

딥러닝 기반 고성능 얼굴인식 기술 동향

Research Trends for Deep Learning-Based
High-Performance Face Recognition Technology

김형일 (H.I. Kim, hikim@etri.re.kr)

시각지능연구그룹 연구원

문진영 (J.Y. Moon, jymoon@etri.re.kr)

시각지능연구그룹 책임연구원

박종열 (J.Y. Park, jongyoul@etri.re.kr)

시각지능연구그룹 책임연구원/그룹장

As face recognition (FR) has been well studied over the past decades, FR technology has been applied to many real-world applications such as surveillance and biometric systems. However, in the real-world scenarios, FR performances have been known to be significantly degraded owing to variations in face images, such as the pose, illumination, and low-resolution. Recently, visual intelligence technology has been rapidly growing owing to advances in deep learning, which has also improved the FR performance. Furthermore, the FR performance based on deep learning has been reported to surpass the performance level of human perception. In this article, we discuss deep-learning based high-performance FR technologies in terms of representative deep-learning based FR architectures and recent FR algorithms robust to face image variations (i.e., pose-robust FR, illumination-robust FR, and video FR). In addition, we investigate big face image datasets widely adopted for performance evaluations of the most recent deep-learning based FR algorithms.

* DOI: 10.22648/ETRI.2018.J.330405

* 본 연구는 과학기술정보통신부의 재원으로 정보통신기술진흥센터의 지원을 받아 수행한 연구임[No.B0101-15-0266, 실시간 대규모 영상 데이터 이해 예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발].

2018
Electronics and
Telecommunications
Trends

Deep Neural Network &
Object Recognition 특집

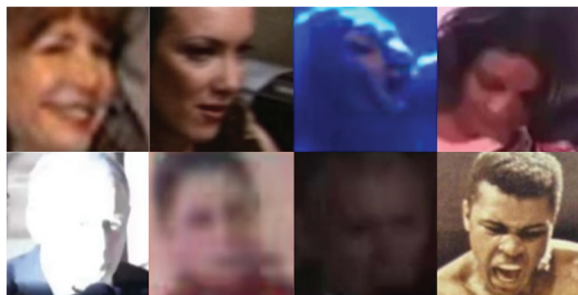
- I. 서론
- II. 얼굴인식 기술 개괄
- III. 딥러닝 기반 얼굴인식
기술 동향
- IV. 얼굴인식 기술 검증에
위한 데이터셋
- V. 결론

I. 서론

얼굴인식 기술은 얼굴을 포함하는 입력 정지영상 또는 비디오에 대해 얼굴 영역의 자동적인 검출 및 분석을 통해 해당 얼굴이 어떤 인물인지 판별해 내는 기술로 패턴인식 및 컴퓨터 비전 분야에서 오랫동안 연구되어 온 분야이다. 많은 연구결과로부터 최근에는 감시시스템, 출입국 관리, 생체인식 등과 같은 실제 환경에 적용되고 있다. 이러한 얼굴인식 기술은 다른 물체인식 기술 대비 상대적으로 성숙된 기술로 높은 성능을 보인다고 알려져 있으나, 실제 환경에서 취득되는 얼굴 영상은 (그림 1)과 같이 포즈 및 조명 변화, 원거리 촬영과 사람의 움직임에 의한 저 해상도/블러(blur) 문제 등으로 인해 여전히 해결해야 할 문제가 많은 것으로 알려져 있다.

최근 딥러닝(Deep Learning) 기술의 발달에 힘입어 영상인식 기술이 매우 빠른 속도로 발전하고 있다. 딥러닝 기술은 매우 많은 수의 층(Layer)으로 구성된 깊은 신경망 구조를 대용량의 데이터를 이용하여 학습시키는 기술로써, 비선형의 계층적 특징 학습 능력은 사람의 인지 메커니즘과 유사하다고 알려져 있다. 이러한 딥러닝 기술이 얼굴인식에 접목됨에 따라 다양한 데이터 환경에서 고성능의 얼굴인식이 가능하게 되고, 사람의 인지 수준을 능가하는 연구 사례 또한 나타나고 있다.

본고에서는 고성능의 얼굴인식을 위한 대표적인 딥러닝 기반 얼굴인식 기술의 연구 사례에 대해 살펴보고, 딥러닝 기반 얼굴인식 기술을 입증하기 위한 도전적인



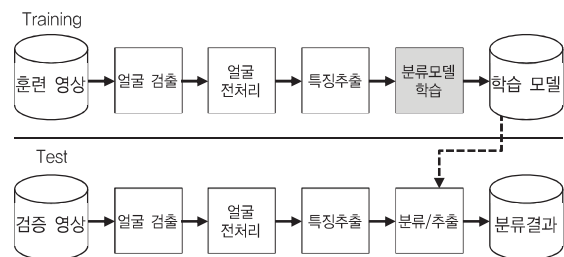
(그림 1) 실제 환경에서 취득되는 얼굴 영상 예제

데이터셋(Dataset), 그리고 최근 연구들의 얼굴인식 성능에 대해 살펴보도록 한다.

본고의 II장에서는 얼굴인식 기술에 대한 개괄적인 설명을 하고, III장에서는 딥러닝 기반 고성능 얼굴인식 기술에 대해 살펴본다. IV장에서는 딥러닝 기반 얼굴인식 구조의 학습 및 검증을 위한 데이터셋과 최근 얼굴인식 성능에 대해 살펴보고, V장에서는 결론을 맺는다.

II. 얼굴인식 기술 개괄

얼굴인식 기술은 (그림 2)와 같이 기본적으로 입력 정지영상 또는 비디오 내에 존재하는 얼굴 영역을 검출하고 얼굴영역에서 추출된 랜드마크(Landmarks) 정보를 이용한 얼굴정렬, 광도보정, 그리고 영상 정규화 같은 전처리를 수행한다. 다음으로 전처리된 얼굴 영상으로부터 특징(Feature)을 추출하고, 이 특징을 이용하여 훈련 단계에서는 분류모델을 학습시키고, 검증 단계에서는 이 모델을 이용하여 어떤 사람인지 판별하게 된다. 얼굴인식 기술은 응용에 따라 얼굴검증(Face Verification) 그리고 얼굴식별(Face Identification) 기술로 구분된다. 얼굴검증은 입력으로 들어오는 두 개의 얼굴 영상이 동일 인물인지 여부를 판단하는 1:1 검증 문제이며, 얼굴식별은 입력으로 들어오는 하나의 얼굴 영상이 사전에 등록된 N 명의 인물 중 어떤 인물에 해당하는지를 판단하는 1: N 의 문제로 볼 수 있다. 보통 얼굴식별 기술은 학습된 모델을 이용하여 검증의 대상이 되는 프로브(Probe) 영상과 사전에 등록된 N 명에 대한 갤러리



(그림 2) 얼굴인식 파이프라인(Pipeline)

(Gallery) 영상과의 비교를 통해 수행한다. 기존 전통적인 얼굴인식 기술에서는 주로 얼굴 영상에 대해 분별력 있는 특징을 추출하기 위한 기술(예: Local Binary Patterns, Gabor 특징과 같은 Handcrafted 특징)과 추출된 특징에 대해 어떤 인물인지 판별하기 위한 분류모델이 사용되었다. 이러한 기존의 컴퓨터 비전 및 패턴인식 기술을 이용한 방법들은 실제 환경에서 얼굴 영상의 다양한 변화가 발생했을 때 학습에 사용된 데이터와 입력된 테스트 영상 사이의 불일치로 인해 성능이 매우 저하되는 것으로 알려져 있다. 대조적으로, 딥러닝 기반 얼굴인식 기술에서는 특징추출과 분류모델 학습을 다양한 환경에서 구축한 대용량의 얼굴 데이터를 이용하여 End-to-End로 학습함에 따라 매우 고차원적인 특징을 스스로 학습하게 된다. 이에 따라 와일드 환경에서의 얼굴인식 성능은 매우 높아지고 있고, 사람의 인지 성능을 뛰어넘는 연구 사례도 나타나고 있다. 다음 절에서는 와일드 환경에서의 딥러닝 기반 얼굴인식 기술의 연구 사례에 대해 상세히 살펴본다.

III. 딥러닝 기반 얼굴인식 기술 동향

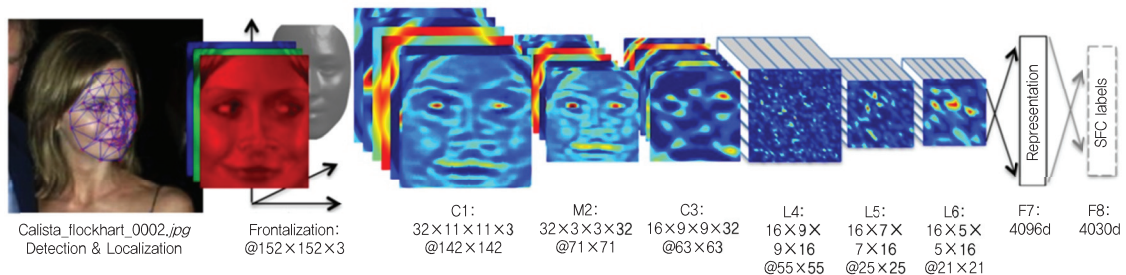
본 장에서는 실제 환경에서 취득한 다양한 변화를 갖는 얼굴 영상을 위한 고성능의 딥러닝 기반 얼굴인식 기술 사례들을 살펴보고, 분별력 있는 특징 학습을 위한 손실함수(Loss Function) 기반 거리 척도 학습(Metric Learning) 기술을 살펴본다. 다음으로, 포즈변

화와 조명변화 문제를 다루기 위한 딥러닝 기반 얼굴인식 사례, 그리고 다중 프레임으로 구성된 비디오를 위한 딥러닝 기반 얼굴인식 사례에 대해 살펴보도록 한다.

1. 대표적인 딥러닝 기반 얼굴인식 구조

가. DeepFace

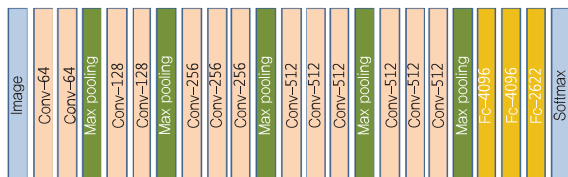
딥러닝 기술이 얼굴인식에 처음으로 접목된 연구는 2014년 CVPR에서 발표된 Facebook의 DeepFace[1] 연구이다[(그림 3) 참조]. DeepFace에서는 사전에 학습된 3D 얼굴 기하 모델을 이용하여 랜드마크 추출 후에 어파인(affine) 변환에 의해 얼굴 정렬을 수행한 후 9개의 층으로 구성된 컨볼루션(Convolution) 신경망을 Facebook이 내부적으로 수집한 대용량의 데이터를 이용해 학습하였다. 이때, 얼굴 영역 내의 지역적인 특징을 효과적으로 추출하기 위하여 국소 연결(Locally Connected) 컨볼루션 층을 사용한 것이 특징이다. DeepFace 구조는 LFW 데이터셋에 대해 97.35% 정확도를 달성하면서 기존 Hand-crafted 특징 기반 방법과 비교했을 때 약 27%의 인식률을 향상시키면서 딥러닝 기반 방법이 매우 성공적임을 보여주는 사례가 되었다. 하지만 당시에는 Facebook이 보유하고 있는 정도의 대용량 데이터를 확보하기 어려웠기 때문에 이러한 딥러닝 구조의 학습에 제한이 있었다. 또한, 국소 연결 컨볼루션 층에 의해 1억2천만 개의 매우 많은 수의 신경망 파라미터(Parameter) 학습이 필요하게 되어 이러한 신경망 구조를 학습하기 어려운 것이 단점이다.



(그림 3) Deepface 딥 네트워크 구조[1]

나. VGGFace

DeepFace 이후로 등장한 구조는 옥스포드 대학의 Visual Geometry Group(VGG)에서 제안한 VGG-Face[2](또는 DeepFR) 딥 네트워크 구조이다. VGG-Face에서는 인터넷 검색을 통해 직접 만든 대용량의 얼굴인식을 위한 데이터셋인 VGG 얼굴 데이터셋을 공개하고, 이 데이터를 이용하여 (그림 4)와 같이 15개의 컨볼루션 층으로 구성된 딥 네트워크 구조를 학습시켰다. VGG에서는 VGGFace 학습 모델을 제공할 뿐만 아니라, ImageNet 영상인식 챌린지(Challenge)에서의 VGG 구조와 마찬가지로 상대적으로 간단한 3×3 컨볼루션 필터를 이용하여 학습시킴으로써 VGGFace는 LFW 데이터셋에 대해 DeepFace보다 약 1% 정도 개선된 98.95% 성능을 달성하였다. 이 외에도 얼굴인식 성능 개선을 위해 DeepID[3], DeepID2[4], DeepID2+[5], 그리고 DeepID3[6]와 같은 얼굴인식을 위한 다양한 딥 네트워크 구조가 제안된 바 있다.

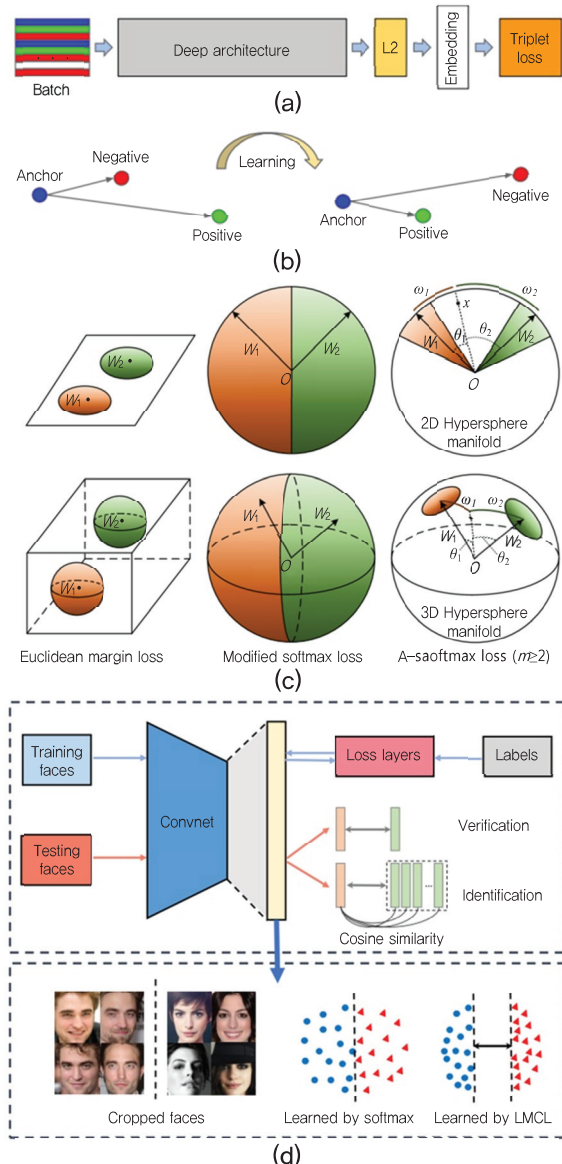


(그림 4) VGGFace 딥 네트워크 구조

2. Distance Metric Learning

앞에서 언급한 다양한 얼굴인식 딥 네트워크 구조에 대한 연구 사례와 더불어서 손실함수(Loss Function)의 재정의에 통한 분별력 있는 특징을 학습하기 위한 거리 척도 학습(Distance Metric Learning)에 대한 연구가 수행되었다. 대표적으로, 구글에서 발표한 FaceNet[7]에서는 (그림 5a)-(b)와 같이 동일한 인물에 대해 추출된 특징들 사이의 유클리디언(Euclidean) 거리가 다른 인물들로부터 추출된 특징들 사이의 유클리디언 거리보다 작다는 Triplet Loss를 정의하여 딥 네트워크를 학

습시킨 사례가 있다. 즉, (그림 5b)와 같이 학습을 통해 anchor 특징과 동일인에 대한 Positive 특징 사이의 거리가 negative 특징과 비교했을 때 가까워지는 것이다. 이때, 기본 딥 네트워크 구조는 22개 층으로 구성된 Inception 네트워크를 사용하였고, 내부적으로 수집된 500M 장의 얼굴 영상 데이터셋으로 학습하였다. 학습



(그림 5) (a) FaceNet 딥 네트워크 구조[7], (b) FaceNet에 의한 학습 결과 예제[7], (c) A-Softmax 손실 함수 시각화[8], (d) CosFace 기술 개요[9]

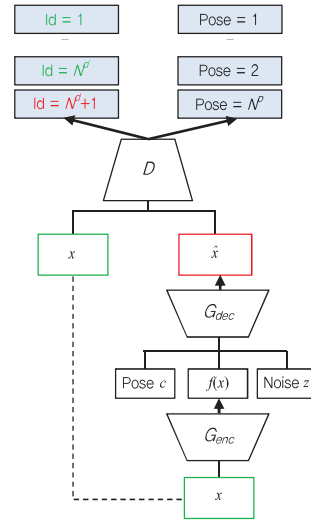
을 통해 얼굴 영상 입력에 대해 높은 분별력을 갖는 128 바이트(bytes)의 특징 embedding을 수행하였고, LFW 데이터셋에 대해 99.6%의 정확도를 보였다.

FaceNet 연구의 성공에 따라 [7]에서는 hypersphere 공간에서의 고차원의 영상 데이터 맵핑(mapping)을 위한 angular softmax(A-Softmax, (그림 5c)) 손실함수를 정의하여 학습을 수행한 SphereFace[8] 연구가 수행되었다.

이러한 손실함수를 재정의하는 것의 주요 목적은 동일 인물로부터 추출된 특징의 분산은 작게 하고, 다른 인물로부터 추출된 특징의 분산은 크게 하는 것이다. 그 밖에 기존 방법들의 성능을 개선하기 위한 Large Margin Cosine(LMCL) 손실함수를 정의한 CosFace (그림5d)[9], Ring Loss[10], 그리고 ArcFace[11] 등의 연구가 있다.

3. 포즈변화에 강인한 얼굴인식 기술

와일드 얼굴인식에서 포즈변화 문제는 얼굴인식 성능을 저하시키는 주요 원인 중 하나로 알려져 있다. 보통 얼굴인식 시스템의 갤러리 영상은 정면의 얼굴 영상으로 등록되는데 반해 프로브 영상은 다양한 포즈를 갖는 영상들이 입력으로 들어오게 되어 갤러리와 프로브 사이의 불일치 문제로 성능이 매우 저하된다. 이러한 포즈변화 문제를 다루기 위해 포즈변화에 강인한 특징 학습 기법 및 정면 얼굴 복원 연구가 수행되고 있다. 특히, 최근에는 Generative Adversarial Network(GAN) 기술의 발달에 따라 다양한 영상 생성 및 인식 문제에 활용되고 있다. [12]에서는 포즈변화에 강인한 얼굴인식을 위한 GAN 모델인 Disentangled Representation Learning-Generative Adversarial Network(DR-GAN)를 제안하였다. DR-GAN에서는 (그림 6)과 같이 인코더(Encoder)-디코더(Decoder) 기반 생성기와 컨볼루션 신경망 기반 분류기로 구성된다. 입력 영상(x)에 대해 생성



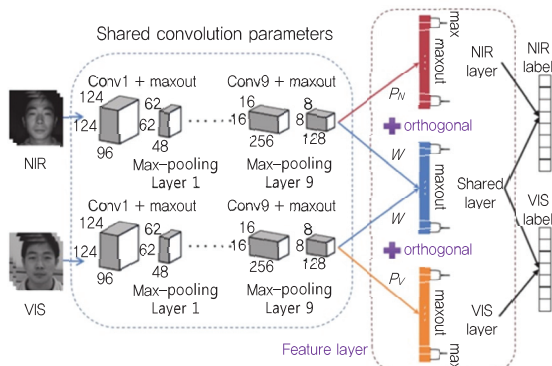
(그림 6) DR-GAN 딥 네트워크 구조[12]

기의 인코더(G_{enc})를 통해 특징($f(x)$)을 추출하게 되고, 추출된 특징, 원하는 포즈(c), 그리고 잡음(z)을 이용하여 디코더(G_{dec})의 출력으로 특정 포즈의 영상($\hat{x}$)을 생성하게 된다. 분류기(D)에서는 생성된 특정 포즈를 갖는 영상에 대해 실제 포즈 라벨(Label)과 인물정보를 구분하도록 학습을 수행하게 된다. 즉, 특정 포즈를 갖는 입력 영상이 들어왔을 때 원하는 포즈를 갖는 얼굴 영상을 생성하여 포즈변화에 강인한 인식을 수행하는 것이다. 이때, 다양한 포즈변화가 존재하는 MultiPIE 데이터셋에 대해 89.2% 성능을 달성하였다.

4. 이종의 얼굴인식 기술

다음으로, 조명변화에 강인한 딥러닝 기반 얼굴인식 기술 사례에 대해 살펴보도록 한다. 와일드 환경에서의 얼굴을 포함하는 영상은 촬영 시간과 설치 위치에 따라 변화하는 조명에 의해 매우 다른 영상을 얻는다. 이러한 문제를 다루기 위해 조명(Illumination)과 반사(Reflectance) 모델링을 통해 조명 정보를 추정하는 연구 사례가 있고, 조명에 의해 열화된 영상의 정규화/전처리 또는 조명변화에 강인한 특징 추출 기법과 관련된 연구 사례가 있다. 또한 근적외선(NIR: Near-Infrared) 영상이

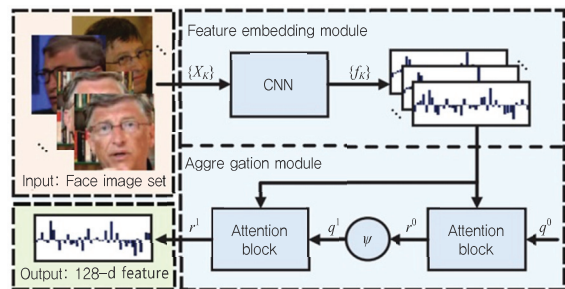
조명변화에 민감하게 반응하지 않는다는 점에 착안하여 근적외선 영상과 가시광(VIS: visible) 영상의 매칭(Matching)을 통한 얼굴인식 기법도 주목 받고 있다. 이러한 문제를 이종의(Heterogeneous) 얼굴인식 기술로 부르며, 핵심 기술은 서로 다른 영상들 사이의 Modality 차이를 줄이고 인물의 ID 정보만 포함하는 공통된 특징을 학습하도록 하는 것이다. 2017년 AAAI에서 발표된 [13]에서는 (그림 7)과 같은 이종의 얼굴인식을 위한 딥 네트워크를 제안하고 있다. 먼저 Low-Level 층에서는 대용량의 근적외선 얼굴 영상을 확보하는 것이 어렵기 때문에 얼굴의 특징을 효과적으로 학습하기 위한 대용량의 가시광 영상으로 학습된 딥 네트워크를 특징 추출에 활용하게 된다. 다음으로, High-Level 층에서는 각 Modality와 관련된 특징과 인물에 관련된 특징 학습을 위해 Maxout 기반 딥 네트워크를 구성하였고, NIR 층 그리고 VIS 층에서 얻어지는 특징과 shared 층에서 얻어진 특징 사이의 Orthogonality Constraint를 부과함으로써 근적외선 및 가시광 영상의 공통된 인물의 특징을 추출하는 방법을 제안하였다. 본 연구를 통해 CASIA 2.0 데이터셋을 이용하여 95.82%의 정확도를 달성하였다. 이는 간단한 Handcrafted 특징 중 하나인 주성분 분석 알고리즘과 비교했을 때 약 70%의 성능 향상, 그리고 일반 가시광 영상으로 학습된 VGG 특징 추출기와 비교했을 때 약 30%의 성능 향상을 보였다.



(그림 7) NIR-VIS 이종의 얼굴인식 프레임워크에서 shared 특징 추출을 위한 딥 네트워크 구조[9]

5. 비디오 얼굴인식

실세계에서 얼굴인식을 위한 입력은 얼굴의 시간 변화에 따른 다중 프레임의 비디오 형태로 들어오게 된다. 비디오 내에는 다중의 얼굴 영상 취득이 가능하고, 정지 영상에서 놓칠 수 있는 인물 정보가 포함되어 있기 때문에 다중 정보를 효과적으로 다루기 위한 비디오 기반 얼굴인식 연구 역시 활발히 수행되고 있다. 고전적인 비디오 얼굴인식(또는 영상 셋(Set) 매칭) 기술은 여러 개의 프레임 정보로부터 단일의 특징 또는 대표 영상을 생성하여 얼굴인식을 수행하거나, 다중의 얼굴영상 또는 특징에 대한 집합(Set, 예: convex hull, affine hull)을 생성하여 집합 대 집합(Set-to-Set) 매칭을 수행하게 된다. 2017년 CVPR에서 발표된 [14]는 비디오 얼굴인식을 위해 여러 개의 프레임 정보로부터 단일 특징을 추출



(a)



(b)

(그림 8) (a) Neural aggregation network(NAN) 딥 네트워크 구조[14] (b) 비디오 내에 존재하는 얼굴 영상과 NAN attention 모듈에 의해 얻게 되는 가중치[14]

하는 딥 네트워크인 Neural Aggregation Network (NAN)을 제안하였다. 제안된 NAN 딥 네트워크는 (그림 8a)와 같이 크게 두 모듈로 구성된다. 첫 번째 모듈은 특징 임베딩(Feature Embedding) 모듈로써 각 프레임에 대해 공통된 딥 네트워크 모델을 이용하여 특징을 추출한 후 다중의 특징 벡터를 구성하는 모듈이고, 두 번째 모듈은 통합(Aggregation) 모듈로써 이 특징들을 두 개의 주의(Attention) 블록을 이용하여 하나의 융합된 특징 벡터를 생성하게 된다. 이때, 통합 모듈의 주의 블록은 각각의 영상 프레임으로부터 획득된 특징 벡터에 대해 딥 네트워크에 의한 적응적인 가중치를 이용하여 다중의 특징 벡터를 융합하여 하나의 특징 벡터를 생성한다. 이때, 상대적으로 좋은 품질을 갖는 얼굴 영상은 높은 가중치를 갖게 되고, 다양한 변화에 의해 열화된 얼굴 영상에 대해서는 낮은 가중치가 할당되는 것을 실험적으로 보였다. NAN을 통해 YouTube Face 데이터셋 기준 95.72%의 성능을 달성하였다.

IV 얼굴인식 기술 검증을 위한 데이터셋

1. 대용량의 공개 데이터셋

딥러닝 기반 얼굴인식 구조 학습 및 성능을 검증하기 위해 <표 1>과 같이 다양한 대용량의 데이터셋이 활용되고 있다. 이 중 CelebFaces, DeepFace(페이스북), 그리고 FaceNet(구글)은 대용량 데이터셋으로 딥 네트워크 학습에 적합하지만, 비공개 데이터셋으로 일반적으로 활용이 제한된다. 반면, VGGFace 데이터셋의 경우에는 공개 데이터셋으로 현재 두 가지 버전(VGGFace 및 VGGFace2)이 딥 네트워크 학습을 위해 활용 가능하다고 알려져 있다. 대용량의 학습 데이터와 더불어 최근 얼굴인식 기술들의 성능 검증을 위해 다양한 챌린지가 개최되며 이에 따라 다양한 변화를 포함하는 와일드 검증 데이터셋이 공개되고 있다. 그 중 가장 널리 사용되는 데이터셋은 LFW[15], YouTube Face(YTF)[16],

<표 1> 얼굴인식을 위한 대용량의 데이터셋 통계

데이터셋	영상개수	인물수	Source	공개여부
MegaFace	1,027,060	690,572	Flickr	공개
MegaFace2	4,753,320	672,057	Flickr	공개
CASIA WebFace	494,414	10,575	Celebrity Search	공개
LFW	13,233	5,749	Yahoo News	공개
FaceScrub	106,863	530	Celebrity Search	공개
YouTube Faces	3,425(V*)	1,595	YouTube	공개
CelebFaces	202,599	10,177	Celebrity Search	비공개
DeepFace	4,400,000	4,000	Internal	비공개
FaceNet	500,000,000	10M	Internal	비공개
IJB-A	5,712 2,085(V*)	500	Internal	공개
IJB-B	67,000 7,000(V*)	1,845	Internal	공개
IJB-C	138,000 11,000(V*)	3,531	Internal	공개
VGGFace	2,600,000	2,622	Google& YouTube	공개
VGGFace2	3,310,000	9,131	Google& YouTube	공개

* V: Video

IJB[17]–[19], 그리고 MegaFace[20]–[21]로 알려져 있다. 본 절에서는, LFW, YTF, IJB, 그리고 MegaFace 데이터셋에 대해 자세히 살펴보도록 한다.

가. LFW

2009년에 공개된 LFW 데이터셋은 웹에서 5,749명의 유명인에 대해 취득된 13,233장의 영상으로 구성된다. 기존에 제약된 환경에서 촬영하여 획득된 얼굴인식 데이터셋(예: FERET, MultiPIE)과 비교했을 때 상대적으로 일상생활에서 나타나는 조명, 표정, 그리고 포즈변화 등 다양한 변화가 포함되어 있기 때문에 얼굴인식 기술 성능 검증을 위해 널리 활용되어 왔다. LFW 데이터셋에는 인물 한 명당 평균적으로 2.31장의 영상으로 구성되어 갤러리 및 검증 영상 셋이 따로 존재하지 않기 때문에, 얼굴식별보다는 얼굴검증 기술의 성능 검증에 주

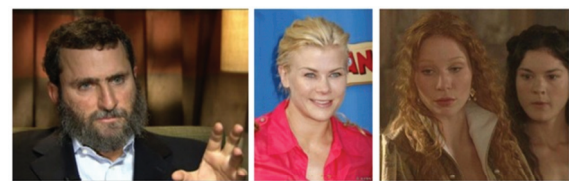
로 활용되고 있다. 최근 얼굴인식 연구에서 보여주는 LFW 데이터셋에 대한 성능은 약 99.73%[5]로 매우 높은 성능으로 포화된 것으로 알려져 있다.

나. YouTube Face-YTF

다음으로 비디오 기반 얼굴인식 또는 영상 셋 매칭에서 주로 활용되는 YTF 데이터셋은 YouTube에서 검색된 유명인의 비디오 클립으로 구성된 데이터셋으로 1,595명의 인물에 대한 3,425개의 비디오로 구성되어 있다. YTF 데이터셋은 LFW와 마찬가지로 뉴스와 같은 TV 프로그램에서 취득된 데이터셋으로 알려져 있다.

다. IJB

The IARPA Janus Benchmark(IJB) 데이터셋은 NIST(National Institute of Science and Technology)에서 얼굴인식 챌린지를 위해 공개한 데이터셋으로 정지 영상 및 비디오로 구성되어 있다. IJB 데이터셋은 A 버전(IJB-A)에서 출발하여 최근 C 버전(IJB-C)까지 공개되었다. 각 버전에 따라 정지영상, 비디오, 그리고 사람의 수가 서로 다르며, 각 버전에 따라 정의되는 챌린지의 내용이 다르게 구성된다. 또한, IJB 데이터셋은 매우 많은 포즈변화, 서로 다른 인종 및 국가에 해당하는 인물, 그리고 다양한 해상도의 얼굴을 포함하고 있다. 초기 버전인 IJB-A[17]에서는 500명의 인물에 대해 5,712장의 정지 영상과 2,085개의 비디오를 포함하고 있고, 1:1의 얼굴검증 및 1:N 얼굴 검색 태스크(task)로 구성되어 있다. IJB-B[18]는 IJB-A의 확장된 버전으로 1,345명의 새로운 인물이 추가되었다. 이에 따라, 1,845명의 인물에 대한 67,000장의 정지 영상, 7,000개의 비디오, 그리고 10,000장의 얼굴이 아닌 영상으로 구성되어 있다. 이 데이터에 대해서는 얼굴검증, 얼굴식별, 얼굴 검출, 얼굴 군집화, 그리고 crowded 영상의 처리와 관련된 8가지의 챌린지 태스크로 구성이 되어 있



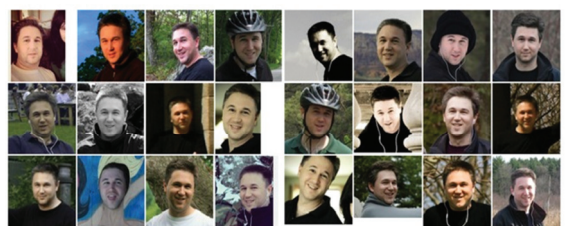
(a)



(b)



(c)



(d)

(그림 9) 각 데이터셋의 예제 얼굴 영상. (a) LFW, (b) YouTube Faces, (c) IJB, (d) MegaFace

[출처] LFW Face Database, <http://viswww.cs.umass.edu/lfw/>

[출처] YouTube Faces,

<https://www.cs.tau.ac.il/~wolf/ytfaces/>

[출처] IJB, IJB-C Dataset Request Form,

<https://www.nist.gov/itl/iad/image-group/ijb-c-dataset-request-form-0>

[출처] MegaFace, <http://megaface.cs.washington.edu>

다. 최근 공개된 IJB-C[19] 데이터셋은 IJB-B의 확장판으로 3,531명의 인물에 대해 31,334장의 정지 영상, 11,779개의 비디오, 그리고 10,000장의 얼굴이 아닌 영상으로 구성되며, IJB-B와 마찬가지로 8개의 챌린지 태스크로 구성된다.

라. MegaFace

MegaFace는 2016년 CVPR 학회에서 발표된 챌린지 및 데이터셋을 일컫는 것으로, 다양한 일반인들의 얼굴 영상이 포함되어 있으며, 백만 스케일의 얼굴인식 기술을 검증하기 위해 만들어졌다. 본 벤치마크에서는 크게 두 가지의 챌린지로 구성된다. 첫 번째 챌린지에서는 훈련용 데이터 없이 690,572명에 대한 1,027,060장의 얼굴 영상으로 갤러리 그리고 프로브 영상들로 얼굴인식 검증을 수행한다. 이때, 10명에서 1,000,000명까지의 갤러리에 등록되지 않은 인물이 나타났을 때, 얼굴인식 성능을 측정하는 것이 기존 챌린지와 차이점이다. 즉, 등록되지 않은 다양한 인물들을 비교 검색함으로써 얼굴인식 알고리즘의 강인성을 검증하기 위해 효과적인 데이터셋으로 볼 수 있다. MegaFace2[21]라고 불리는 두 번째 프로토콜에서는 672,057명의 인물에 대해 4,753,320장의 훈련 영상이 공통으로 주어진 상황에서 첫 번째 프로토콜과 마찬가지로 등록되지 않은 인물이 나타났을 때의 얼굴인식 알고리즘의 강인성을 평가한다.

2. 딥러닝 기반 얼굴인식 성능

〈표 2〉는 앞에서 언급한 네 가지의 검증 데이터셋에 대한 최근 얼굴인식 성능을 정리한 것이다. LFW 및 YTF 데이터셋에 대해서는 얼굴검증 성능이 정리되었으며, IJB-A, MegaFace, 그리고 MegaFace2 데이터셋에 대해서는 얼굴검증 및 얼굴식별 성능을 보여주고 있다. 얼굴인식 성능을 측정하기 위해서 얼굴식별의 경우 분류를 통해 얻은 라벨 정보와 실제 라벨이 동일한지 여부

를 측정하는 rank-1 식별율(Identification Rate)이 사용된다. 얼굴검증의 경우에는 입력 두 영상이 동일 인물인지 여부를 임계값에 따른 True/False로 측정하여 Receiver Operating Characteristic(ROC) 곡선 분석을 통해 타인 오수락률(FAR: False Acceptance Rate) 대비 본인 수락률(TAR: True Acceptance Rate)을 측정하고 있다. 이때, FAR은 전체 잘못 판단된 경우 중 다른 인물인데 동일 인물로 판단한 비율을 나타내고, TAR은 옳게 판단된 경우 중 동일 인물을 동일 인물로 판단한 비율을 나타낸다. 〈표 2〉의 IJB-A 데이터셋에 대해서는 0.01 FAR에 대한 TAR과 rank-1 식별율, 그리고 MegaFace 및 MegaFace2 데이터셋에 대해서는 10^{-6} FAR에 대한 TAR과 rank-1 식별율을 보여주고 있다.

〈표 2〉 데이터셋 및 얼굴인식 방법에 따른 최근 얼굴인식 성능(%)

데이터셋	얼굴인식 알고리즘	Face Verification	Face Identification
LFW	DeepFace	97.4	—
	VGGFace	98.9	—
	SphereFace	99.4	—
	FaceNet	99.6	—
	CosFace	99.7	—
YTF	DeepFace	91.4	—
	VGGFace	95.0	—
	SphereFace	95.1	—
	FaceNet	97.8	—
	CosFace	97.6	—
IJB-A	DR-GAN	77.4	85.5
	VGGFace	80.5	91.8
	Deep Multi-pose[22]	87.6	84.6
	Template Adaptation[23]	93.9	92.8
	NAN	94.1	95.8
Mega Face	NTechLAB	85.1	78.8
	Light CNN-29[24]	85.1	78.7
	FaceNet	86.5	70.5
	SphereFace	89.1	75.8
	CosFace	96.7	82.7
Mega Face2	3DiVi	66.5	57.0
	NEC	66.8	62.1
	GRCCV	74.8	75.8
	SphereFace	84.2	71.2
	CosFace	86.8	74.1

표 2에서 볼 수 있듯이, LFW 데이터셋에 대해서는 딥러닝 얼굴인식의 첫 사례인 DeepFace가 97.35%를 달성한 이후 최근에는 약 99%로 성능이 매우 포화되었음을 확인할 수 있다. YTF 데이터셋에 대해서는 LFW 데이터셋에 비해 어려운 데이터로 알려져 있음에도 불구하고 최근에는 97%의 인식률을 보여주고 있다. 마찬가지로 IJB-A 데이터셋에 대해서도 높은 성능을 보여주고 있으나, 최근 IJB-B 및 IJB-C 데이터셋이 공개됨에 따라 어려운 얼굴 영상들이 포함되어 상대적으로 낮은 성능을 보이고 있다. 반면, MegaFace 데이터셋에 대해서는 앞서 설명한 세 가지의 데이터셋과 비교했을 때는 rank-1 식별률 70%, TAR 80% 후반대로 상대적으로 낮은 성능을 보여주고 있다. 이는 매우 많은 수의 인물이 등록되어 있을 뿐만 아니라 백만 스케일의 등록되지 않은 인물이 인식에 어려움을 주기 때문이다. 최근 다양한 얼굴인식 기술 개발에 따라 이 성능 역시 빠른 시간 내에 고도화 될 것으로 전망되고 있다.

V. 결론

본고에서는 딥러닝 기반 고성능 얼굴인식 기술 연구 동향에 대해 살펴보았다. 기존의 패턴인식 및 컴퓨터 비전 기술과 비교했을 때, 딥러닝 기반 얼굴인식 기술이 다양한 데이터셋에 대해 높은 성능을 보여주어 다양한 환경에서의 강인함이 입증되고 있다. 향후에는 기존 기술을 활용하여 실제 환경에서 효율적인 얼굴인식을 위해 사람이 움직이는 상황에서의 고도화된 비디오 얼굴인식 기술이 요구될 것이다. 또한, CCTV 기반 관제/감시 시스템에서 나타나는 저 해상도, 블러, 그리고 가리움(Occlusion) 문제에 강인한 실용적인 딥러닝 기반 얼굴인식 기술에 관한 연구가 요구된다. 마지막으로, 얼굴인식 시스템이 실제 환경에서 적용될 때, 환경에 따른 알고리즘의 적응적인 조정이 필요하지만, 연구개발자의 많은 노동이 요구되는 부분이다. 이러한 문제를 다루기

위한 도메인 적응(Domain Adaptation) 기술도 활발히 연구될 필요가 있다.

용어해설

얼굴검증 얼굴인식 기술 중 두 장의 입력 얼굴 영상에 대해 두 영상의 유사도 평가를 통해 동일 인물인지 여부를 판단하는 1:1 추론 방법.

얼굴식별 얼굴인식 기술 중 한 장의 입력 얼굴 영상에 대해 사전에 등록된 N명의 인물과 비교를 통해 입력 영상이 어떤 인물 인지를 판단하는 1:N 추론 방법.

약어 정리

DRGAN	Disentangled Representation Learning-Generative Adversarial Network
FAR	False Acceptance Rate
GAN	Generative Adversarial Networks
IJB	IARPA Janus Benchmark
LFW	Labeled Faces in the Wild
NAN	Neural Aggregation Network
NIR	Near-Infrared
NIST	National Institute of Science and Technology
ROC	Receiver Operating Characteristic
TAR	True Acceptance Rate

참고문헌

- [1] Y. Taigman et al., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *IEEE Conf. Comput. Vision Pattern Recogn.*, Columbus, OH, USA, June 2014, pp. 1701-1708.
- [2] O. Parkhi et al., "Deep face recognition," *British Machine Vision Conf.*, Swansea, UK, Sept. 7-10, 2015, pp. 6-17
- [3] Y. Sun, X. Wang, and X. Tang, "Deep Learning face Representation from Predicting 10,000 Classes," *IEEE Conf. Comput. Vision Pattern Recogn.*, Columbus, OH, USA, June 2014, pp. 1891-1898.
- [4] Y. Sun et al., "Deep Learning Face Representation by Joint Identification-Verification," *Neural Inform. Process. Syst.*, Montreal, Canada, Dec. 2014, pp. 1988-1996.
- [5] Y. Sun, X. Wang, and X. Tang, "Deeply Learned Face Representations are Sparse, Selective, and Robust," *IEEE Conf. Comput. Vision Pattern Recogn.*, Boston, MA, USA, June 2015, pp. 411-438.

- [6] Y. Sun et al., "DeepID3: Face Recognition with Very Deep Neural Networks," arXiv preprint arXiv:1502.00873, 2015.
- [7] F. Schroff et al., "FaceNet: A Unified Embedding for Face Recognition and Clustering," *IEEE Conf. Comput. Vision Pattern Recogn.*, Boston, MA, USA, June 2015, pp. 815–823.
- [8] W. Liu et al., "SphereFace: Deep Hypersphere Embedding for Face Recognition," *IEEE Conf. Comput. Vision Pattern Recogn.*, Honolulu, HI, USA, July 2017, pp. 212–220.
- [9] H. Wang et al., "CosFace: Large Margin Cosine Loss for Deep Face Recognition," arXiv preprint arXiv:1801.09414, 2018.
- [10] Y. Zheng et al., "Ring loss: Convex Feature Normalization for Face Recognition," arXiv preprint arXiv:1803.00130, 2018.
- [11] J. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," arXiv preprint arXiv:1801.07698, 2018.
- [12] L. Tran et al., "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," *IEEE Conf. Comput. Vision Pattern Recogn.*, Honolulu, HI, USA, July 2017, pp. 1415–1424.
- [13] R. He et al., "Learning Invariant Deep Representation for NIR-VIS Face Recognition," *AAAI Conf. Artif. Intell.*, 2017, pp. 2000–2006.
- [14] J. Yang et al., "Neural Aggregation Network for Video Face Recognition," *IEEE Conf. Comput. Vision Pattern Recogn.*, Honolulu, HI, USA, July 2017, pp. 4362–4371.
- [15] G. Huang et al., "Labeled Faces in the wild: A Database for Studying Face Recognition in Unconstrained Environments," Univ. of Massachusetts, Amherst, Technical Report 07-49, Oct, 2007.
- [16] L. Wolf, T. Hassner, and I. Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," *IEEE Conf. Comput. Vision Pattern Recogn.*, Colorado Springs, CO, USA, June 2011, pp. 529–534.
- [17] B. F. Klare et al., "Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A," *IEEE Conf. Comput. Vision Pattern Recogn.*, Boston, MA, USA, June 2015, pp. 1931–1939.
- [18] C. Whitelam et al., "IARPA Janus Benchmark-B Face Dataset," *IEEE Conf. Comput. Vision Pattern Recogn.*, Honolulu, HI, USA, July 2017, pp. 592–600.
- [19] B. Maze et al., "IARPA Janus Benchmark-C: Face Dataset and Protocol," *Intell Conf. Biometrics*, 2018 (To appear).
- [20] I. Kemelmacher-Shlizerman et al., "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," *IEEE Conf. Comput. Vision Pattern Recogn.*, Las Vegas, NV, USA, June 2016, pp. 4873–4882.
- [21] A. Nech and I. Kemelmacher-Shlizerman, "Level Playing Field for Million Scale Face Recognition," *IEEE Conf. Comput. Vision Pattern Recogn.*, Honolulu, HI, USA, July 2017, pp. 3406–3415.
- [22] W. AbdAlmageed et al., "Face Recognition Using Deep Multi-pose Representations," *IEEE Winter Conf. Applicat. Comput. Vision*, Lake Placid, NY, USA, Mar. 2016, pp. 1–9.
- [23] N. Crosswhite et al., "Template Adaptation for Face Verification and Identification," *IEEE Intell. Conf. Automatic Face Gesture Recogn.*, Washington, DC, USA, 2017, pp. 1–8.
- [24] X. Wu et al., "A light CNN for Deep Face Representation with Noisy Labels," *IEEE Trans. Inform. Forensics Security* (Early Access), 2018.