청각장애인용 방송에서 화자 식별을 위한 얼굴 인식 알고리즘 및 전처리 연구

*김나연, 조숙희, *배병준, 안충현
한국전자통신연구원, *과학기술연합대학원대학교
{boboss, shee, 1080i, hyun}@etri.re.kr

Face Recognition and Preprocessing Technique for Speaker Identification in hard of hearing broadcasting

*Nayeon Kim Sukhee Cho *Byungjun Bae ChungHyun Ahn
Electronics and Telecommunications Research Institute
*Korea University of Science and Technology

요 약

본 논문에서는 딥러닝 기반 얼굴 인식 알고리즘에 대해 살펴보고, 이를 청각장애인용 방송에서 화자를 식별하고 감정 표현 자막을 표출하기 위한 배우 얼굴 인식 기술에 적용하고자 한다. 우선, 배우 얼굴 인식을 위한 방안으로 원샷 학습 기반의 딥러닝 얼굴 인식 알고리즘인 ResNet-50 기반 VGGFace2 모델의 구성에 대해 이해하고, 이러한 모델을 기반으로 다양한 전처리 방식을 적용하여 정확도를 측정함으로써 실제 청각장애인용 방송에서 배우 얼굴을 인식하기 위한 방안에 대해 모색한다.

1. 서론

얼굴 인식이란 얼굴을 포함하고 있는 이미지 혹은 비디오에서 얼굴 영역을 검출하고, 특징을 분석하여 인물을 식별하는 기술이다[1]. 딥러닝 기술의 발전과 함께 얼굴 인식 기술 또한 사람의 얼굴 인식 능력을 상회하게 되었으며[2], 출입국 시스템, 보안, 결제시스템, 미디어에서의 유명인 인식 등 다양한 분야에서 활용되고 있다[3]. 특히, 영상 미디어 분야에서는 화면 정보 분석을 통한 하이라이트 클립 자동 생성, 배우 인식을 통한 화자 식별에 얼굴 인식 기술이 활용되고 있으며, 이를 자막방송, 수화방송, 화면해설방송 등 시·청각장애인을 위한 장애인방송에 적용하기 위한 연구가 진행되고 있다[4, 5].

본 논문에서는 청각장애인용 감정표현 자막방송에서 화자를 식별하기 위한 배우 얼굴 인식 기술을 연구한다. 먼저, 널리 사용되는 얼굴 인식 알고리즘과 정확도 향상을 위한 다양한 이미지 전처리 방식에 대해 살펴보고, 전처리 방식별 각 얼굴 인식 모델의 정확도를 실험하여 실제 자막방송에서 이용하기 위한 방안을 모색한다.

2. 원샷 학습 기반 얼굴 인식 알고리즘

본 논문에서는 끊임없이 창작되는 영상 콘텐츠에서 수많은 배우를 인식하기 위한 이미지 인식 방식으로 일반적인 분류(classification) 학습이 아닌 원샷(one-shot) 학습을 이용한다. 표준 분류 학습은 모든 클래스에 대해 확률 분포가 생성되며 새로운 클래스를 분류하기 위해서 새로운 훈련이 필요하지만, 원샷 학습은 동일한 가중치(weights)와 아키텍처를 공유하는 두 개의 합성곱 신경망을 이용하는 샴 네트워크(siamese network) 방식[6]을 통해 두 개의 입력 이미지가 서로 구분될 수 있는 특성(feature)을 추출하여 유사성을 측정하거나 군집(clustering)하여 클래스를 분류할 수

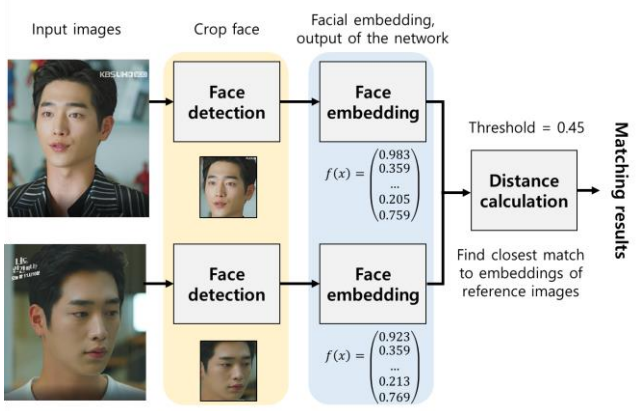

그림 1 샴 네트워크 기반 원샷 학습 얼굴 인식기 구조

있으므로, 새로운 배우가 지속적으로 추가되거나 적은 수의 이미지로 얼굴을 인식해야 하는 시스템에 유용하다[7].

그림 1 은 샴 네트워크 기반 원샷 학습 얼굴 인식기 구조를 나타내며, 크게 얼굴 검출(detection), 임베딩(embedding), 유사도(similarity) 및 거리(distance) 측정 단계로 이루어진다. 이러한 구조의 대표적인 모델로는 FaceNet[8], VGGFace[9]을 들 수 있으며, 본 논문에서는 ResNet-50[10] 아키텍처를 기반으로 VGGFace2 데이터셋을 학습시킨 모델, 통칭 VGGface2 [11]을 사용한다.

3. 실험 결과

본 논문에서는 샴 네트워크 기반 원샷 학습 얼굴 인식기 실험을 위해 S^3FD 얼굴 검출 모델[12]과 tensorflow-keras 기반으로 사전 학습(pre-trained)된 VGGface2 얼굴 인식 모델을 이용하고, 한국 드라마에서 검출한 주연 배우 4 명의 정면 얼굴 총 2,559 개와 각 배우의 인터뷰 사진 한 장씩을 성능 검증을 위한 데이터셋으로 사용하였다.

먼저 영역 보간법으로 224*224 리사이즈된 컬러 배우 얼굴 이미지를 원본(original) 이미지로 두고 성능을 측정한 뒤, 이를 얼굴 정렬(alignment), 샤프닝(sharpening), 패딩(padding), CLAHE 등 전처리를 적용하여 얻은 성능 결과와 비교하였다. 실험 결과, 그림 2 와 같이 배우 및 전처리 기법에 따른 accuracy 는 80.32%에서 98.03%로 다양한 편차를 보였다. 얼굴 정렬 전처리의 경우 배우 1, 2 는 원본 이미지를 사용했을 때와 대비하여 0.33 %에서 1.55% 더 높은 정확도를 보였고, 배우 3, 4 는 원본 이미지 결과가 더 높은 정확도를 보였다. CLAHE 는 다른 전처리 기법들뿐만 아니라 원본 실험 결과에 비해서도 낮은 성능을 보였다.

배우별 정확도 또한 큰 편차를 보였으며, 특히 보조축을 통해 표현된 배우 4 의 경우 60.07%-80.32%의 정확도로 다른 배우들에 비해 최고 37% 더 낮은 성능을 나타냈다. 해당 배우의 경우 격한 감정을 드러내거나 눈을 감은 데이터셋의 비율이 다른 배우에 비해 30%에서 40% 더 많았으며, 해당 데이터의 경우 대부분 False-Negative 의 결과를 보였다.

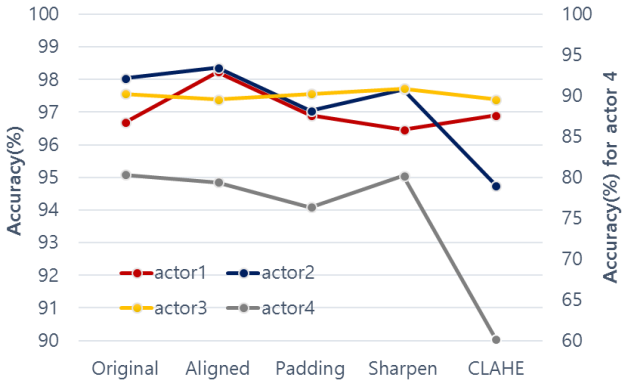

그림 2 배우별 샴 네트워크 기반 원샷 학습 얼굴 인식 실험

4. 결론

본 논문에서는 여러 전처리 기법을 적용한 배우 얼굴 데이터셋을 샴 네트워크 기반 원샷 학습 얼굴 인식기로 인식하고 실험 결과를 보여준다. 전체적으로 봤을 때, 얼굴 정렬 전처리를 한 얼굴 이미지가 모든 실험에서 Top-2 정확도를 나타냈으며, 배우 1 의 경우 Top-1 정확도가 1.55% 향상되는 결과를 보여 얼굴 정렬 기법이 얼굴 인식 실험에서 성능을 개선시킬 수 있음을 확인했다. 그러나, 데이터셋이 옆모습을 제외한 정면에 가까운 순수한 얼굴만으로 구성되었으므로, 다른 각도의 얼굴 포즈와 모자, 안경 등 소품, 격한 표정 이미지에서도 얼굴을 식별할 수 있는 방안을 모색해야한다. 또한, 본 논문에서는 VGGFace2 모델만을 이용하였으나, DeepFace[14], Sphere 및 Pose-Robust 얼굴 인식기인 DREAM 등 다양한 모델에서의 성능 평가가 필요하며, pre-trained 모델 뿐만 아니라, 국내 데이터셋으로 튜닝된 모델에 대한 실험이 필요하다.

Acknowledgement

"이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2019-0-00447, 시청각 장애인의 방송시청을 지원하는 감성표현 서비스 개발)"

참 고 문 헌

[1] 김형일, 문진영, 박종열. 딥러닝 기반 고성능 얼굴인식 기술 동향. [ETRI] 전자통신동향분석, 33(4), 2018.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[3] 황원준. 딥러닝 기반 얼굴 인식 최신 기술 동향. 전자공학회지, 46(8), 15-22, 2019.

[4] 안충현. 장애인방송 기술개발 현황. [ETRI] 전자통신동향분석, 34(3), 2019.

[5] 김성훈, 안충현, 서봉석, 현창중, 김동. ATSC3.0 기반 청각장애인을 위한 감정표현자막방송 핵심기술에 관한 연구, 대한전자공학회 추계학술대회 논문집, 2019.

[6] CHOPRA, Sumit; HADSELL, Raia; LECUN, Yann. Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, p. 539-546. 2005.

[7] Zhou J., Chen J., Liang C., Chen J. One-Shot Face Recognition with Feature Rectification via Adversarial Learning. In: Ro Y. et al. (eds) MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science, vol 11961. Springer, Cham, 2020.

[8] SCHROFF, Florian; KALENICHENKO, Dmitry; PHILBIN, James. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 815-823. 2015.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman Deep Face Recognition British Machine Vision Conference, 2015.

[10] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p. 770-778, 2016.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman VGGFace2: A dataset for recognising face across pose and age International Conference on Automatic Face and Gesture Recognition, 2018.

[12] ZHANG, Shifeng, et al. S3fd: Single shot scale-invariant face detector. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 192-201.