2021-2 컴퓨터비전 기말고사 대체 리포트

딥러닝 기반 얼굴 랜드마크 검출 연구 분석 및 동향

지능기전공학부 스마트기기공학전공
18011817 홍주영

1. 서론

사람들은 얼굴을 통해 의도, 감정 등과 같은 수많은 비언어적 메시지를 알아차릴 수 있다. 이러한 얼굴 정보를 자동으로 추출하기 위해 얼굴 핵심 지점의 위치 파악하는 '얼굴 랜드마크 검출 (Facial Landmark Detection)' 연구가 활발하게 진행되고 있다. 얼굴 랜드마크란 얼굴을 대표하는 눈, 코, 입 등을 의미하며, 랜드마크 검출 기술에서는 (그림 1)과 같이 사전 정의된 랜드마크 포인트의 위치를 찾는다.

얼굴과 관련된 다른 연구 분야는 검출된 랜드마크 기반으로 구축되기도 한다. 예를 들어 얼굴 표정 인식[1]과 머리 포즈 추정 알고리즘[2]은 랜드마크 위치에서 제공하는 얼굴 형태 정보에 의존한다. 또한 눈 주위의 얼굴 랜드마크 포인트는 눈 감지 및 시선 추적을 위한 동공 중심 위치를 알아낼 수 있다[3].

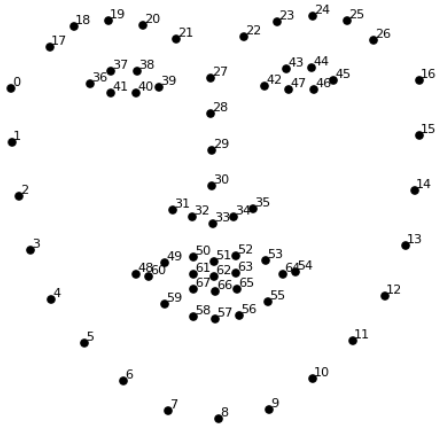

그림 1 사전 정의된 랜드마크 위치

그러나 얼굴 랜드마크 감지는 조명 및 포즈 변화, 사람의 움직임과 원거리 촬영으로 인한 저해 상도 및 블러 문제 등의 도전적인 이유로 성능 향상에 어려움이 있다고 알려져 있다.

최근 딥러닝의 발전으로 랜드마크 검출 방식 또한 딥러닝 기반의 방식으로 바뀌면서 기존에 달성하지 못하던 성능을 갱신하고 있다. 이런 발전에 크게 기여한 것은 기존보다 더 많아진 공개 얼굴 데이터셋과 다양한 환경을 능동적으로 학습할 수 있는 딥러닝 기술의 공개 등을 들 수 있다.

본 보고서에서는 이러한 딥러닝 기술을 이용한 '얼굴 랜드마크 검출' 기술의 대표적인 연구 사례에 대해 살펴보고, 딥러닝 기반 얼굴 랜드마크 검출 기술을 입증하기 위한 도전적인 데이터셋(Dataset)에 대해 살펴보도록 한다.

2. 얼굴 랜드마크 검출 기술 동향

얼굴 랜드마크 검출을 위한 딥러닝 방식은 2013 년 CVPR 에서 발표된 Cascade 기반의 CNN 을 이용한 방식[4]이다. 해당 논문에서는 랜드마크 값 예측 시 Local Minimum 에 빠지는 것을 방지하기 위해서 (그림 2)와 같이 Cascade 형태로 초기 예측 값을 기준으로 복수의 네트워크를 통해

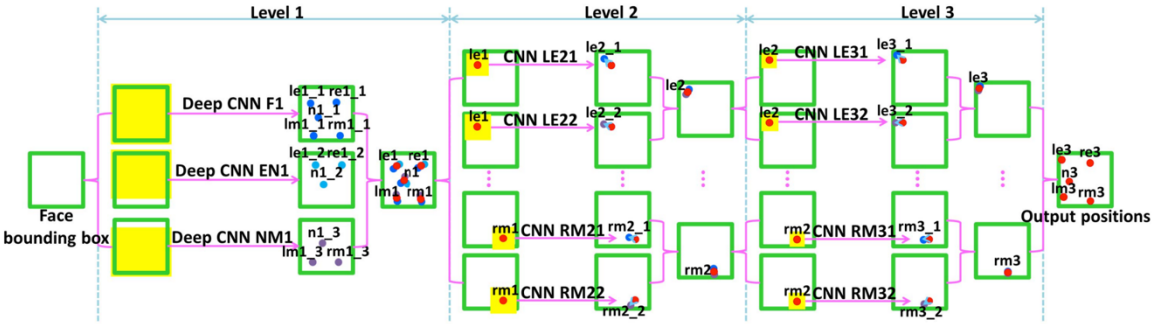

그림 2 3개의Level로 구성된 Deep Convolutional Network Cascade 구조[4]

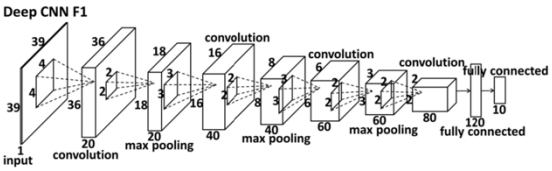

그림 3 네트워크 [4]에서 Level-1의 구조

결과를 보정한다. 즉, (그림 3)이 나타내는 첫번째 레벨에서는 4 개의 컨볼루션 레이어를 가지는 CNN 모델을 적용하여 얼굴 경계 상자에 의해 결정된 얼굴 이미지를 주어진 랜드마크 위치를 예측한 후, 여러 얕은 네트워크가 각 개별 지점을 로컬로 개선하는 Cascade 형태로 결과를 보정하게 된다. 39x39 입력 영상에서 5 개의 랜드마크를 추출한다.

해당 논문 이후 두 가지 방향으로 랜드마크 검출 성능 개선이 이루어졌다. 하나는 Multi-task 학습 기법을 사용하는 방식이다. 2014년 ECCV에서 얼굴의 특성(Attribute)을 Multi-task 기법으로 사용한 얼굴 랜드마크 검출기[5]가 발표되었다. 얼굴 특성(Attribute)은 예를 들어 포즈, 성별, 안경 착용 유무 및 웃음으로 정의된다. 얼굴 인식 성능 개선을 위해서 사용하던 특성 기반 Multi-task 학습 기법을 (그림 4)와 같이 랜드마크 검출기에 적용하여서 성능을 향상시켰다.

랜드마크 검출기를 개선시키는 두번째 방향은 Cascade 형태를 개선하는

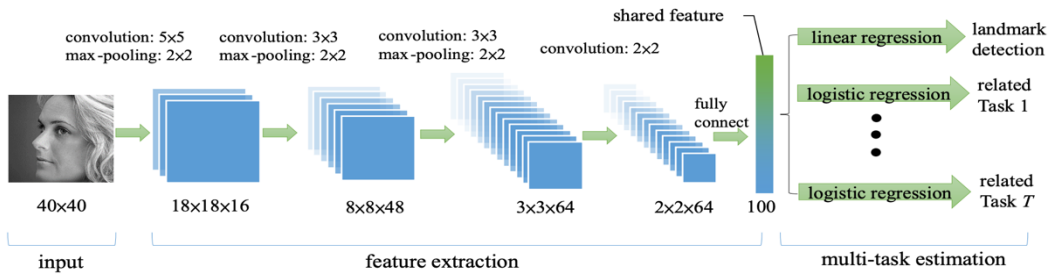

그림 5 Multi-task 학습 기법의 네트워크 구조 [5]

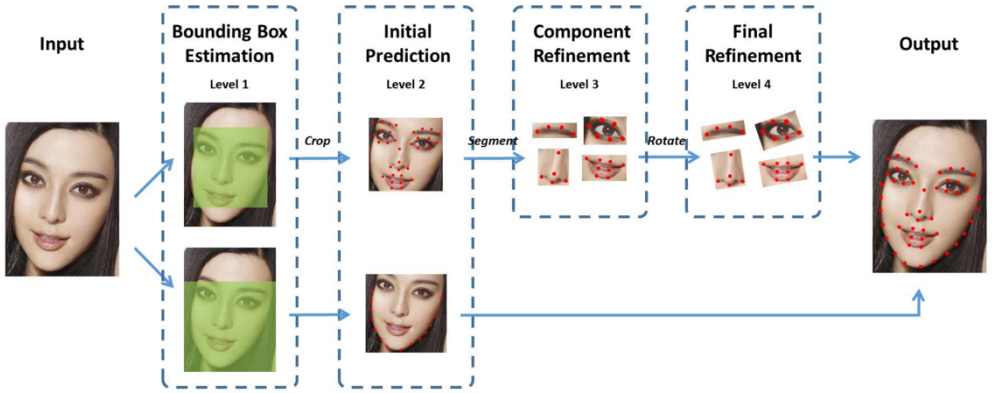

그림 6 Casecade 형식을 개선한 모델의 개요[6]

방식이다. [6]에서는 유사한 계단식 CNN 모델이 더 많은 점(5개가 아닌 68개의 랜드마크)을 예측하도록 구성되었다. (그림 6)과 같이 랜드마크 모든 68개의 점 예측에서 시작하여 점차 예측을 국소적인 얼굴 구성 요소로 분리하는 방식이다.

2014년 ECCV에서 발표된 논문[7]은 Auto Encoder 아키텍처를 랜드마크 검출기에 적용하였고, 랜드마크값이 Local Minimum에 빠져 다른 결과를 도출 하는 것을 피하기 위해 (그림 7)과 같이 영상 크기에 따른 순차적 검출을

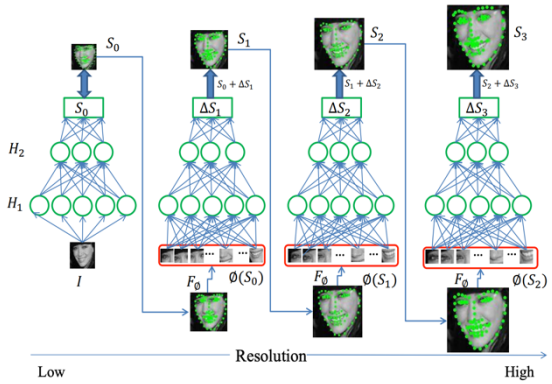

그림 7 Auto Encoder 아키텍처를 사용한 네트워크 구조 [7]

적용하였다. 이러한 Coarse-to-Fine 방식의 개념을 랜드 마크 검출기에 적용하여 50x50 입력 영상에서 68개의 랜드마크를 검출한다.

기존 방법들과는 달리 하이브리드 심층 방법은 투영 모델 및 3D 변형 형상 모델과 같은 3D Vision에 CNN을 결합한다. 2D 얼굴 랜드마크 위치를 직접 예측하는 대신 3D 형상 변형 모델 계수와 머리 포즈를 예측한다. 이후 투영 모델을 통해 2D 랜드마크 위치를 결정한다. 2016년 CVPR에 발표된 [8]에서는 밀도 높은 3D 얼굴 모양 모델이 구성되며 반복적인 계단식 회귀 모델과 심층 CNN 모델을 사용하여 3D 얼굴 모양 및 포즈 매개 변수의 계수를 업데이트한다. 각 반복에서 현재 추정된 3D 매개 변수를 통합하기 위해 비전 투영 모델을 사용하여 3D 형상을 2D로 투영하고 회귀 예측을 위한 CNN 모델의 추가 입력으로 2D 형상을

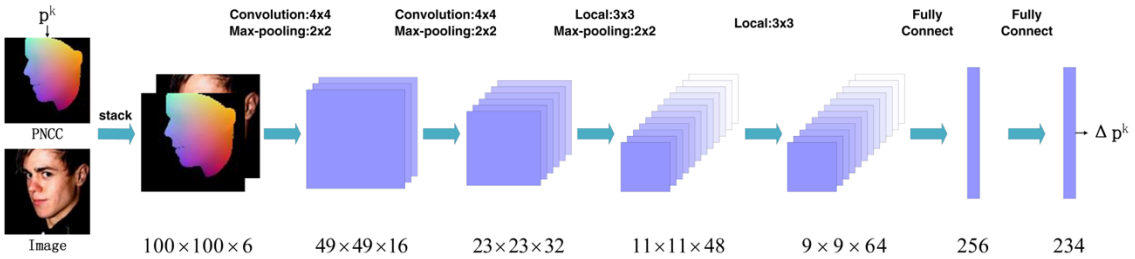

그림 8 3D 형상 모델을 사용한 네트워크 구조 [8]

사용한다.

같은 해, CVPR에 발표된 논문 [9]에서는 [8]과 동일한 계단식 방식을 가지며, 첫 번째 계단식 CNN 모델에서 전체 얼굴 생김새가 3D 모양 매개 변수와 포즈를 업데이트를 예측하는 데 사용되고, 이후 계단식 CNN 모델에서는 로컬 패치를 사용하여 랜드마크를 개선한다.

3. 얼굴 랜드마크 검출 데이터셋

얼굴 랜드마크와 관련된 얼굴 관련 공개 데이터셋은 크게 두 가지 유형이 존재한다. “제어된” 조건에서 수집된 데이터와 “현실에 가까운” 이미지가 있는 데이터이다. 여기서 “제어된” 조건은 사전 정의된 표현식, 헤드 포즈 등 실내에서 수집된 비디오/이미지가 있는 데이터베이스를 의미한다.

가. BioID [11]

제어된 조건에서 수집된 대표적인 데이터로, 23명의 얼굴에서 384 × 286 해상도의 1521개의 그레이 스케일 실내 이미지가 포함되어 있다. 이미지는 다양한 조명과 배경 하에서 촬영된다. 데이터에 존재하는 얼굴은 크지 않은

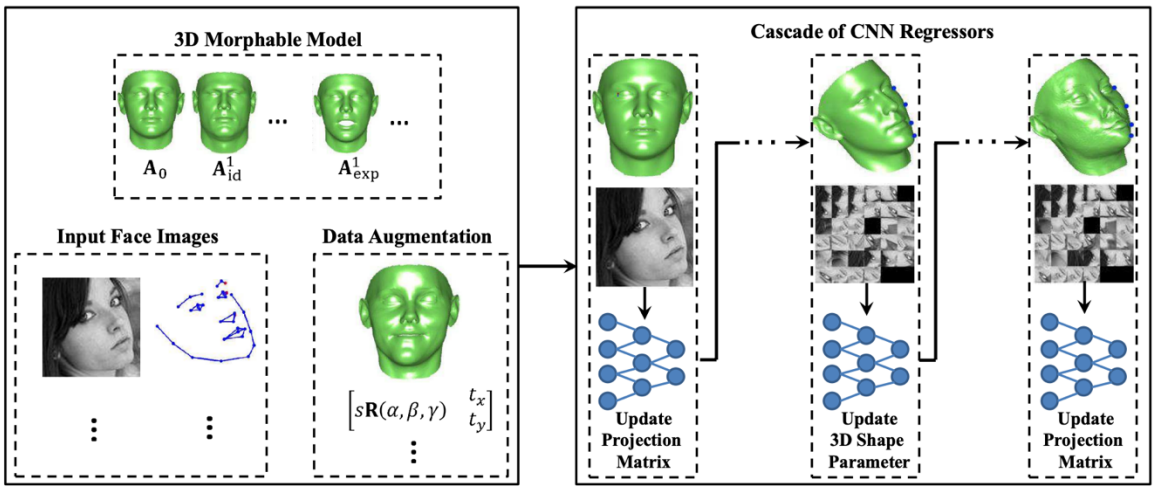

그림 9 3D 형상 모델을 사용한 네트워크 구조 [9]

표정 변화가 존재한다. 해당 데이터는 20개의 랜드마크로 구성된다.

나. AFLW [12]

현실에 가까운 야생(AFLW) 데이터로 달린 얼굴 랜드마크에는 약 25,000개의 이미지가 포함되어 있다. 가시성에 따라 최대 21개의 랜드마크로 구성된다.

다. LFPW [13]

현실에 가까운 데이터로, LFPW(Wild) Labeled Face parts(레이블링된 얼굴 부위)는 29개의 랜드마크가 존재한다. 1,132개의 학습 영상과 300개의 테스트 영상에 대한 68개의 랜드마크로 다시 정의된 것은 [14]에 의해 제공된다..

4. 결론 및 시사점

본 보고서에는 얼굴 랜드마크 검출 연구들을 분석하였다. 딥러닝 기반 방법론들의 등장으로 다양한 데이터셋에 높은 성능을 보여주어 다양한 환경에서의 강인함이 입증되고 있지만, 실제 환경에 적용하기 위해서는 한계가 존재한다. 이런 문제를 해결하기 위해서는 데이터셋 취득부터 고도화된 기술까지 지속적인 연구가 필요하다.

5. 참고문헌

[1] Murphy-Chutorian, E., Trivedi, M. “Head pose estimation in computer vision: A survey.” IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 607-626 (2009)

[2] Pantic, M., Rothkrantz, L.J.M.: “Automatic analysis of facial expressions”, The state of the art. IEEE Transanctions on Pattern Analysis and Machine Intellgence 22(12), 1424-1445 (2000)

[3] Hansen, D.W., Ji, Q.: “In the eye of the beholder: A survey of models for eyes and gaze.” IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 478-500 (2010)

[4] Y. Sun, X. Wang, X. Tang, “Deep Convolutional Network Cascade for Facial Point Detection,” CVPR 2013

[5] Z. Zhang, P. Luo, C. C. Loy, X. Tang, “Facial Landmark Detection by Deep Multi-task Learning,” ECCV 2014

[6] Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q. “Extensive facial landmark localization with coarse-to-fine convolutional network cascade.” In IEEE International Conference on Computer Vision Workshops, pp. 386-391 (2013)

[7] J. Zhang, S. Shan, M. Kan, X. Chen, "Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time face Alignment," ECCV 2014

[8] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S. "Face alignment across large poses: A 3d solution." In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV (2016)

[9] Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV (2016)

[10] BioID. https://www.bioid.com/About/BioID-Face-Database. Accessed: 2015-08-30

[11] Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)

[12] Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(12), 2930-2940 (2013)

[13] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision, 300 Faces in-the-Wild Challenge (300-W). Sydney, Australia (2013)