The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 24, No. 1, pp.155-161, Feb. 29, 2024. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2024.24.1.155
JIIBC 2024-1-23

기하학적 특징 추가를 통한 얼굴 감정 인식 성능 개선

Improvement of Facial Emotion Recognition Performance through Addition of Geometric Features

정호영*, 한희일**

Hoyoung Jung*, Hee-Il Hahn**

요 약 본 논문에서는 기존의 CNN 기반 얼굴 감정 분석 모델에 랜드마크 정보를 특징 벡터로 추가하여 새로운 모델을 제안한다. CNN 기반 모델을 이용한 얼굴 감정 분류 연구는 다양한 방법으로 연구되고 있으나 인식률이 매우 저조한 편이다. 본 논문에서는 CNN 기반 모델의 성능을 향상시키기 위하여 CNN 모델에 ASM으로 구한 랜드마크 기반 완전 연결 네트워크를 결합함으로써 얼굴 표정 분류 정확도를 향상시키는 알고리즘을 제안한다. CNN 모델에 랜드마크를 포함시킴으로써 인식률이 VGG 0.9%, Inception 0.7% 개선되었으며, 랜드마크에 FACS 기반 액션 유닛 추가를 통하여 보다 VGG 0.5%, Inception 0.1%만큼 향상된 결과를 얻을 수 있음을 실험으로 확인하였다.

Abstract In this paper, we propose a new model by adding landmark information as a feature vector to the existing CNN-based facial emotion classification model. Facial emotion classification research using CNN-based models is being studied in various ways, but the recognition rate is very low. In order to improve the CNN-based models, we propose algorithms that improves facial expression classification accuracy by combining the CNN model with a landmark-based fully connected network obtained by ASM. By including landmarks in the CNN model, the recognition rate was improved by several percent, and experiments confirmed that further improved results could be obtained by adding FACS-based action units to the landmarks.

Key Words : ASM, Facial Expression Recognition, Inception, VGG

Ⅰ. 서 론

사람의 얼굴 감정을 컴퓨터로 인식하는 연구는 컴퓨터 비전과 기계 학습 등의 발전으로 인하여 많은 주목을 받고 있다. 사람들은 얼굴 표정을 통하여 다양한 감정을 표현하고 이러한 정보를 기반으로 상황을 이해하고 상호 작용한다. 따라서 이에 대한 응용 분야는 환자의 표정 분석을 통한 건강 진단 등의 의료 분야, 인간과 컴퓨터의 상호 작용 및 감정 인식 로봇 분야, 마케팅 및 광고 분야, 고객 서비스 자동화 등 다양한 분야에서 활용 가능하다.

얼굴 표정을 통해 사람의 감정을 인식하기 위하여 과거에는 HOG(Histogram of Oriented Gradients)[1], ASM(Active Shape Model)[2], AAM(Active Appearance Model)[3] 등, 전통적인 특징 기반 분류 방법이 널리 사

*학생회원, 한국외국어대학교 정보통신공학과
**정회원, 한국외국어대학교 정보통신공학과(교신저자)
접수일자 2024년 1월 9일, 수정완료 2024년 1월 30일
게재확정일자 2024년 2월 9일

Received: 9 January, 2024 / Revised: 30 January, 2024 / Accepted: 9 February, 2024
*Corresponding Author: hihahn@hufs.ac.kr
Dept. Information and Communications Eng., College of Engineering, Hankuk University of Foreign Studies, Korea

용된다. 이러한 기법들은 주로 지역적인 정보에 초점을 맞추어, 일차원 벡터로 변환하는 과정에서 공간적인 정보를 손실하고 이미지 내에서 화소 간의 상관관계를 고려하지 못하거나, 낮은 수준의 특징을 위주로 추출하여 고수준의 추상적인 패턴을 효과적으로 인식하지 못하는 등의 문제점을 드러낸다. Sarnarawickrame[4]와 장길진[5]은 얼굴 이미지에서 특징 정보인 랜드마크를 추출하기 위해 ASM을 이용하고 서포트 벡터 머신으로 감정을 분류한다. 기존의 얼굴 감정 인식 모델은 모호한 표정, 조명, 얼굴 각도 등 다양한 외부요인들로 인하여 응용할 수 있을 정도의 성능 확보가 쉽지 않다. 또한, 특정 환경과 개체에 대해서만 효과적으로 작동하고 다양한 환경 및 개체에 대한 감정 분석에서 일반화 능력이 저하될 수 있다. 예를 들어, 모델이 특정 인종, 연령, 또는 문화적 배경에 대하여 훈련이 되면, 다른 인종, 연령 및 문화적 배경의 개체에 대한 감정 분석에서 제한 사항이 발생할 수 있다. 따라서 기존의 얼굴 감정 인식 모델은 범용성이 제한되며, 다양한 환경과 다양한 인구에 대한 감정 분석의 정확성이 저하된다는 문제점이 있다. 따라서 모델의 성능을 향상시키고 일반화 능력을 확장하기 위해서는 다양한 데이터와 다양한 조건을 고려하여야 한다. 이러한 문제점을 해결하기 위하여 최근에는 합성곱 신경망(CNN) 등과 같은 딥러닝 기술이 사용되고 있다.

CNN을 활용한 얼굴 감정 분석은 저수준에서 고수준의 추상적인 특징을 추출한다. 또한 합성곱 필터를 통해 주변 화소와의 연관 관계를 유지하며 학습하기에 얼굴 이미지의 공간적 특징을 고려할 수 있다. Wan[6]은 캐글(Kaggle)에서 제공하는 FER2013 데이터셋에 대하여 AlexNet과 VGGNet의 모델을 성능 비교하여 각각 54.8%, 63.1%의 정확도를 보인다. 그러나, 네트워크의 깊이가 깊은 CNN 모델은 특정 데이터셋에 과도하게 최적화되기 쉬워 새로운 환경이나 다른 문화적 배경에서 일반화하기 어려운 한계가 발생할 수 있다. 이에 VGGNet과 유사하게 네트워크를 구성하되 층의 깊이를 11개의 층으로 줄여 모델을 구성함으로써 65.3%의 얼굴 감정 분류 정확도를 보였다. 이와 유사하게 Christopher et al.[7]은 FER2013 과 같이 데이터의 크기가 작고 클래스별 데이터 샘플의 개수가 차이가 많은 데이터셋에 층이 깊은 모델을 사용하는 것은 불필요하다고 제안한다. 이에 입력 데이터에 대해 상하 반전, 랜덤 크롭 등의 데이터 증강 기법을 통해 데이터 불균형 문제를 해결하고, Inception, VGG와 같은 CNN 모델에 기반하여 구조를 단순화시킴으로서 69.6%의 정확도를 보여 준다. Yen et al.[8]은 입력 데이터에 대하여 감정 클래스의 개수와 샘플 개수의 비율을 가중치로 주고 Inception 모델을 학습하여 68.0%의 성능을 보여 준다. 하지만 이미지 샘플의 크기가 작아 특징 정보를 효과적으로 추출하지 못했다는 한계점이 지적된다. 이러한 한계점을 극복하기 위해 조찬영[9]은 얼굴 이미지에 추가적으로 음성 데이터를 추가하는 멀티모달 방식의 감정 분석 시스템을 제안하였한다. 이미지 분석 네트워크와 음성 데이터를 텍스트로 변환 후 네트워크를 구축하고 각 계산된 결과에 대해 우선순위가 높은 감정을 최종감정으로 선택하여 74%의 정확도를 보인다. 이에 본 논문은 작은 데이터셋에서의 특징 정보 부족의 문제를 해결하기 위하여 얼굴 이미지에서 추출한 랜드마크 정보를 추가하는 방법을 제시한다. ASM 알고리즘을 활용하여 기본적인 얼굴 특징점을 추출하고, 이에 더하여 FACS[10](Facial Action Coding System)를 기반으로 의미 있는 특징점을 추가로 선정한다. 최종적으로 특징점을 활용하여 완전 연결 기반의 네트워크를 구성하고, 이미지 데이터를 입력으로 사용하기 위한 CNN을 결합한다. 이를 통하여 감정 분석 모델의 정확도와 다양한 환경에서의 일반화 능력을 향상시킬 수 있는지를 실험으로 확인한다.

본 논문의 구성은 다음과 같다. 논문의 핵심 개념을 이해하기 위한 배경 지식으로 ASM, VGG, Inception을 Ⅱ절에서 설명한다. Ⅲ절에서는 제안 알고리즘을 소개하고, 얼굴 감정 인식에 대한 실험 결과 및 분석을 Ⅳ절에 제시한다. 마지막으로 Ⅴ절에서는 결론을 맺고 향후 연구 진행방향에 대하여 논의한다.

Ⅱ. 제안 모델 설계

1. ASM을 이용한 얼굴 형태 모델 추출

얼굴 형태 모델을 구현하기 앞서 이미지에서 얼굴 영역을 검출해야 한다. 검출하는 방법으로는 모델 기반 알고리즘[11]과 비올라 존스 알고리즘 등이 있다. 본 논문에서는 OpenCV에서 제공하는 Haar cascade 분류기를 사용하여 이미지에서 얼굴 영역을 검출한다. 검출된 영역에서 얼굴 형태를 표현하기 위한 모델을 구현하기 위해서는 각 이미지 내의 얼굴 주위에 특정한 랜드마크를 표시하고 라벨링한 다음, 동일한 라벨의 점들에 대한 통계적 특성을 구함으로써 얼굴형태에 대한 모델을 구할 수 있다. 각 훈련 이미지에서 추출한 랜드마크들 간의 통계정보를 추출하기 위 모델을 구성하기 위하여 다음과

같이 각 이미지의 랜드마크 좌표로 구성된 $N\times 2n$ 행렬을 구한다.

$$X=\begin{bmatrix}\boldsymbol{x_1}\\ \boldsymbol{x_2}\\ \vdots\\ \boldsymbol{x_N}\end{bmatrix}=\begin{bmatrix}x_{11}\ y_{11}\ x_{12}\ y_{12}\cdots x_{1n}\ y_{1n}\\ x_{21}\ y_{21}\ x_{22}\ y_{22}\cdots x_{2n}\ y_{2n}\\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots\\ x_{N1}\ y_{N1}\ x_{N2}\ y_{N2}\cdots x_{Nn}\ y_{Nn}\end{bmatrix} \quad (1)$$

두 개의 형태 벡터 $\boldsymbol{x_i}$와 $\boldsymbol{x_j}$가 주어질 때, $\boldsymbol{x_j}$를 $\boldsymbol{x_i}$의 포즈에 정합하기 위해서 다음과 같은 닮음 변환을 이용하여 형태벡터 $\hat{\boldsymbol{x}}_{\boldsymbol{j}}$를 구하는 방법[2]을 채택한다.

$$\begin{bmatrix}\hat{x}_{jk}\\ \hat{y}_{jk}\\ 1\end{bmatrix}=\begin{bmatrix}s_j\cos\theta_j & -s_j\sin\theta_j & t_x\\ s_j\sin\theta_j & s_j\cos\theta_j & t_y\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_{jk}\\ y_{jk}\\ 1\end{bmatrix} \quad (2)$$

본 논문에서는 N 개의 훈련 이미지에서 구한 형태 벡터를 정렬하기 위해 첫번째 형태 $\boldsymbol{x_1}$을 기준 형태 벡터로 정하고, 나머지 각 형태 벡터($\boldsymbol{x_2},\cdots,\boldsymbol{x_N}$)를 $\boldsymbol{x_1}$과 정렬시킨 다음, 정렬된 형태 벡터로부터 평균 형태 벡터와 공분산 행렬 등의 통계정보를 다음과 같이 각각 구한다.

$$\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{x}}_{\boldsymbol{i}} \quad (3)$$

$$\boldsymbol{S}=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\boldsymbol{x}}_{\boldsymbol{i}}-\bar{\boldsymbol{x}}\right)\left(\hat{\boldsymbol{x}}_{\boldsymbol{i}}-\bar{\boldsymbol{x}}\right)^T \quad (4)$$

여기서 $\hat{\boldsymbol{x}}_i$는 $\boldsymbol{x_i}$를 $\boldsymbol{x_1}$ 중심으로 정렬한 형태 벡터이고 $\hat{\boldsymbol{x}}_1=\boldsymbol{x_1}$이다. 이와 같이 구한 정렬된 형태 벡터와 평균 형태 벡터를 이용하여 PCA(principal component analysis)를 적용하기 위하여 $\boldsymbol{S}$의 고유값(eigenvalue)과 고유벡터(eigenvector)를 다음과 같이 구한다.

$$\boldsymbol{S}\boldsymbol{e_k}=\lambda_k\boldsymbol{e_k},\qquad k=1,\cdots,2n \quad (5)$$

λ_k는 공분산 행렬 $\boldsymbol{S}$의 k 번째 고유값으로서 내림차순으로 정렬되었다고 가정한다. 즉, $\lambda_k\geq\lambda_{k+1}$를 만족한다. 고유값 λ_k는 N 개의 형태 벡터에서 고유벡터 $\boldsymbol{e_k}$ 방향으로 변하는 정도의 분산에 비례한다. 즉, 가장 큰 고유값 λ_1에 해당되는 고유벡터 $\boldsymbol{e_1}$은 N개의 형태 벡터에서 가장 크게 변하는 성분을 나타내고 그 분산은 λ_1에 비례한다. 이러한 수학적 특성을 얼굴 검출에 적용하면, 얼굴에서 어느 성분의 변화가 크고 작은지를 확인할 수 있다. 그림 1은 초기 기준형태 벡터와 30개의 형태 벡터의 분포를 보여준다.

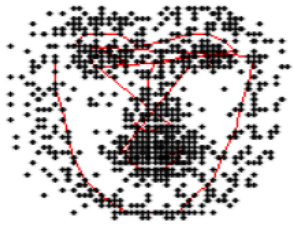


그림 1. 기준 형태벡터와 30개의 형태벡터의 분포
Fig. 1. Initial reference shape vector and distribution of individual 30 shape vectors

실험에 의하면 $2n$개의 고유값 중에서 일부만 큰 값을 갖고 나머지는 매우 작은 값을 갖는다. 따라서, $2n$ 차원의 형태 벡터를 $2n$개의 고유 벡터 대신 그보다 훨씬 적은 수의 벡터로 표현할 수 있기 때문에 자유도를 크게 줄일 수 있어 문제를 단순화시킬 수 있는 효과를 얻을 수 있다. 이를 이용하면, $t\ll 2n$개의 고유벡터의 선형 결합으로 임의의 형태 벡터 $\boldsymbol{e}$를 합성할 수 있다.

$$\boldsymbol{e}=\bar{\boldsymbol{e}}+\boldsymbol{\Phi b} \quad (6)$$

여기서, $\Phi=\{e_0,e_1,...,e_{t-1}\}$는 t개의 고유벡터로 구성된 행렬이고 $\boldsymbol{b}=(b_1,\cdots,b_t)^T$는 고유 벡터의 비례상수로 구성된 열벡터이다. 위 식은 파라미터($b_1\sim b_t$)를 적절히 변화시켜 줌으로써 임의의 형태 벡터를 생성할 수 있음을 보여 준다. 또한, b_k의 분산은 고유값 λ_k와 일치하므로 b_k가 가우시안 분포를 갖는다고 가정하면 대부분이 표준편차의 3배 범위 내에 존재하므로 b_k의 범위를 다음과 같이 한정시키면,

$$-3\sqrt{\lambda_k}\leq b_k\leq 3\sqrt{\lambda_k} \quad (7)$$

대부분의 얼굴 형태에 대한 변화량을 감당할 수 있다. 그림 2는 ASM을 이용하여 검출한 형태 벡터의 예를 보여준다.

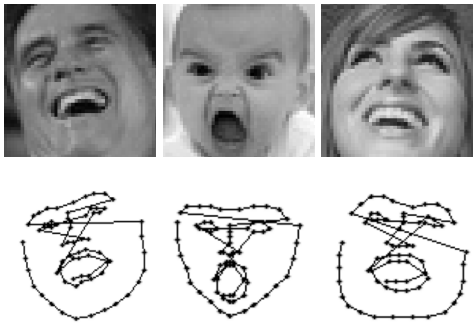


그림 2. ASM을 이용한 형태 벡터 검출
Fig. 2. Examples of detection of shape vectors

2. VGG 네트워크

VGG, Inception 등과 같은 CNN 기반 모델을 이용하여 얼굴 감정을 분류하는 연구가 활발히 진행되고 있다. VGGNet[12]는 모든 합성곱 계층에서 3×3 합성곱 필터를 사용하여 AlexNet이 가진 8개의 층보다 2배 깊은 신경망을 구축하였음에도 Top-5 에러를 절반 이상 줄임으로써 성공적인 학습 능력을 보여 준다. 7×7 합성곱 필터는 3×3 합성곱 필터 3개로 추상화하여 같은 결과 특성맵의 크기로 추상화할 수 있으며, 연산량이 절반 이상 감소하는 효과가 있다. 또한 하나의 필터를 여러 개의 필터로 분리하는 과정에서 모델 분류의 비선형성이 증가되어 이미지 분류에서 향상된 성능을 보이는 것으로 알려져 있다. VGG 네트워크의 대표적인 모델인 VGG16는 13개의 합성곱 층과 3개의 완전 연결 층으로 이루어진다. VGG 네트워크는 합성곱 연산에서 특징 맵의 크기가 변함 없고 풀링층에서만 특징 맵의 크기가 조절되는 단순한 네트워크라는 점에서 이미지 분류 모델에서 많이 사용되는 CNN 모델 중 하나이다.

3. InceptionV3 네트워크

일반적으로, CNN은 성능을 향상시키기 위하여 모델의 깊이와 너비를 동시에 증가시키는데, 이로 인하여 연산량이 크게 증가하여 모바일 환경이나 제한된 메모리에서 활용하여야 하는 등의 상황에서 큰 단점으로 작용한다. 이를 해결하기 위하여 Inception 모델은 더 작은 합성곱으로의 분해. 보조 분류기 활용, 그리드 크기 축소, 레이블 스무딩을 이용한 모델 정규화 등의 기법을 적용한다. 이 모델은 N×N 합성곱 필터를 N×1 필터와 1×N 필터로 대체하고, 대부분의 합성곱 필터로 3×3 필터를 사용한다. 또한 1×1 합성곱 필터를 통해 차원의 수를 줄여 연산량을 감소시키고, 이러한 필터를 여러 번 통과시킴으로써 모델의 비선형도를 증가시켜 복잡한 패턴에 강인해지는 효과를 얻는다. 따라서 InceptionV3[13,14]는 VGGNet과 유사한 정도의 연산량으로 42 계층의 깊은 신경망으로 구현이 가능하여 특징 정보를 효과적으로 추출해낼 수 있는 동시에 학습해야 할 파라미터의 수는 VGGNet과 유사하다는 점에서 이미지 분류 모델에서 많이 사용된다.

Ⅲ. 제안 알고리즘

본 논문에서는 CNN 모델에 랜드마크 기반 완전연결 네트워크를 결합함으로써 얼굴 표정 분류 정확도를 향상시키는 알고리즘을 제안한다. CNN의 기본 아키텍처로 VGG16과 InceptionV3를 채택하고, 7가지 감정을 분류하는데 용이하도록 각 네트워크를 학습시킨다. 그림 3은 제안 알고리즘의 구조를 보여 준다.

이 그림의 하단에 있는 CNN 블럭은 VGG16 또는 InceptionV3을 사용하는데 이들의 출력 특징 맵의 크기는 각각 512, 2048개이다. 상단의 MLP 블럭은 얼굴 이미지에서 ASM 알고리즘을 통해 추출된 랜드마크와 FACS 기법을 참고하여 특징점을 추가한 벡터(90×2)가 입력으로 사용된다. 사용된 CNN 모델에 따라 512 또는 2048개의 특징맵을 출력시킨다.

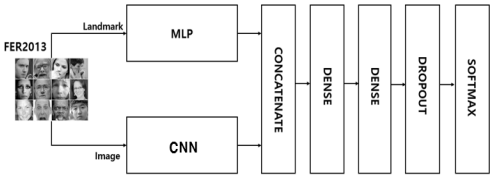


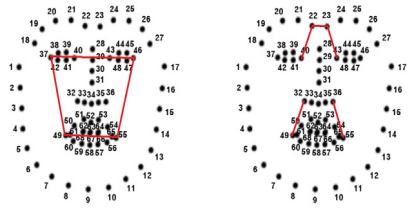
그림 3. 제안 알고리즘 구조
Fig. 3. Block diagram of our proposed algorithm

본 논문에서는 이미지 랜드마크 정보 뿐만 아니라 Ekman[10]이 제안한 FACS 방법론을 기반으로 하여 랜드마크에 유의미한 정보를 추가한다. FACS는 얼굴 표정을 체계적으로 분석하고 이해하기 위한 방법론으로 얼굴 표정을 설명하는 여러 액션 유닛(AU)을 정의한다. 액션 유닛은 특정 얼굴 근육 또는 근육 그룹의 활성화 또는 억제로 인해 발생하는 얼굴 표정의 세분화된 단위이다.

액션 유닛은 1부터 44까지의 번호로 분류되며, 액션 유닛에 대응되는 기본 감정[15]을 기반으로 선정한 랜드마크 간 거리 정보는 표 1과 같다. 행복함과 화남에 대한 추가 특징점의 정보를 시각적으로 나타내면 그림 4와 같다.

표 1. 액션유닛에 대응되는 랜드마크
Table 1. Landmarks corresponding to the action units

감정	액션 유닛 조합 (랜드마크 시작점-끝점)
행복함	6(37-46), 12(49-55), 37-49, 46,55
화남	4(22-23), 5(39-41, 44-48), 7(40-43), 23(9-58)
역겨움	9(32-40, 36-43), 15(32-49, 36-55), 16(9-49, 9-55)
두려움	1(22-40, 23-43), 2(18-37, 27-46), 4(22-23), 5(39-41, 44-48), 20(58-67), 26(34-52, 9-58)
중립	-
슬픔	1(22-40, 23-43), 4(22-23), 15(32-49, 36-55)
놀람	1(22-40, 23-43), 2(18-37, 27-46), 26(34-52, 9-58)



(a) (b)

그림 4. 액션 유닛의 조합에 대응되는 랜드마크 간 거리정보 선정의 예 (a) 행복함 (b) 슬픔
Fig. 4. Examples of determination of distances between landmarks corresponding to combination of the action units. (a) Happiness (b) Sadness

예를 들어, 표 1의 '행복함'에서 액션 유닛 6(37-46)과 12(49-55)가 지정된다. 액션 유닛 6은 랜드마크 37과 46 간의 조합으로, 액션 유닛 12는 랜드마크 49와 55간의 조합으로 각각 나타내고 이들 간의 거리정보가 그림 4(a)에 붉은 선으로 표시된다. 이와 같이 기존 68개의 랜드마크에 FACS 기반 액션 유닛 22가 추가되어 총 90개의 좌표를 특징 정보로 이용한다.

Ⅳ. 실험 결과 및 분석

본 논문에서는 Kaggle이 제공한 공개 데이터셋인 FER2013을 이용한다. 이 데이터셋은 모두 48x48 크기의 그레이 스케일 이미지로 표 2와 같이 훈련 데이터와 테스트 데이터로 구성되어 있다. 얼굴의 크기가 작거나 해상도가 낮은 이미지, 그리고 얼굴이 부분적으로 잘려 얼굴 형태를 명확하게 구분하기 어려운 이미지는 랜드마크 인식 과정에 어려움을 겪을 수 있으므로, 이러한 이미지들은 임의로 제거함으로써 모델의 훈련 데이터의 품질을 향상시키고자 한다.

표 2. FER2013 전처리 후 클래스 별 데이터 분포
Table 2. Number of data belonging to each class after preprocessing of FER2013

	화남	역겨움	두려움	행복함	중립	슬픔	놀람
훈련	2757	340	2537	5620	3731	2664	2358
테스트	642	86	646	1381	931	665	611

FER2013 데이터셋에서 발견된 주요 문제 중 하나는 역겨움 클래스에 대한 훈련 데이터가 340개로, 다른 클래스 중 하나인 행복함의 5620개의 훈련 데이터에 비해 현저히 적다는 것이다. 이러한 데이터 불균형 문제는 모델의 학습 과정에서 특정 클래스에 과적합되는 경향을 보일 수 있다. 이로 인해 전체적인 정확도는 향상시킬 수 있지만, 역겨움 클래스와 같이 데이터가 부족한 클래스에 대한 예측 성능이 저하될 수 있다. 위 문제를 해결하기 위하여 데이터 증강 기법을 도입한다. 좌우 반전, 회전, 확대, 축소와 같은 데이터 증강 기법을 적용하여 훈련 이미지를 다양한 방식으로 변형한다. CNN 기반 네트워크인 VGG, InceptionV3는 각각 512, 2048의 특징 맵의 크기를 갖는다. 이와 결합하기 위하여 MLP 네트워크의 출력 특징 맵의 크기를 결정하여야 한다. 본 논문에서는 실험을 통하여 MLP의 적절한 특징 맵 크기를 결정하는데, 1024개의 출력 특징 맵에서 가장 높은 성능을 보여 준다. 예를 들어, VGG와 결합한 경우에 MLP의 출력 노드의 수가 128일 때에는 64.8%의 정확도를 보이나 1024개의 출력 노드에서는 65.7%로 성능이 개선된다. 마찬가지로 InceptionV3에 대하여 동일한 실험을 수행할 때 65.2%에서 65.7%로 향상된 정확도를 확인할 수 있다.

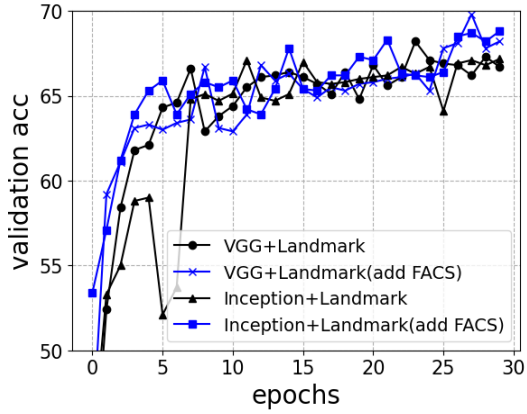


그림 5. 랜드마크 정보에 액션 유닛의 추가로 인한 성능 개선 효과
Fig. 5. Improvement of performance when the action units are added to the feature vectors of landmark.

68개의 랜드마크 정보는 이미지의 정보를 효과적으로 표현하기에 충분하지 않다. 이에 FACS 방법론에 기반하여 랜드마크 간 거리 정보를 추가할 경우 이에 대한 성능 개선 효과를 그림 5에 제시한다.

CNN 네트워크로 VGG를 사용한 경우, 검증 정확도 67.7%에서 FACS에 따라 거리 정보를 추가함으로써 69.4% 검증 정확도로 1.7% 상승하였다. 또한 Inception 네트워크를 사용한 경우 66.3%에서 67.1%로 0.8% 상승함을 확인할 수 있다. 정리하면, VGG나 Inception 등의 CNN을 이용한 알고리즘보다는 랜드마크를 추가함으로써 성능 개선효과를 확인하였다. 본 논문에서 제안한 알고리즘의 성능은 표 3에 제시한다. 실험 결과, 랜드마크를 추가한 모델이 상대적으로 데이터 샘플이 적은 클래스에 대해서도 높은 인식률을 보인다. 이는 랜드마크가 감정 분류에 중요한 정보를 제공하며, 특히 데이터가 부족한 클래스에 대한 예측 성능을 향상시키는 데 중요한 역할을 하는 것으로 분석된다.

표 3. 네트워크 구성에 따른 성능 평가
Table 3. Evaluation of performances corresponding to the structures of the networks

네트워크 구성	정확도
VGG	67.75%
Inception	66.32%
VGG+MLP(제안)	69.42%
Inception+MLP(제안)	67.11%

V. 결 론

본 논문에서는 기존의 CNN 기반 얼굴 감정 분석 모델에 랜드마크 정보를 통합하여 새로운 모델을 제안하였다. 눈, 코, 입 등 얼굴의 중요한 랜드마크와 주변 특징을 정교하게 추출하고 이미지 정보와 결합하여 얼굴 감정 분석에 유용한 특징 정보를 보강하였다. 이 추가 정보는 얼굴 표현을 보다 강화시켜 모델의 성능을 향상시킨 것으로 분석된다. 향후에는 랜드마크 검출 의 정확성에 따른 성능 변화를 분석하고, 다양한 데이터와 다중 모달 데이터를 활용하여 일반화 성능을 향상시킴과 동시에 다양한 연령, 인종의 데이터에 대한 감정 분류 성능을 향상시키는 연구를 수행할 계획이다.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, pp. 886-893, 2005.
DOI: https://doi.org/10.1109/CVPR.2005.177

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham "Active shape models - their training and application". Computer Vision and Image Understanding, Vol. 61, Issue 1, pp. 38-59, 1995.
DOI: https://doi.org/10.1006/cviu.1995.1004

[3] T. F. Cootes, G. J. Edwards, C. J. Taylor, "Active Appearance Model," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No.6, pp.681-685, 2001.
DOI: https://doi.org/10.1109/34.927467

[4] K. Sarnarawickrame and S. Mindya, "Facial expression recognition using active shape models and support vector machines", International Conference on Advances in ICT for Emerging Regions(ICTer), pp. 51-55, 2013.
DOI: https://doi.org/10.1109/ICTer.2013.6761154

[5] Jang, Gil-Jin, Jo, Ahra, Park, Jeong-Sik,, Seo, Yong-Ho, "Video-based Facial Emotion Recognition using Active Shape Models and Statistical Pattern Recognizers", The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 14, No. 3, pp 139-146, 2014.
DOI: https://doi.org/10.7236/JIIBC.2014.14.3.139.

[6] W. Wan, C. Yang, Y. Li, "Facial Expression Recognition Using Convlutional Neural Network :A Case Study of The Relationship Between Dataset Characteristics and Network Performance", Report 20, Stanford Vision Lab

[7] C. Pramerdorfer, M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art", arXiv:1612.02903, 2016.
DOI: https://doi.org/10.48550/arXiv.1612.02903

[8] C. -T. Yen and K. -H. Li, "Discussions of Different Deep Transfer Learning Models for Emotion Recognitions," in IEEE Access, Vol. 10, pp. 102860-102875, 2022.
DOI: https://doi.org/10.1109/ACCESS.2022.3209813

[9] Chanyoung Jo, Hyunjun Jung, "Multimodal Emotion Recognition System using Face Images and Multidimensional Emotion-based Text", The Journal of Korean Institute of Information Technology, Vol.21 No.5, pp. 39-47, 2023.
DOI: https://doi.org/10.14801/jkiit.2023.21.5.39

[10] Ekman, P., & Friesen, W. V., "Facial Action Coding System (FACS) [Database record]", APA PsycTests, 1978.
DOI: https://doi.org/10.1037/t27734-000

[11] Seok-Woo Jang, "Robust Facial Area Aquisition Based on Decision Tree", Journal of the Korea Academia-Industrial cooperation Society, Vol.22, No.7, pp. 183-189, 2021
DOI: https://doi.org/10.5762/KAIS.2021.22.7.183

[12] Karen Simonyan, Andrew Zisserman “Very Deep Convolution Networks for Large Scale Image Recognition”, arXiv:1409.1556, 2014.
DOI: https://doi.org/10.48550/arXiv.1409.1556

[13] C. Szegedy et al., "Going deeper with convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, 2015.
DOI: https://doi.org/10.1109/CVPR.2015.7298594.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.
DOI: https://doi.org/10.1109/CVPR.2016.308.

[15] Ruicong Zhi, Mengyi Liu and Dezheng Zhang. "A comprehensive survey on automatic facial action unit analysis" The Visual Computer: International Journal of Computer Graphics, Vol. 36, Issue. 5, pp. 1067-1093, 2020,
DOI: https://doi.org/10.1007/s00371-019-01707-5

저 자 소 개

정 호 영(학생회원)

- 2024년 2월 : 한국외국어대학교 공과대학 정보통신공학과 학사과정
- 주관심분야 : 정보통신, 머신러닝

한 희 일(정회원)

- 1984년 : 서울대학교 제어계측공학과 학사 졸업.
- 1988년 : 서울대학교 제어계측공학과 석사 졸업.
- 1995년 : University of Arizona 전기및컴퓨터공학과 박사 졸업.
- 2023년 ~ 현재 : 한국외국어대학교 공과대학 정보통신공학과 교수.
- 주관심분야 : 신호처리, 컴퓨터비전, 머신러닝, 게이지 이론

※ 본 논문은 2023년도 한국외국어대학교 교내 학술연구지원에 의하여 연구되었음