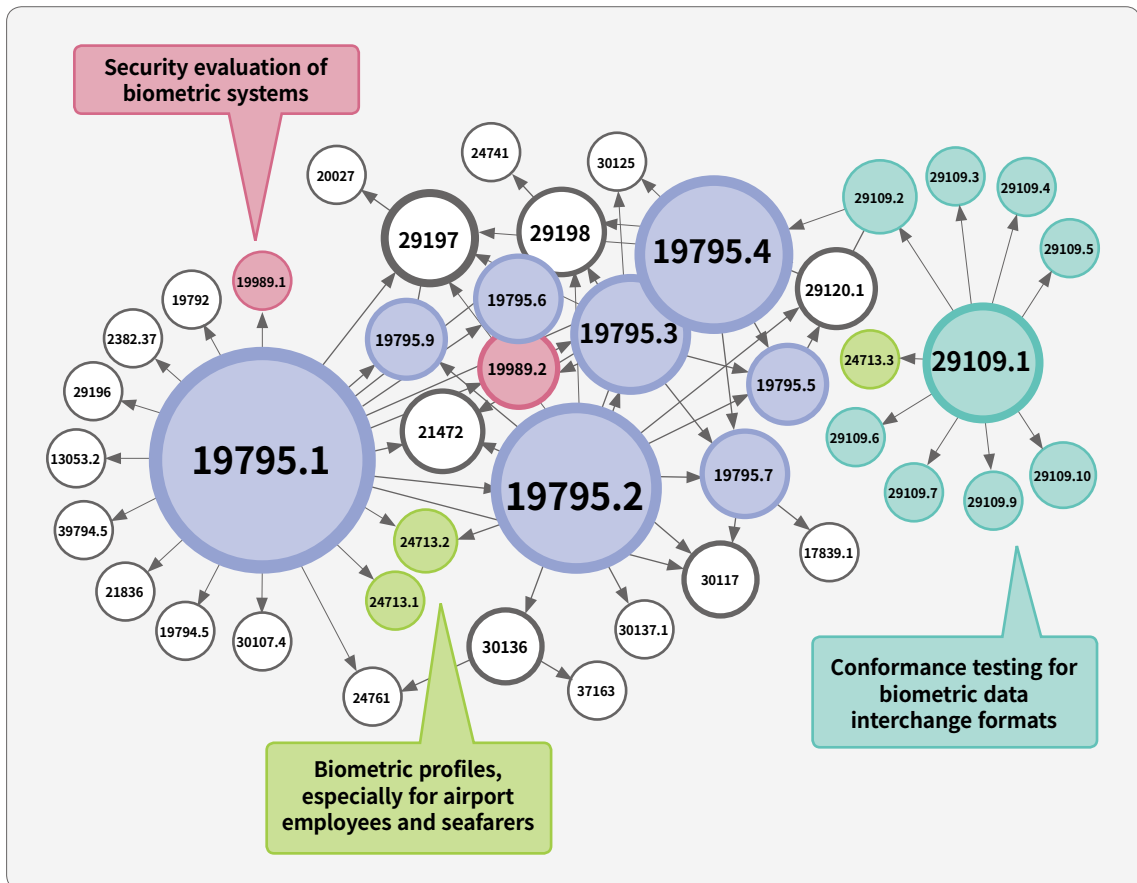
ISO/IEC 19795 제1부는 생체인식 성능 테스트와 보고를 위한 원리와 프레임워크(framework)를 제공한다. 2006년에 처음 발표됐으며, 올해 새로운 버전이 배포될 예정이다. 이 표준은 생체인식 시스템 테스트에 관한 전

반적인 지침을 제공하고 있다. 이런 지침은 지문, 목소리, 얼굴 등 사용하는 생체정보와 관계없이 모두 적용할 수 있다. 이 표준의 주요 내용 중에는 생체인식 시스템의 검증에 적용할 수 있는 핵심 성능 평가지표 정의가 있다.

생체인식 표준 간 연관성

[그림 8]은 생체인식 테스트와 관련된 주요 표준을 원으로 표현하고 있다. 원을 서로 연결하는 화살표는 하나의 표준에서 다른 표준을 참조하는 것을 나타내고 있으며 원

[그림 8] 생체인식 표준 간 연관성



의 크기는 해당 표준이 얼마나 자주 참조되는지 보여준다.

앞에서 언급한 ISO/IEC 19795 제1부가 눈에 띄게 크다는 것을 볼 수 있으며, 이 표준이 22개의 다른 표준에서 참조되고 있다는 것을 알 수 있다. 하나의 표준은 평균적으로 2개의 다른 표준에서 참조한다.

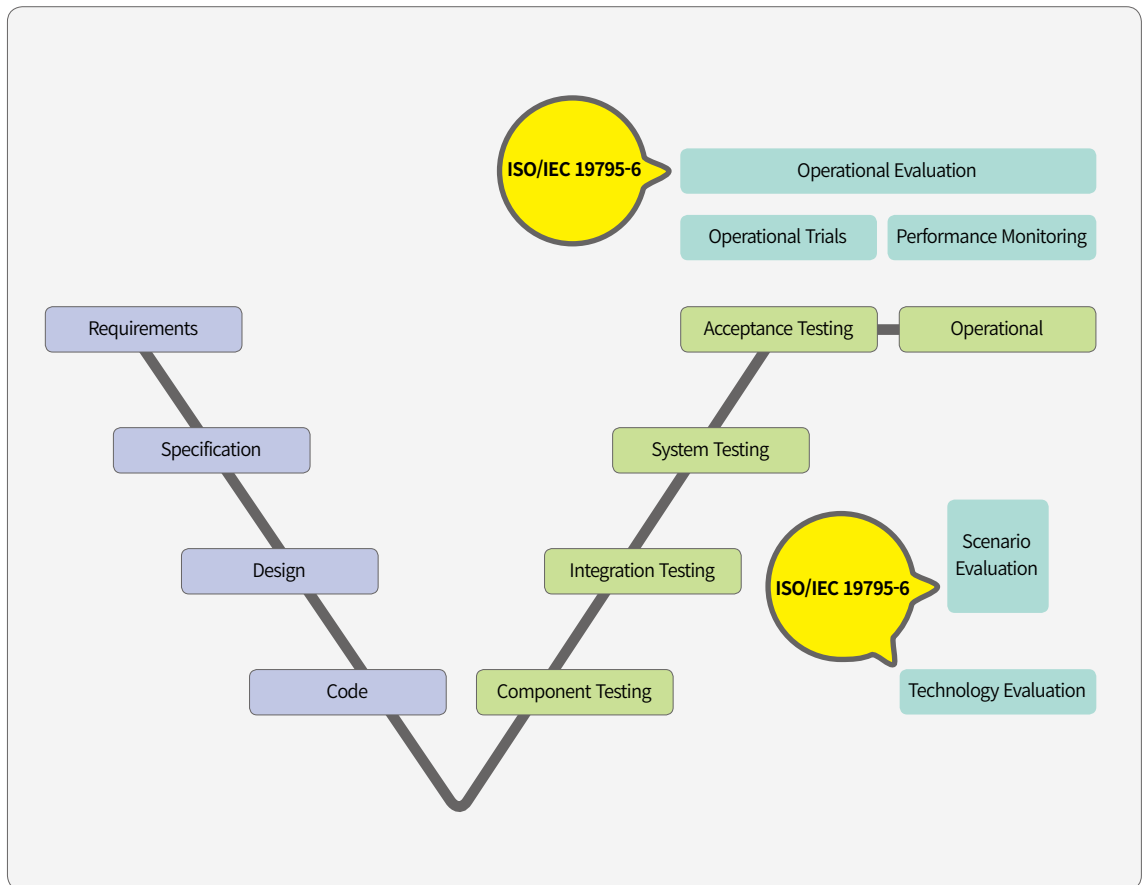
같은 시리즈의 표준은 같은 색으로 표시돼 있다. 19795 프레임워크(architecture) 표준은 노란색으로 표시돼 있고, 준수 테스트(conformance testing)와 관련된 표준은 보라색으로, 보안 평가 표준은 황색으로, 생체 인식 프로파일 관련 표준은 파란색으로 표시돼 있다. 특정 시리즈에 포함되지 않은 표준은 하얀색으로 표시돼

있다. 여기서 확연하게 중심이 되는 19795 시리즈 표준의 중요성을 확인할 수 있다.

생체인식 테스트 3레벨

생체인식 시스템 테스트에서 사용하는 3단계 평가를 살펴보고 앞서 언급한 표준이 어떻게 활용되고 있는지 살펴볼 수 있다. [그림 9]의 좌측은 각 소프트웨어 개발 단계가 있고 우측은 연관된 테스트 단계를 표시한 일반적인 V 모델을 보여주고 있다. 생체인식 시스템 테스트

[그림 9] V 모델과 생체인식 시스템 테스트 레벨 매핑



은 그것만의 독특한 3레벨이 있다.

가장 아래에는 기술 평가(Technology Evaluation)가 있다. 컴포넌트 테스트로 생각할 수 있으며 기술 평가에서는 테스트 데이터베이스(database)에 저장된 생체 인식 데이터로 사용해서 이루어진다. 일반적으로 얘기하는 통합 테스트와 시스템 테스트는 시나리오 평가(Scenario Evaluation)에 해당된다. 이 단계에서는 실제 센서와 사람을 사용해서 더 사실적인 테스트가 이뤄지게 된다. 기술 및 시나리오 평가는 19795 시리즈의 제2부에서 자세히 다루고 있다. 운영 평가(Operational Evaluation)는 시범 활용과 운영되는 시스템의 성능 모니터링을 포함한다. 그러다 보니 일반적으로 실제 운영되는 시스템을 가지고 고용된 테스트 인원이 아닌 실제 사람들을 대상으로 이뤄지게 된다. 19795 제6부는 운영 평가에 관한 구체적인 지침을 제공한다.

생체인식 업계에서는 생체인식 시스템의 핵심이 되는 비교 알고리즘의 성능 신뢰성을 강조한다. 하지만 생체인식

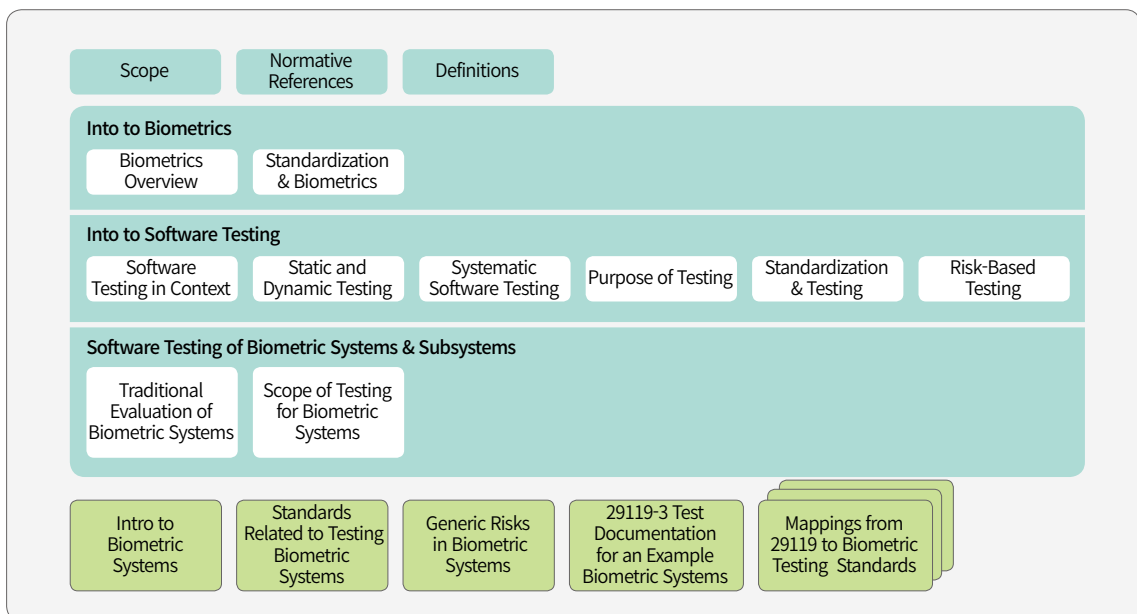
시스템에 존재하는 다른 리스크를 외면하면서까지 비교 알고리즘 테스트에 과한 노력을 들이고 있다고 볼 수도 있다.

리스크 기반 접근 방식을 통해 전반적인 생체인식 시스템에 대해 균형 잡힌 테스트가 이루어질 수 있도록 하는 것이 좋을 것으로 보인다. 비교 알고리즘을 테스트하면서는 테스트 결과가 그 알고리즘이 운영 시스템에 통합됐을 때도 성능을 유지할지 고민해야 한다.

ISO/IEC 29119-13

기존 생체인식 테스트 표준은 인식 성능 측정이라는 면을 중심으로 이야기하고 있다. 하지만 사용성, 신뢰성, 반응 시간을 포함한 여러 가지 비기능적 품질특성을 고려해야 할 필요가 있다. 마지막으로 SC37의 생체인식 테스트 표준은 리스크 중심의 진행을 명시하고 있지는 않다. ISO 29119 소프트웨어 테스트 표준 시리즈의 핵심

[그림 10] ISO/IEC 29119-13 TR



이 되는 리스크 기반 테스트 접근 방식을 생체인식 시스템에 적용하는 것은 매우 유용할 것으로 보인다.

이런 이유로 생체인식 시스템의 테스트에 ISO 29119 소프트웨어 테스트 표준의 사용을 다루는 ISO 기술 보고서가 개발되고 있다. 이 기술보고서는 우선 생체인식 분야를 처음 접하는 소프트웨어 테스터를 위해 생체인식 분야에 관한 소개를 담고 있으며, 또 SC37 생체인식 테스트 표준에서 얘기하는 기능 성능 측정지표를 넘어서 생체인식 시스템을 테스트하고자 생체인식 분야 전문가를 위해서 소프트웨어 테스트에 관한 소개를 제공한다.

생체인식 시스템 테스트와 관련된 모든 SC37, SC27, SC17 표준을 소개하고 있으며, 가장 중요한 SC37 테스트 표준과 ISO 29119 시리즈 표준 간의 매핑도 제공한다.

생체인식 시스템을 테스트할 때 ISO 29119 표준을 준수하고자 한다면, 즉 리스크 기반 테스트 접근 방식을 활용하기 위해 테스트 전략을 수립 과정에서 활용할 수 있는 리스크 및 그것에 대한 완화 방법을 체크리스트로 제공하고 있다. 마지막으로 실제 공항에서의 생체인식 시스템을 테스트하면서 개발한 테스트 문서도 예시로 제공하고 있다.

안면인식을 포함한 생체인식 시장은 빠르게 발전하고 있다. 공공분야 외에 민간에서도 도입이 늘어나고 있으며 머신러닝의 발전과 함께 활용 사례가 점점 늘어날 것으로 예상된다. 그러나 아직은 이런 시스템의 성능을 충분히 검증할 수 있는 기준은 명확하게 확립되어 있지 않다. 지금까지 존재하는 평가 방법은 대부분 인식/매칭 성능 자체의 측정에 머무르고 있다.

앞으로 이런 생체인식 시스템을 사용자가 믿고 사용할 수 있기 위해서는 생체인식 엔진의 성능을 다양한 측면에서 검증할 방법뿐만이 아닌 그런 컴포넌트가 포함된 전체 시스템을 테스트하는 표준적인 방법이 필요할 것으로 보인다.

참고 자료

AI식별추적시스템 구축 사업 의의와 성과, 강승준, 정보통신 산업진흥원.

인공지능 식별추적시스템 성능 검증 및 실증랩 구축 운영 지원 사업을 통한 평가 모델 고도화, 손영수, 어니컴 & STA테스팅컨설팅

AI기반의 안면인식 검증 및 생체인식 테스트 표준 동향 세미나, Stuart Reid, 어니컴 & STA테스팅컨설팅