https://doi.org/10.7840/kics.2022.47.12.2156

트랜스포머 기반의 판소리 소리꾼 검출과 얼굴 표정 인식

오 문 흠◆, 이 혜 정*, 이 준 환°

Transformer-Based Detection and Facial Expression Recognition of a Pansori Singer

Wenqin Wu◆, Hyejeong Lee*, Joonwhoan Lee°

요 약

인공지능 기술을 이용하는 음악 연구가 활발하게 진행되고 있다. 본 연구에서는 한국 전통음악인 판소리의 발림 동작과 얼굴 표정의 인식을 목표로 한다. 이를 위해 동영상을 구성하는 매 프래임에서 소리꾼 객체를 검출하고, 소리꾼을 포함하는 프래임에서 동작 및 얼굴 표정 인식을 시도한다. 단 동작 인식은 이전 논문[1]에서 언급하였기 때문에 본 논문에서는 소리꾼 얼굴 표정 인식에 관련된 과정만을 강조하여 언급한다. 본 논문에서 부족한 주석 데이터 때문에 소리꾼 검출에서는 MS COCO 데이터를 혼합하여 활용하였으며, 얼굴 표정 분류에는 ImageNet 데이터셋과 RAF-DB 데이터를 활용하여 MAE 자기 지도 학습 모델로 구성한 ViT 트랜스포머를 백본으로 활용하였다. 소리꾼 검출의 경우 IoU 문턱치 75%에서 약 91.2%의 평균 정밀도를 얻었으며, 4개 범주의 얼굴 표정의 분류에서 78.44%의 정확도를 달성했다. 본 연구의 결과는 문화 콘텐츠 보존이나 교육에 의미 있는 활용이 예상된다.

Key Words : Pansori, Transformer, Self-supervised training, Facial expression classification

ABSTRACT

Our work aims to recognize the singer's motions and facial expression of Pansori, a traditional Korean music. In order to achieve our goal, the region of a singer is at first detected from every video frame, then the motion and facial expression are classified based on still image analysis in the detected region. In order to overcome the difficulty due to insufficient labelled data, the person class of MS COCO data is mixed up with the collected Pansori dataset for detecting Pansori singer, and ViT(Visual Transformer) pre-trained backbone is adopted to construct a singer's facial expression classifier after self-supervised training in MAE model with ImageNet and RAF-DB dataset. For singer detection, 91.2% of AP has been achieved under 75% of IoU threshold. Also, 78.44% of accuracy has been obtained in Pansori singer's facial expression classified into 4 categories. The results of this study are expected to be meaningful for preservation or education of cultural content.

※ 본 연구는 한국연구재단 중견연구과제(NRF-2021R1A2C200689512)에 의해 지원되었음.
◆ First Author : Chonbuk National University Department of Computer Science and Engineering, ryanwu200@gmail.com, 학생회원
° Corresponding Author : Chonbuk National University Department of Computer Science and Engineering, chlee@chonbuk.ac.kr, 종신회원
* Chonbuk National University Department of Korean Music, jina2747@naver.com
논문번호 : 202207-147-C-RE, Received July 21, 2022; Revised September 20, 2022; Accepted October 20, 2022

Ⅰ. 서 론

인류에게 있어 노래는 문화 현상이며 감정표현의 수단으로 어울리는 감정을 표현하기 위해 가락이나 내용에 따라 필요한 동작과 표정을 취하기도 한다. 이는 동서양을 막론하고 어떤 음악에서도 예외가 없이 관찰되는 현상이며 1인극 형태인 우리 고유의 판소리에서도 연기자이며 가수인 소리꾼은 발림 동작과 얼굴표정을 통해 관객들의 호응을 얻어내며 아니리와 소리를 이어간다. 본 연구는 판소리 동영상 비디오로부터 소리꾼을 찾아내고 발림 동작을 인식하고 소리꾼의 얼굴을 찾아 얼굴 표정을 자동으로 분류하는 인공지능 개발을 목표로 한다. 이전 논문에서의 발림 동작 인식 [1]을 언급하였기 때문에 본 논문에서는 소리꾼 얼굴 표정 인식과 이를 위한 소리꾼과 소리꾼의 얼굴 검출 과정을 다룬다. 따라서 이들 방법들을 조합하면 소리꾼 동작과 표정을 인식하여 소리꾼의 연기 특징을 분석하는 등 판소리 연구에 활용할 수 있다. 이러한 연구는 인공지능을 활용하여 판소리 소리꾼 등의 연기를 분석함으로써 분석결과와 함께 우리 전통음악을 보존하는데 도움이 되며, 소리꾼 교육 등 전승에도 활용할 수 있으며, 오페라 또는 뮤지컬 배우, 가수의 노래하는 몸짓과 얼굴을 인식하고 표정을 분류하는 연구로 확장될 수 있다.

한편 얼굴표정 인식은 컴퓨터 비전의 전형적인 문제로 전통적인 인식 방법과 최근의 데이터 기반의 딥러닝을 이용하는 접근방법 등이 활용되고 있다. 일반적으로 판소리 비디오부터 다량의 소리꾼 얼굴표정과 주석을 얻기 쉽지 않다. 따라서 본 논문에서는 데이터가 많지 않을 경우 사용하는 MAE(Masked Auto-Encoder)[2] 자기 지도(self-supervised)모델을 이용하여 ViT(Visual Transformer) 백본을 구성하고 이를 활용하여 분류기를 구성하였다.

일반적으로 얼굴표정을 인식하기 위해서는 클립 단위의 동영상 기반의 인식 방법[3-6]과 동영상을 구성하는 매 프레임에서 소리꾼을 찾아내고 이 소리꾼 얼굴 이미지를 분류하는 방법을 활용할 수 있다. 그러나 얼굴표정이 시작되고 끝나는 비디오 프레임을 찾고 이를 클립으로 묶어내어 해당 클립의 표정을 인식하는 동영상 기반의 인식 방법은 정확성과 표정의 동적인 다양성 측면에서 유리하지만 표정의 시작과 끝이 모호하기 때문에 동영상 세그먼트 데이터 구성이 어렵다.

따라서 본 논문에서는 비교적 단순한 방법인 소리꾼이 존재하는 프레임을 찾고, 정지된 프레임에서 얼굴을 찾아 소리꾼의 표정을 인식하는 방식을 활용한다. 즉 동영상 프레임을 구성하는 매 프레임에서 소리꾼이 들어있는 프레임과 소리꾼 영역을 트랜스포머 기반의 객체 검출 기법인 DETR[7]를 이용하여 자동으로 찾아내고, 소리꾼 얼굴을 추출하여 ViT(Visual Transformer) 기반의 표정 인식을 수행한다.

일반적으로 소리꾼 검출과 얼굴 표정 인식에 필요한 주석이 부가된 규모가 큰 데이터 확보가 용이하지 못하다. 따라서 소리꾼 검출의 경우 청중의 경우는 MS COCO 데이터 셋의 “사람(person)” 데이터를 추가적으로 활용하여 객체 검출을 시도하였으며, 소리꾼 얼굴 표정인식에는 ImageNet과 RAF-DB(Real-world Affective Faces Database) 데이터 셋으로 MAE(Masked Auto-Encoder) 방법으로 자기 지도 학습방법을 통해 ViT 트랜스포머 백본을 얻고 이를 활용하여 분류기를 구성하였다.

본 논문에서 제안된 방법은 안정적인 소리꾼 프레임과 소리꾼 영역 검출을 제공하며, 얼굴 검출 결과와 표정 인식 결과를 제공한다. 본 논문의 동영상에서의 소리꾼 검출은 IoU 문턱치 75%에서 91%의 검출 성능을 보였으며 판소리 소리꾼의 네 부류의 얼굴 표정 분류에서는 정확도 78.44%를 달성하였다.

본 연구는 인공지능 기법으로 판소리 비디오에서 얼굴 표정을 자동으로 분류하는 최초의 논문으로 문제를 제기하는 데 의미가 있으며 향후 데이터 셋 구축과 보다 정확하고 효율적인 알고리즘 개발을 통해 판소리 소리꾼 얼굴 표정 분석과 판소리 교육 등에 활용될 것으로 기대된다.

Ⅱ. 관련 연구

성공적인 표정 인식을 위해서는 계산량, 시공간 정보의 사용, 촬영 각도, 빛의 세기나 강도 등 도전적인 문제들을 해결해야 한다. 고전적인 방법에서는 LBP(Local Binary Patterns)[8], LBP-TOP[9], NMF(Non-negative Matrix Factorization) [10], sparse 학습[11] 등으로 잘 설계된 handcraft 특징을 이용해 왔다. 반면 딥러닝에서는 보통 얼굴 정렬(face alignment), 얼굴 정규화(face normalization) 등 전처리를 수행한 후 표정을 분류하기 위해 목적에 맞게 설계된 레이어를 장착한 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), Deep auto-encoder, GAN(Generative Adversarial Network) 등 딥러닝 모델로 특징을 추출해서 분류한다[12]. 본 논문에서는 주석이 붙은 학습 데이터가 제한적이기 때문에 MAE 모델로 자기 지도 학습으로 얻은 ViT 트

랜스포머 구조를 백본으로 활용하여 특징을 추출하고 표정 분류기 구성에 활용한다.

심리학에서 표정을 모델링하는 방법은 CES(Categorical Emotion States)와 DES (Dimensional Emotion Space)로 분리된다. CES는 표정을 몇 가지 확정된 범주로 나누고 분류하는 방식으로 에크만의 여섯 가지 기본 표정("행복", "놀람", "분노", "혐오", "공포", "슬픔") 등 방법 등의 부류로 분류한다. 반면에 DES는 연속적인 다차원의 공간에 표정을 매핑하는데, 러셀의 VA(Valence-Arousal) 공간[13]를 많이 사용한다. DES는 연속적인 값으로 표정을 표하기 때문에 미세한 표정도 표현할 수 있으나 연속적인 표정을 표현하는 값이 주관적이고 모호하기 때문에 주석을 얻기 어렵다.

본 논문에서는 CES를 채택하고 동아시아의 감성 표현인 희로애락(喜怒哀樂) 에 "무표정"을 추가해서 적용하고자 하였으나 "노여움(怒)"에 해당하는 표정이 판소리에 많이 나타나지 않아 "기쁨(喜)","슬픔(哀)","즐거움(樂)","무표정"으로 4개의 범주로 한정하였다.

공식적인 얼굴 표정 인식 데이터가 많지만 여기서 본 연구에 사용했거나 관련된 데이터 두 종류를 소개한다. 첫째, RAF-DB[14] 데이터 셋은 인터넷에서 수집했으며, 29,572 장 영상과 주석(에크만의 여섯 가지 기본 표정)을 포함하는 데이터 셋이다. SFEW(Static Facial Expressions in the Wild)[15] 데이터 셋은 본 논문의 소리꾼 얼굴과 같이 영화 또는 동영상에서 캡처하여 수집한 영상으로 다양한 촬영 각도를 가지며 조명도 변화가 심한 1,398 장 영상과 주석(에크만의 여섯 가지 기본 표정)을 포함하는 데이터 셋이다.

한편 최근 딥러닝 분야에서 주목 많이 받고 있는 자기 지도 학습은 사전 학습(proxy task) 단계와 내림 과제(downstream task)단계로 구성되어 있다. 사전 학습에서는 주석 없는 데이터를 기반으로 데이터의 특징으로부터 스스로 학습을 수행하여 필요한 특징을 추출하며, 내림 과제 단계에서 적은 양의 주석 데이터를 활용하여 미세조정(finetuning)을 수행한다. 주석 있는 데이터 수집의 어려움 때문에 자기 지도 학습은 수 많은 주석 없는 데이터를 활용 적은 주석 데이터를 활용하여 성능을 향상 시키는 데 도움을 준다. 본 연구에서는 얼굴표정 인식을 위한 백본인 ViT 트랜스포머 구조를 MAE 모델을 활용하여 자기 지도 학습을 수행하며, 이때 ImageNet과 RAF-DB 데이터 셋은 주석이 없는 데이터로 간주된다.

본 논문은 판소리 발림 동작 및 얼굴 표정에 대한 인공지능 기술을 개발하기 위한 논문으로 현재까지 국내외적으로 유사한 연구를 찾아보기 어렵다. 따라서 본 연구에서는 가장 최신의 인공지능 모델인 트랜스포머 모델과 자기 지도 학습을 적용하여 그 가능성을 확인해 보았다.

Ⅲ. 제안된 소리꾼 얼굴 표정 인식 방법

본 논문에서 제안하는 소리꾼 얼굴표정 인식 방법의 파이프라인은 1 단계의 소리꾼 검출과 2 단계의 얼굴 검출과 표정 분류로 구성된다. 제안된 방법에서는 우선 소리꾼만 있는 이미지뿐만 아니라 고수 등 다른 등장 인물도 포함한 판소리 공연 전체 비디오에서 소

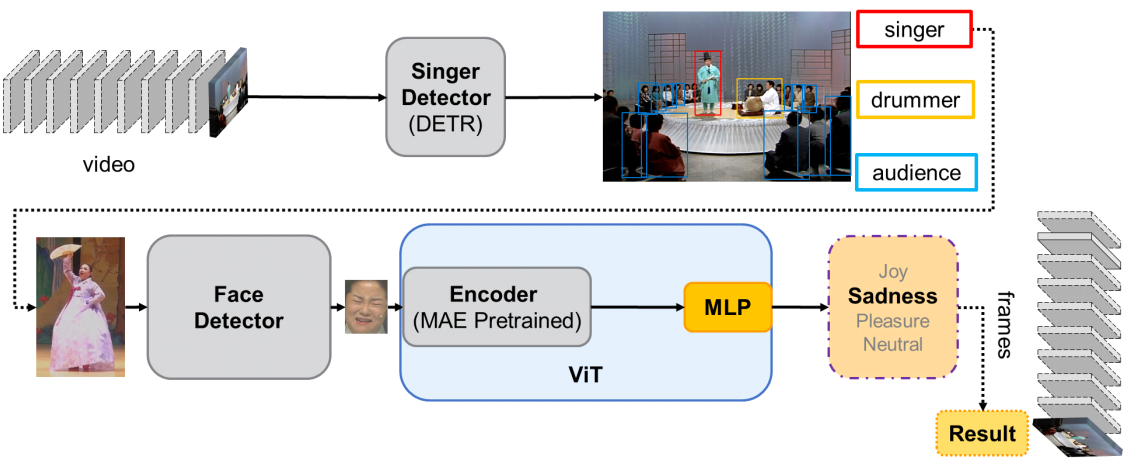

그림 1. 제안된 얼굴 표정 인식 방법의 파이프라인
Fig. 1. Pipeline of Proposed Facial Expression Recognition of Pansori Singer

리꾼을 검출하며, 소리꾼의 얼굴과 얼굴 표정을 분류한다.

그림 1의 소리꾼 검출기(Singer Detector)는 연속되는 비디오 프레임에서 소리꾼, 고수, 청중 등을 검출하고 소리꾼 부분의 영상은 발림동작 인식[1]에 활용되고 얼굴 검출기(Face Detector)를 통해 얻은 소리꾼의 얼굴은 ViT 분류기를 통해 그 인식 결과를 제공하며, 전체적인 발림과 얼굴 표정을 화면에 표시한다.

3.1 DETR를 이용한 소리꾼 검출

동작 및 표정 분류를 수행하기 전에 먼저 비디오 프레임에서 동작과 표정을 취하는 소리꾼을 찾아야 한다. 판소리 동영상에 등장하는 인물은 소리꾼, 고수, 청중(사회자, 스텝들도 청중으로 간주함)들이 있다. 본 논문에서는 소리꾼만 찾은 것이 목적이지만 소리꾼을 잘 찾기 위한 소리꾼 검출을 위해 소리꾼, 고수, 청중 3개의 클래스로 나누어 객체 검출을 수행하였다.

3.1.1 소리꾼 검출을 위한 데이터 셋의 구성

소리꾼과 고수 클래스의 데이터는 인터넷에서 수집하여 총 2,013장으로 구성되었다. 소리꾼만 포함한 영상이 739장이었으며 고수까지 포함한 영상은 1,274장이었다. 청중 클래스의 학습 데이터의 부족을 보충하기 위해 그림 2와 같이 MS COCO 데이터셋에서 크기가 큰인 "사람" 클래스 이미지들을 청중 클래스로 간주해서 추가하여, 총 3,013장을 7대3으로 학습 데이터와 테스트 데이터로 나누었다. 그림 2는 각각의 객체에 대한 데이터셋 및 학습과 검증에 사용한 데이터 수를 표현한다.

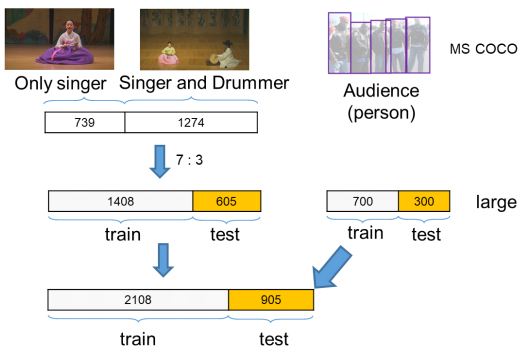

그림 2. 소리꾼 검출 데이터의 구성
Fig. 2. Data Construction of Pansori Singer Detection

3.1.2 소리꾼 검출 방법[1]

언급한 바와 같이 소리꾼 동작 및 표정을 분류하기 위해서는 동영상을 구성하는 매 프레임에서 소리꾼을 검출해야 한다. 객체 검출에 관한 연구는 2-stage의 R-CNN 모델 [16] 에서 1stage의 YOLO(You Only Look Once) 모델[17]까지 수많은 연구가 있어 왔으며, 최근 몇 년간 자연어 처리 분야에 많이 적용된 트랜스포머[18] 모델이 컴퓨터 비전 분야에서도 활발하게 적용되고 있다. 본 연구에서는 객체의 크기 변화가 심하지 않을 경우 성능이 우수하며 NMS(Non-Maximum Suppression)연산이 제외되어 검출 객체 수가 작은 경우 속도 면에서도 유리한 DETR 모델을 사용하였다.

DETR 모델은 그림 3 과 같이 검출 문제를 직접 end-to-end방식인 집합 예측(set prediction) 문제로 간주하여 트랜스포머의 encoder-decoder 구조로 N개의 객체 바운딩 박스를 예측한다. 여기서 N은 사전에 설정한 이미지 중에 존재하는 객체의 개수보다 큰 정수이다. 학습은 예측한 바운딩 박스와 참값(ground truth)의 비교는 이분 그래프(bipartite graph)의 매칭을 수행하도록 손실 함수를 정의하고 이를 감소시키는 방식으로 진행된다.

본 논문에서는 소리꾼을 검출하기 위해 객체 크기 변화가 크지 않고 검출할 객체수가 한정된 경우, Faster R-CNN 모델의 성능과 비슷하며 실행 속도도 더욱 빠른 DETR를 사용하였다. 즉 DETR는 작은 객체에 대한 식별력이 Faster R-CNN보다 못하지만 본 논문에서 검출해야 할 소리꾼 객체는 일반적인 판소리 동영상에서 작지 않으며 비슷한 크기를 가진다.

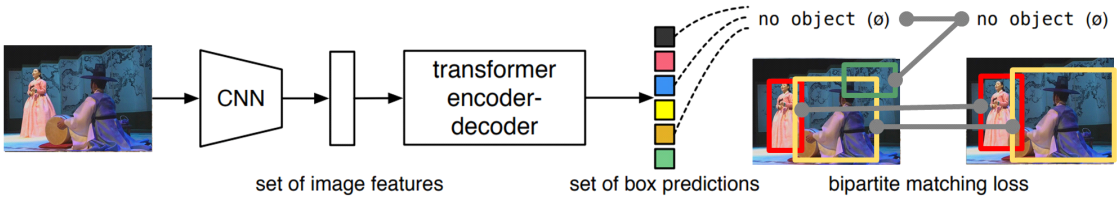

그림 3. DETR 모델의 파이프라인
Fig. 3. Pipeline of DETER Object Detector

3.2 소리꾼 얼굴 표정 인식 방법

검출된 소리꾼의 전체에 해당하는 영상은 발림 동작을 인식[1]하는데 활용되고, 얼굴 표정을 인식하기 위해서는 얼굴 부분의 영상을 추출해야 한다. 본 논문에서는 SCRFD(Sample and Computation Redistribution for Efficient Face Detection) 방법[19]를 적용하여 영상에서 얼굴부분을 추출하였다. ResNet 구조의 이 방법에서는 학습데이터의 재 배분(sample redistribution)과 백본, 넥, 헤드에 계산량 재분배(computation redistribution) 등을 통해 적은 계산량으로 최고(state-of-art)의 얼굴검출 성능을 구현한다. 이렇게 검출된 얼굴은 본 논문에서 제안하는 ViT 기반의 얼굴 표정 인식 방법을 통해 분류된다.

3.2.1 얼굴 표정 데이터 셋의 구성

서론에서 언급한 바와 같이 본 논문에서는 판소리 소리꾼의 얼굴표정을 "기쁨(喜)", "슬픔(哀)", "즐거움(樂)", "무표정" 4개 부류로 나눈다. 즉 喜怒哀樂에서 "화냄(怒)" 표정이 판소리 소리꾼에서는 잘 나타나지 않기 때문에 이를 제외하였다. 이들은 러셀의 VA 감성공간[13]에서 喜哀樂은 각각 1상한, 3상한, 4상한에, "무표정"은 공간의 중앙인 원점 부분에 해당한다. 그림 4는 이들 4개 범주의 대표적인 표정을 보여주고 있다. 여기서 주목해야 할 점은 "기쁨"과 "즐거움"으로 의미적인 차이가 크지 않다는 것이다. 본 연구에서 "즐거움"은 단순하게 잔잔한 웃음으로 정의하였다.

실제 대규모 판소리 소리꾼의 표정 데이터를 얻기 위해서는 많은 노력이 필요하다. 본 연구에서는 YouTube 사이트에서 판소리 공연 동영상을 얻었으며, 이들 동영상으로부터 연구에 무관한 프레임들을 제거하기 위해 소리꾼 별로 편집을 수행하였다. 편집된 동영상들의 관련 정보는 표 1과 같다. 이렇게 얻은 편집 영상은 주석자들이 두 번의 관찰을 통해 주석 작업을 수행하였다. 첫 번째 관찰에서는 주석을 부여하지 않고 보기만 하면서 공연에 등장하는 소리꾼의 전체적인 감정 흐름을 대강 파악하도록 하였으며, 두 번째 관찰에는 동영상을 보면서 프레임을 추출하고 선

그림 4. 정의된 얼굴 표정의 범주
Fig. 4. Categories of Defined Facial Expressions

표 1. 편집된 동영상 정보
Table 1. Information of Obtained Video Frames

항목	내용
동영상 길이	02분41초 ~ 124분39초
동영상 수	31
소리꾼 수	26
마당	흥부가,심청가,춘향가,수궁가
대표 동영상 사이트	https://youtu.be/KxBWPzu5VRc
	https://youtu.be/xYTyZBXrd-Y
	https://youtu.be/PdU4f7Gdnuo
	https://youtu.be/CsYq2LTPzj8

표 2. 얼굴 표정 데이터
Table 2. Facial Expression Data of Pansori Singers

영상	프레임 수/인원
이미지 수("기쁨")	153
이미지 수("슬픔")	194
이미지 수("즐거움")	267
이미지 수("무표정")	173
주석자 수	2(명)

정된 4개의 범주로 주석을 부여하도록 하였다. 주석을 붙인 이미지 데이터의 정보는 표 2와 같다.

3.2.2 데이터 부족을 극복하기 위한 MAE (Masked Auto-Encoder) 자기 지도 학습

표 2의 데이터로는 수많은 파라메터를 포함한 딥러닝 모델의 충실한 학습에 부족하다. 그러나 더 많은 데이터를 얻기 위해서는 많은 인력과 시간이 필요하기 때문에 본 논문에서는 자기 지도 학습을 통해 ViT의 백본을 구성하고 이를 활용하여 부족한 데이터 문제를 극복하려 하였다. 특별히 본 연구에서는 MAE를 이용한 ViT 백본을 학습하고 이를 활용하여 얼굴표정 분류기를 구성하였다.

MAE에서는 자연어 처리 분야에서 마스크 언어 모델링(Masked Language Modeling) 과제를 수행하는 식으로 BERT[20] 학습을 통해 영상 패치를 예측하는 방식을 취한다. 즉 영상을 패치 단위로 쪼개고 일부를 마스킹해서 인코더-디코더 모델을 통해 마스킹해기 전에 영상 체를 예측하는 마스크 영상 모델링(Masked Image Modeling)과업을 수행함으로써 자기 지도 학습을 수행한다. MAE에서는 BEiT[21]는 다르게 예측하려는 표적이 영상 토큰이 아니고 패치라는 점이며 직접 마스킹된 화소를 예측함으로써 본 얼굴 표정 인

식 ViT 백본을 구성하기에 더 적합하다. 그림 5는 MAE의 구조를 보여준다. 여기서 인코더 ViT의 트랜스포머이며 디코더 역시 트랜스포머 구조를 가지며 일반적으로 인코더에 비해 단순한 구조로 장착된다.

본 논문에서는 부족한 판소리 소리꾼 표정 학습 데이터 대신에 ImageNet 데이터 셋과 RAF(Real-world Affective Faces) 데이터베이스를 사용하여 자기 지도 학습을 수행했다. 여기서 ImageNet 데이터 셋은 범용의 영상 데이터 셋으로 일반적으로 사전 학습(Pre-training)에 사용되는 데이터이며, RAF는 에크만의 6가지 기본 표정을 포함하는 29,572장의 얼굴영상이다. 즉 이들 데이터 셋으로 MAE를 학습하고 인코더 부분의 ViT에 MLP 헤드를 부가하여 얼굴 표정 분류기를 구성하였다. 이렇게 대략 구성된 표정 분류기는 부족한 판소리 소리꾼 얼굴표정 데이터로 미세 조정하여 분류기를 완성하였다. 이러한 과정이 자기 지도 학습의 내림 과제(down-stream task)에 해당하며, 그림 6은 이러한 과정을 통해 얻은 소리꾼 얼굴표정 인식기를 보여준다.

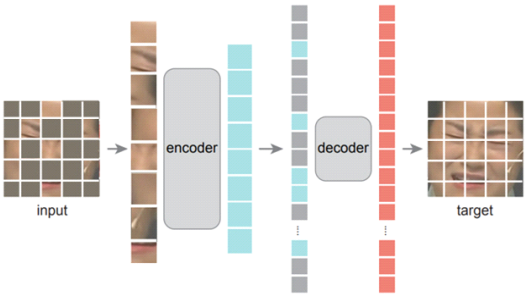

그림 5. MAE의 구조
Fig. 5. Structure of MAE

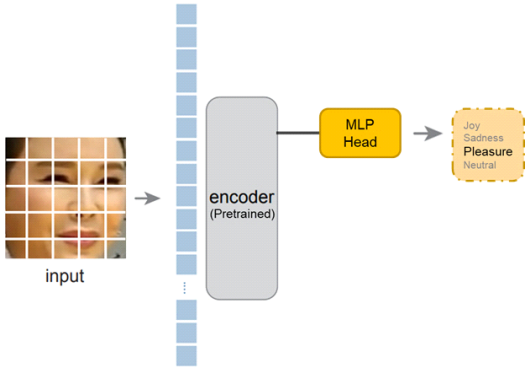

그림 6. 학습된 MAE 인코더를 이용하는 판소리 소리꾼 얼굴 표정 분류기
Fig. 6. Classifier for Facial Expressions of Pansori Singer Using Pretrained MAE Encoder

Ⅳ. 실험 및 결과

4.1 소리꾼 검출 실험 결과

학습된 트랜스포머 DETR를 활용 테스트 데이터의 추론을 진행하여 표 3 같은 결과를 얻을 수 있었다. 객체 검출의 평가 지표로 사용된 IoU(Intersection over Union)의 문턱치를 95%, 75%, 50%로 설정할 때 각각 74%, 91%, 93.5%의 소리꾼 객체의 AP(Average Precision)를 얻을 수 있다. 여기서 주목할 것은 본 실험에서 나온 소리꾼 객체는 큰 사이즈의 객체로 크기 변화가 심하지 않다는 점이며 IoU 75% 이상이면 약 91%의 소리꾼 검출이 가능하고 이어지는 발림 동작 분류 및 소리꾼 얼굴 검출 및 분류에 큰 영향을 주지 않는다는 점이다.

만약 더 많은 객체를 검출하여 발림 동작을 판단하기 위해서는 IoU 문턱치를 줄일 수 있는데 이 경우는 발림 동작 분류 및 소리꾼 얼굴 검출을 수행하기 전에 후 처리를 통해 객체를 확장해 소리꾼 영역을 크게 만들어서 반드시 소리꾼이 발림 동작 분류에 포함되어야 한다. 그러나 매 프레임 발림 동작이나 얼굴 표정의 부류가 변하지는 않기 때문에 후처리 과정 없이도 91%의 프레임에서 소리꾼 객체를 찾으므로 빠진 프레임이 있더라도 분석에 큰 영향을 주지 않을 것으로 예상된다.

표 3. 소리꾼 검출의 실험 결과
Table 3. Singer Detection Results

	AP (IoU=0.95)	AP (IoU=0.75)	AP (IoU=0.50)
소리꾼 검출	0.74	0.91	0.935

4.2 얼굴 표정 분류

본 연구의 얼굴 표정 분류도 그림 1의 파이프라인에서 소리꾼을 검출하고 얼굴을 검출한 뒤 특징을 추출해서 표정을 분류한다. 여기서 얼굴 검출기는 성능도 우수하고 속도도 빠른 ResNet기반 SCRFD 검출기를 사용하였다[19]. 판소리 소리꾼의 얼굴 표정 데이터 셋에 실험 결과는 얼굴 표정에 관한 공식 데이터 셋에서의 결과와도 비교해 검토하였다.

4.2.1 단계적인 자기 지도 학습 실험

우선 ViT 모델을 ImageNet 데이터셋에서 MAE 방법으로 자기 지도 학습을 수행한 뒤 그 결과를 전이 학습에 활용하고 뒤이어 공식 얼굴표정 데이터 셋인 RAF-DB로 학습을 진행하였다. 그림 7은 사전 학습이

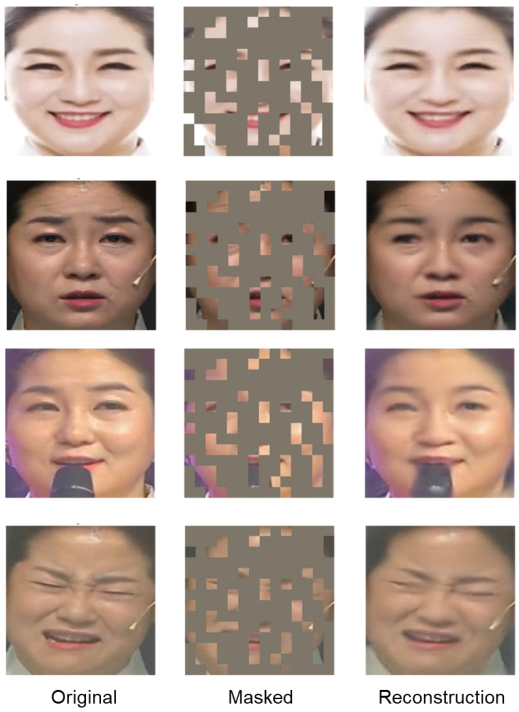

그림 7. MAE 사전 학습 추론의 예
Fig. 7. Example of Inference Results of MAE Pretraining

끝난 MAE 모델을 통한 판소리 소리꾼 얼굴 표정의 복원 결과를 보여주며, 이는 MAE 방법이 사전 학습을 통해 훼손된 입력 얼굴 영상의 패치들을 충실히 복원할 수 있음을 의미하고, 인코더는 얼굴 영상을 구성하는 패치 간의 관계를 충실히 학습하였음을 의미한다.

이렇게 데이터 수가 많고 주석 정보 없는 데이터로 사전 학습 끝난 뒤 수집한 판소리 소리꾼 얼굴 표정 데이터 셋에 인코더 부분의 ViT 가중치를 초기화하고 FC(Fully Connected) 층의 분류기를 붙여서 미세조정으로 교차 검증 학습을 수행하였다.

4.2.2 얼굴 표정 분류 실험 결과

얼굴 표정 분류 실험은 사용된 부류의 데이터 불균형이 심각하지 않고 단순한 단일 레이블 분류 문제이기 때문에 식 (1) 의 정확도(accuracy)를 평가 척도로 사용하였다.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

MAE 사전 학습의 유용성을 검증하기 위해 RAF-DB 데이터, ImageNet 데이터, 선 ImageNet 데이터 후 RAF-DB 데이터로 각각 사전 학습한 뒤 MAE 학습에 사용하지 않은 RAF-DB 데이터 셋으로 미세 조정을 수행하여 학습 성능을 측정 하였다.

그 결과 표 4와 같이 각각 90.19%, 90.45%, 91.46%의 정확도를 얻었으며, 선 ImageNet 후 RAF-DB로 사전 학습하는 것이 제일 높은 정확도를 얻었다. 그 뒤를 이어 ImageNet, RAF-DB로 각각 사전 학습한 경우가 우수하였다. 보편적인 ImageNet 데이터보다 얼굴 표정 데이터인 RAF-DB 데이터가 내림과제 도메인을 더 적합하지만 데이터 크기 차이의 벽을 넘지 못하는 것이 원인이라고 여긴다.

선 ImageNet 후 RAF-DB로 사전 학습을 수행한 ViT 백본을 활용 판소리 데이터셋에 미세 조정하여 얻은 실험결과는 표 5와 같이 평균 정확도가 78.44%에 달하였다.

표 4. 사전 학습 데이터 셋의 비교
Table 4. Comparison of MAE Pretraining with Different Datasets

데이터셋	정확도	사전학습 데이터셋
RAF-DB	0.9019	RAF-DB
RAF-DB	0.9045	Imagenet
RAF-DB	0.9146	Imagenet +RAF-DB

표 5. 판소리 소리꾼 얼굴 표정 실험 결과
Table 5. Experimental Results of Pansori Singer's Facial Expressions

Fold	정확도	사전학습 데이터셋
Fold 0	0.7748	Imagenet+RAF-DB
Fold 1	0.7901	Imagenet+RAF-DB
Fold 2	0.7843	Imagenet +RAF-DB
Average	0.7844	

4.2.3 얼굴 표정 인식 결과의 검토

4.2.3.1 얼굴 표정 인식 성능 비교

표 6에는 본 논문의 제안된 MAE 자기 지도 학습을 이용하는 얼굴 표정 인식의 성능을 가름해 보기 위해 공개된 데이터 셋인 RAF DB에 대한 여러 방식의 모델의 표정 인식 성능을 비교해 보았다. 표에서 알 수 있듯이 본 논문에서 제안하는 방법이 동일 데이터 셋에 대해 상대적으로 가장 우수한 결과를 제공하는 것을 알 수 있다. 이는 MAE 방법으로 ImageNet 데이터 셋이나 RAF-DB 데이터로 ViT 백본을 사전 학습하고 나머지 RAF-DB 데이터로 미세 조정하는 방식

표 6. RAF-DB 데이터 셋에 모델 성능 비교
Table 6. Performance Comparison of Public RAF-DB Dataset

모델	정확도
Separate-Loss[22]	0.8638
DDA-Loss[23]	0.8690
SCN[24]	0.8703
PSR[25]	0.8898
DACL[26]	0.8778
IF-GAN[27]	0.8833
EfficientFace[28]	0.8836
MViT[29]	0.8862
DAN[30]	0.8970
MAE(Pretrained)	0.9146

이 바람직한 방법이었음을 의미한다.

또한 본 연구와 비슷한 영화 등 동영상에서 캡처하는 수집 방법을 이용하는 SFEW(Static Facial Expressions in the Wild)의 SOTA(state-of-art)[31] 정확도가 58.50%임을 고려할 때, 비슷하게 얻은 본 연구의 데이터에 대한 78.4%의 정확도는 어느 정도 유용한 결과로 판단된다. 다만 SFEW 데이터 셋은 에크만의 여섯 가지 기본 표정을 포함하고 있다는 점은 4가지 표정을 포함하는 본 연구와는 다르다.

4.2.3.2 얼굴 표정 분류 혼동 행렬과 ViT 추출 특징의 tSNE를 통한 검토

그림 8는 얼굴 표정 분류의 테스트 데이터에 대한 혼동 행렬을 보여 주고 있다. “슬픔” 표정인 샘플들이 “기쁨”이나 “무표정”으로 오분류되었지만 이 샘플들은 정지영상을 기반으로 판단하기 때문에 미세한 변화로 사람도 오분류할 수 있다. 그림 9은 이들 혼동 행렬의 예를 표현하며, FP(False Positive)와 FN(False Negative)의 예를 보면, 동영상의 전후 맥락을 따지지 않고 다양한 형태의 기하학적인 변환과 조명 변화가 있는 정지영상에서 표정을 분류하는 일이 어려운 작업임을 알 수 있다. 이는 정지 영상을 활용하여 CES 방법으로 감정을 한정된 범주로 분류하는 것의 한계로 판단된다.

또한 그림 8의 혼동 행렬에서는 “기쁨” 과 “즐거움”에 해당하는 표정에서 “기쁨” 이 “즐거움”으로 혼동되는 경향이 나타났다. 이는 “기쁨” 표정과 “즐거움” 표정의 경계가 다른 표정에 비해 상대적 명확하지 않음에 기인하는 것으로 판단된다.

그림 10는 ViT 백본을 주석이 있는 소리꾼 얼굴 데

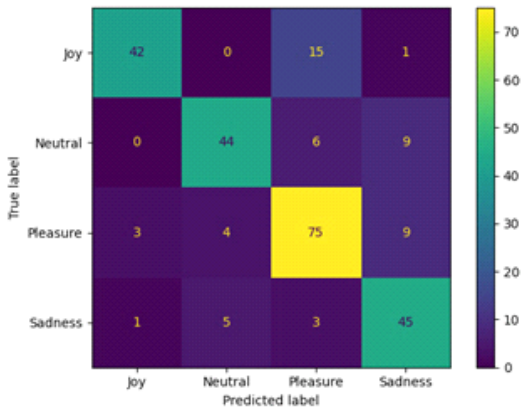

그림 8. 판소리 얼굴 표정 분류 결과의 혼동 행렬
Fig. 8. Confusion Matrix of Classification Results of Pansori Singer’s Facial Expression

이터로 미세조정 한 후의 추출한 특징의 tSNE를 보여주고 있다. 여서도 기쁨” 표정과 “즐거움” 표정의 경계가 다른 표정에 비해 상대적 명확하지 않음을 알 수 있다. 다만 동영상의 연속적인 프레임에서 얼굴 표정을 인식하기 때문에 오인식이 연속적으로 일어나지 않는다면 분류 결과를 스무딩(smoothing)하여 오인식

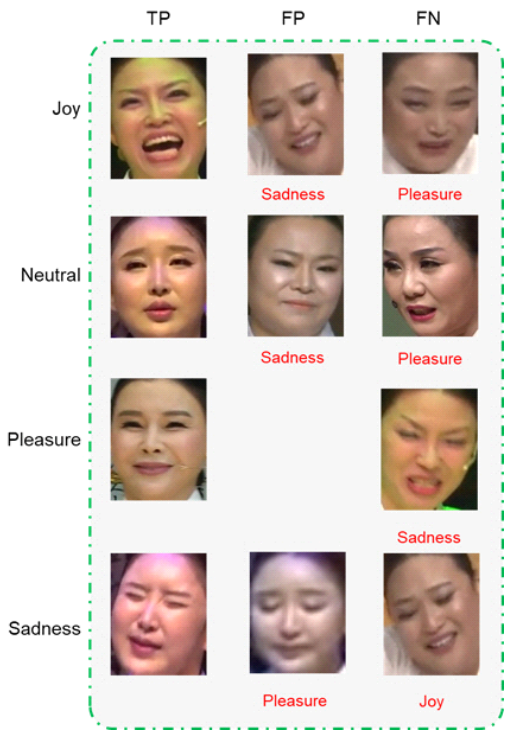

그림 9. 판소리 소리꾼 얼굴 표정 분류 실험결과의 예
Fig. 9. Examples of Classification Results of Pansori Singer’s Facial Expressions

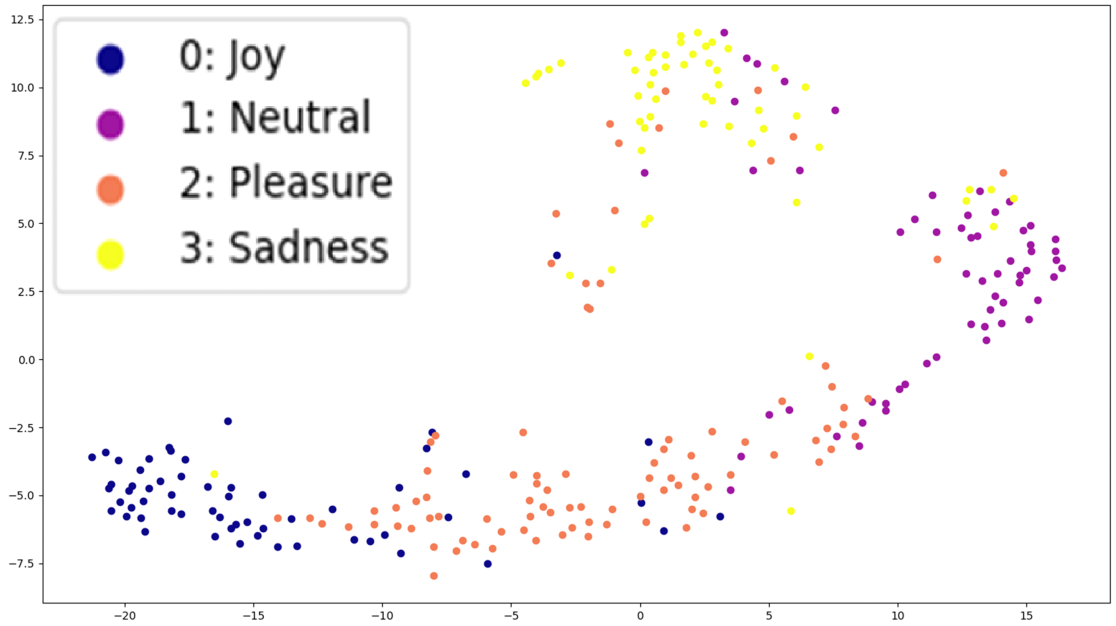

그림 10. 판소리 소리꾼 얼굴 표정의 특징을 TSNE 가시화 결과
Fig. 10. tSNE Visualization for Features of Pansori Singer's Facial Expressions

을 바로잡을 수 있을 것으로 기대된다.

4.3 발림 동작 및 얼굴 표정 분류 인터페이스

본 연구에서는 소리꾼 검출, 발림 동작 분류, 얼굴 표정 분류가 끝난 뒤 동영상 프레임에서 자동으로 소리꾼을 찾고 발림 동작과 얼굴 표정을 분석할 수 있게 그림 11과 같이 GUI 프로그램을 작성하였으며, 길이 2 분 28초이고 프레임 수 총 4,442인 동영상에 적용하여 결과를 출력해 보았다.

그림 11에서와 같이 소리꾼이 있는 영역과 얼굴 영역에 바운딩 박스를 그렸고 구체적인 발림 동작과 표정의 확률을 표현하였다. 통합 실행 결과 소리꾼을 포함하는 비디오 프레임 검출과 범주별 발림 동작 프레임 수는 표 7에 나타냈고 소리꾼 얼굴 표정의 분류 결과는 표 8에 나타내었다.

실험 결과, 비디오 내에 소리꾼과 함께 고수, 청중, 사회자, 스태프 등 인물들이 등장하고 있음에도 불구하고 소리꾼을 찾아서 해당 발림 동작과 표정 범주로 분류하는 능력이 보였다. 비록 표정 분류 성능이 최고조가 아닌 프레임에 성능이 떨어지는 경향을 보였으나, 스무딩 등의 후처리를 거치면 판소리 공연 분석에 도움이 될 것으로 기대된다.

표 7. 통합 실행 결과를 통한 발림 동작 분석
Table 7. An Integrated Analysis Example of Pasori Video for Singer's Motion

항목	프레임 수
총 프레임 수	4,442
소리꾼 존재	3,378
앉음(sitting)	23
부채 펼침(opening the fan)	826
부채를 두 손으로 듦 (holding the fan with two hands)	987
부채 듦(raising the fan)	265
팔 벌림(opening arms)	342
치마 잡기(grabbing the skirt)	657
부채로 지시(pointing with the fan)	655

표 8. 통합 실행 결과를 통한 얼굴 표정 분류 결과
Table 8. An Integrated Analysis Example of Pasori Video for Singer's Facial Expressions

항목	프레임 수
총 프레임 수	4,442
소리꾼 존재	3,378
기쁨(joy)	222
슬픔(sadness)	1,741
즐거움(pleasure)	688
무표정(neutral)	688

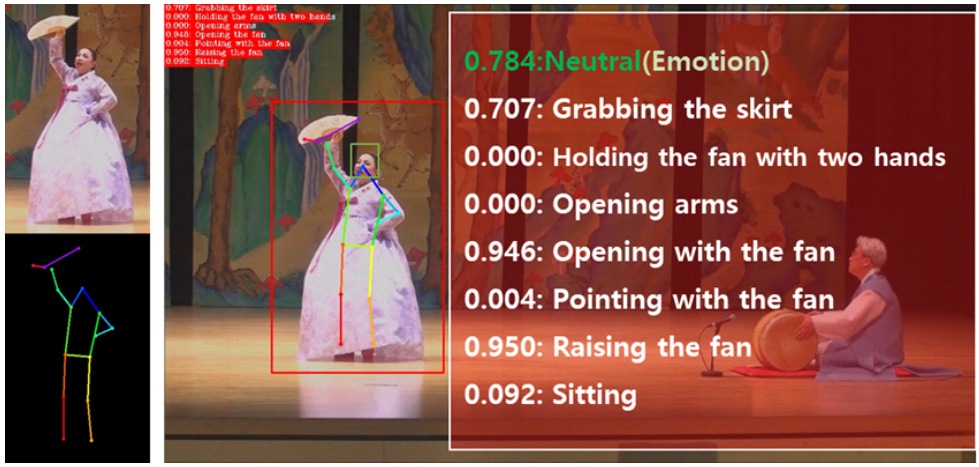

그림 11. 통합 실행의 결과
Fig. 11. Integrated Result of Pansori Singer's Motion and Facial Expression

V. 결 론

판소리 소리꾼은 감정표현 방법의 하나로서 노래 내용에 따라 다른 발림 동작 및 얼굴 표정을 취하여 감정을 표현하고, 이러한 소리꾼의 발림 동작 및 얼굴 표정은 판소리에서 중요한 역할을 한다. 본 논문에서는 동영상의 연속되는 정지된 프레임에서 소리꾼을 찾고, 소리꾼의 얼굴을 찾아내어, 얼굴 표정을 분류하는 방법을 제안하였다.

본 논문에서 소리꾼 검출 방법으로는 트랜스포머 DETR 객체 검출 방법을 활용하였으며, 검출된 소리꾼으로부터 얼굴 검출은 ResNet기반의 SCRFD 검출기를 이용하였다. 판소리 소리꾼 검출에 있어 청중 데이터 부족을 극복하기 위해 MS-COCO 데이터 셋의 "사람" 데이터를 혼합하여 활용하였고, 소리꾼 얼굴 표정 데이터의 제한된 규모를 극복하기 위해 ImageNet 데이터 셋과 RAF-DB 데이터를 활용 MAE 모델에 자기 지도 학습을 수행하여 ViT 백본을 구성하고, 이를 기반으로 소리꾼 얼굴 표정 분류기를 제작하였다. 제안된 방법은 소리꾼 검출은 75% IoU 문턱치에서 약 90%의 정확도를 보였으며, 4개 부류로 분류한 얼굴 표정 인식에서 78.44%의 정확도를 보였다.

제안된 얼굴 표정 인식 방법은 RAF-DB 공개 데이터 셋에 대해 우수한 성능을 보였으며, 비록 판소리 소리꾼의 얼굴 표정 분류의 경우 완벽한 성능은 아니나 매 프레임의 정지 영상을 활용하기 때문에 스무딩 등 후처리 과정을 통해 개선할 여지는 남아있다. 또한 실제 공연 영상에서 얻은 데이터라는 점에서 실용적인 가치도 있다고 판단된다.

이러한 연구 결과는 향후 판소리 뿐만 아니라 다양한전통 문화 콘텐츠 분석 및 보존, 교육 등 분야의 활용을 기대되며, 무용극이나 뮤지컬 등의 다양한 장르의 연구에도 확장할 수 있을 것으로 예상된다.

References

[1] W. Wu, H. Lee, and J.-W. Lee, "Recognition of pansori motion in still images based on keypoint detection," *J. KICS*, vol. 47, no. 4, pp. 575-583, 2022. (https://doi.org/10.7840/kics.2022.47.4.575)

[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," arXiv preprint arXiv:211106377, 2021. (https://doi.org/10.48550/arXiv.2111.06377)

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 35, no. 1, pp. 221-231, 2012. (https://doi.org/10.1109/TPAMI.2012.59)

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. CVPR*, 2014. (https://doi.org/10.1109/CVPR.2014.223)

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition

in videos," *Advances in NIPS*, vol. 27, 2014. (https://doi.org/10.48550/arXiv.1406.2199)

[6] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 40, no. 6, pp. 1510-1517, 2017. (https://doi.org/10.48550/arXiv.1604.04494)

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *Eur. Conf. Comput. Vision,* Springer, 2020. (https://doi.org/10.48550/arXiv.2005.12872)

[8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Comput.*, vol. 27, no. 6, pp. 803-816, 2009. (https://doi.org/10.1016/j.imavis.2008.08.005)

[9] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 29, no. 6, pp. 915-28, 2007. (https://doi.org/10.1109/TPAMI.2007.1110)

[10] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38-52, 2010. (https://doi.org/10.1109/TSMCB.2010.2044788)

[11] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," *2012 IEEE Conf. Comput. Vision and Pattern Recognition,* IEEE, 2012. (https://doi.org/10.1109/CVPR.2012.6247974)

[12] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, 2020. (https://doi.org/10.1109/TAFFC.2020.2981446)

[13] J. A. Russell, "A circumplex model of affect," *J. Personality and Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980. (https://psycnet.apa.org/doi/10.1037/h0077714)

[14] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356-370, 2018. (https://doi.org/10.1109/TIP.2018.2868382)

[15] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," *2011 IEEE ICCV Wkshps.*, IEEE, 2011. (https://doi.org/10.1109/ICCVW.2011.6130508)

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in NIPS,* vol. 28, 2015. (https://doi.org/10.48550/arXiv.1506.01497)

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2016. (https://doi.org/10.48550/arXiv.1506.02640)

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, et al., "Attention is all you need," *Advances in NIPS*, vol. 30, 2017. (https://doi.org/10.48550/arXiv.1706.03762)

[19] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," arXiv preprint arXiv:210504714. 2021. (https://doi.org/10.48550/arXiv.2105.04714)

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:181004 805, 2018. (https://doi.org/10.48550/arXiv.1810.04805)

[21] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," arXiv preprint arXiv:210608254. 2021. (https://doi.org/10.48550/arXiv.2106.08254)

[22] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate loss for basic and compound facial expression recognition in the wild," *Asian Conf. Mach. Learn.*, PMLR, 2019.

[23] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recog. Wkshps.* 2020. (https://doi.org/10.1109/CVPRW50498.2020.00211)

[24] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognition,* 2020. (https://doi.org/10.48550/arXiv.2002.10392)

[25] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access.*, vol. 8, pp. 131988-2001, 2020. (https://doi.org/10.1109/ACCESS.2020.3010018)

[26] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF Winter Conf. Appl. Computer Vision,* 2021. (https://doi.org/10.1109/WACV48630.2021.00245)

[27] J. Cai, Z. Meng, A. S. Khan, J. O'Reilly, Z. Li, S. Han, et al., "Identity-free facial expression recognition using conditional generative adversarial network," *2021 IEEE ICIP*, IEEE, 2021. (https://doi.org/10.48550/arXiv.1903.08051)

[28] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artificial Intell.,* 2021. (https://doi.org/10.1609/aaai.v35i4.16465)

[29] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "MVT: Mask vision transformer for facial expression recognition in the wild," arXiv preprint arXiv:210604520, 2021. (https://doi.org/10.48550/arXiv.2106.04520)

[30] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," arXiv preprint arXiv:210907270, 2021. (https://doi.org/10.48550/arXiv.2109.07270)

[31] F. Wu, C. Pang, and B. Zhang, "FaceCaps for facial expression recognition," *Comput. Animation and Virtual Worlds*, vol. 32, no. 3-4, e2021, 2021. (https://doi.org/10.1002/cav.2021)

오 문 흠 (Wenqin Wu)

2020년 : Zhejiang Gongshang University Hangzhou College of Commerce 공학사

2020년 9월~2022년 8월 : 전북대학교 전자정보공학과(컴퓨터공학) 석사과정

<관심분야> 영상처리, 딥러닝

[ORCID:0000-0002-0322-7142]

이 혜 정 (Hyejeong Lee)

2021년 2월 : 전북대학교 한국음악학과 졸업

2021년 3월~현재 : 전북대학교 한국음악학과 석사과정

<관심분야> 한국음악, 국악, 국악이론, 판소리

[ORCID:0000-0001-7667-3234]

이 준 환 (Joonwhoan Lee)

1980년 2월 : 한양대학교 전자공학과 공학사

1982년 2월 : 한국과학기술원 전자공학과 공학석사

1990년 8월 : 미국 미주리대학 전기 및 컴퓨터공학과 공학박사

1990년 10월~현재 : 전북대학교 컴퓨터공학부 교수

<관심분야> 영상처리, 감성 분석, 인공지능

[ORCID:0000-0003-1854-9643]