특집논문 (Special Paper)
방송공학회논문지 제22권 제2호, 2017년 3월 (JBE Vol. 22, No. 2, March 2017)
https://doi.org/10.5909/JBE.2017.22.2.162
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

깊은 Convolutional Neural Network를 이용한 얼굴표정 분류 기법

최 인 규[a], 송 혁[b], 이 상 용[a], 유 지 상[a)‡],

Facial Expression Classification Using Deep Convolutional Neural Network

In-kyu Choi[a], Hyok Song[b], Sangyong Lee[a], and Jisang Yoo[a)‡]

요 약

본 논문에서는 딥러닝 기술 중의 하나인 CNN(Convolutional Neural Network)을 이용한 얼굴 표정 인식 기법을 제안한다. 기존의 얼굴 표정 데이터베이스의 단점을 보완하고자 질 좋은 다양한 데이터베이스를 이용한다. 제안한 기법에서는 '무표정', '행복', '슬픔', '화남', '놀람', 그리고 '역겨움' 등의 여섯 가지 얼굴 표정 data-set을 구축한다. 효율적인 학습 및 분류 성능을 향상시키기 위해서 전처리 및 데이터 증대 기법(data augmentation)도 적용한다. 기존의 CNN 구조에서 convolutional layer의 특징지도의 수와 fully-connected layer의 node의 수를 조정하면서 여섯 가지 얼굴 표정의 특징을 가장 잘 표현하는 최적의 CNN 구조를 찾는다. 실험 결과 제안하는 구조가 다른 모델에 비해 CNN 구조를 통과하는 시간이 가장 적게 걸리면서도 96.88%의 가장 높은 분류 성능을 보이는 것을 확인하였다.

Abstract

In this paper, we propose facial expression recognition using CNN (Convolutional Neural Network), one of the deep learning technologies. To overcome the disadvantages of existing facial expression databases, various databases are used. In the proposed technique, we construct six facial expression data sets such as 'expressionless', 'happiness', 'sadness', 'angry', 'surprise', and 'disgust'. Pre-processing and data augmentation techniques are also applied to improve efficient learning and classification performance. In the existing CNN structure, the optimal CNN structure that best expresses the features of six facial expressions is found by adjusting the number of feature maps of the convolutional layer and the number of fully-connected layer nodes. Experimental results show that the proposed scheme achieves the highest classification performance of 96.88% while it takes the least time to pass through the CNN structure compared to other models.

Keyword : Convolutional neural network, face expression, data augmentation, data-set

a) 광운대학교 전자공학과(Department of Electrical Engineering, KwangWoon University)
b) 한국전자부품연구원(Department of Electronic Engineering)
‡ Corresponding Author : 유지상(Jisang Yoo)
E-mail: jsyoo@kw.ac.kr
Tel: +82-02)940-5112
ORCID: http://orcid.org/0000-0002-3766-9854
· Manuscript received January 10, 2017; Revised February 28, 2017; Accepted March 20, 2017.

Ⅰ. 서 론

컴퓨터는 인간의 일상생활에 중요한 일부분이 되었을 뿐 아니라,다양한 형태로 편리성을 제공하고 있다. 앞으로도 컴퓨터와 인간과의 밀접성 및 상호작용은 계속해서 증가할 것으로 예상된다. 이에 따라 인간과 컴퓨터와의 상호 작용(Human-Computer Interaction, HCI)에 대한 연구가 인간 공학, 산업 공학, 심리학, 컴퓨터 과학 등 여러 학문 분야에서 진행되고 있다. 인간과 컴퓨터 간의 자연스러운 상호 작용을 위해서 컴퓨터는 사용자의 의도를 종합적으로 판단하고 그에 맞는 반응을 해야 한다. 감정은 인간의 마음 상태를 표출하는 가장 중요한 요소로 사용자의 만족을 극대화하기 위해서는 사용자의 감정 인식이 중요하다. 감정의 형태를 나타내는 중요한 수단 중 하나가 얼굴 표정이고 따라서 얼굴 표정을 분류하는 기술도 반드시 필요하다.

최근에 하드웨어의 발전과 빅데이터의 확보로 빅데이터 안에서 데이터를 기반으로 스스로 학습하고 패턴을 찾아 사물을 구별하는 딥러닝(deep learning) 기술이 주목받고 있다. 복잡한 문제에 대해서 성능이 급격하게 저하되는 기존의 기계학습 모델과는 달리 딥러닝은 깊은 신경망(deep neural networks) 모델을 이용하여 주어진 데이터에 알맞은 고수준의 특징을 추출함으로써 기존의 기계학습의 기술적 한계를 극복할 수 있다. 그 중에서도 인간의 시각 처리 과정을 모방하기 위해 개발된CNN(convolutional neural networks)은 영상 인식 분야에 다양하게 적용되어 높은 성능을 보이고 있다.

매년 개최되는 ILSVRC (ImageNet Large Scale Visual Recognition Competition)에서 2012년도 이후로 성적이 좋은 상위팀은 대부분 CNN 기반 기법을 이용했고, 2015년도 대회에서 우승한 Microsoft Research의 ResNet은 1000개의 부류로 분류하는 문제에 대하여 top5 오차율을 3.54%까지 낮추었다[1]. 또한 페이스북의 CNN 기반의 사진 얼굴 인식 알고리즘인 '딥페이스'의 정확도는 97.25%로 인간 눈의 평균 정확도(97.53%)에 가까운 수치를 보여준다. 기본적인 CNN의 구조는 convolutional layer와 fully-connected layer로 이루어진다. 복수의 convolutional layer를 차례대로 거치면서 특징을 추출하고 추상화하며 점차 고수준의 특징을 추출한다. Fully-connected layer에서 추출한 고수준의 특징으로부터 최종 분류 결과를 결정한다.

기존의 CNN 기반의 표정인식은 확보한 data-set에 맞춰 ILSVRC에서 검증된 모델을 조정하여 사용하거나[2] 얼굴 영역에서 획득하는 특징벡터와 얼굴의 landmark 정보를 결합하여 표정을 분류하는 기법을 이용한다[3]. 전자의 경우 data-set이 얼마나 잘 정제되어있는지에 영향을 받고 객체 분류에서 검증받은 모델이 표정분류에는 성능이 좋지 못한 경우가 있다. 그리고 후자의 경우 얼굴의 landmark를 추정하는 기법이 필요하기 때문에 복잡도가 증가하게 된다. 그리고 CNN 구조를 학습시키려면 많은 데이터가 필요하게 되는데 이러한 문제점을 보완하고자 원본 영상을 인위적으로 변환, 회전, 왜곡하여 데이터의 수를 증가시키는 방법과[4] 적은 데이터를 이용하여 학습하기 위하여 CNN과 Convolutional Autoencoder(CAE)의 두 채널을 이용하여 표정을 분류하는 연구가 진행되었다[5]. 하지만 둘 다 한정된 data-set(CK+, JAFFE)에서 학습과 테스트 데이터를 나누고 실험을 진행했기 때문에 객관적인 결과라고 볼 수 없다.

표정 인식을 위해 CNN 구조를 학습시킬 표정 별로 잘 분리된 다량의 학습데이터가 필요하다. 표정 인식을 위해 많이 쓰이는 Kaggle의 FER2013 data-set의 경우 해상도가 매우 낮고 워터마킹이 삽입된 영상과 다른 표정으로 분류된 영상도 포함되어 있다. 본 논문에서는 10k US Adult Faces Database[6], Indian Movie Face database(IMFDB)[7], Cohn-Kanade AU-Coded Facial Expression(CK+)[8], Chicago Face Database[9], ESRC 3D Face Database[10], Amsterdam Dynamic Facial Expression Set(ADFES)[11], Karolinska Directed Emotional Faces(KDEF)[12], EU-Emotion Stimulus Set[13], Warsaw Set of Emotional Facial Expression Pictures(WSEFEP)[14]의 9개의 data-set을 통합하여 각 표정으로 잘 분류된 고해상도의 data-set을 구성한다. Data-set은 '무표정', '행복', '슬픔', '화남', '놀람', 그리고 '역겨움' 등 여섯 가지 표정으로 구성된다.

본 논문에서는 그림 1의 AlexNet의 기본 구조를 활용한다[15]. 수집한 데이터를 짧은 시간 내에 효율적으로 학습하고 인식하여 높은 정확도의 표정분류를 위해 convolutional layer에서는 특징 지도의 채널 수, fully-connected layer에서는 노드의 수를 조정하여 최적의 구조로 설계하고자 한다.

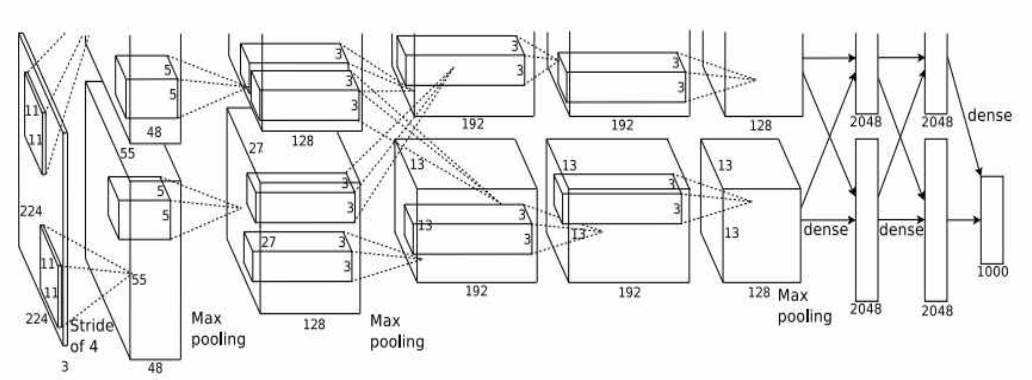

그림 1. AlexNet 구조[13]
Fig. 1. AlexNet structure[13]

본 논문은 다음과 같이 구성된다. 2장에서는 여섯 가지 표정의 data-set을 수집하고, 분류 성능 향상을 위한 데이터 전처리 및 증대 기법에 대해서 소개한다. 3장에서는 수행시간 단축 및 분류 성능 향상을 위한 최적의 학습 구조를 설계하는 과정을 설명하고, 4장에서는 인식 수행시간과 분류 성능 등을 다른 학습 모델들과 비교한다. 마지막으로 5장에서 문제점과 개선사항을 거론하면서 결론을 맺는다.

Ⅱ. Data-set 구성

1. 다양한 DB 수집

먼저 높은 정확도의 얼굴 표정 인식을 위해서는 CNN 구조에 학습시킬 다수의 얼굴영상 data-set이 필요하다. Data-set의 영상들은 표정 인식을 위하여 다양한 감정을 표현하고 있는 얼굴 영상으로 구성되어야 한다. 2013년도에 kaggle에서 개최한 'Facial Expression Recognition Challenge'에서 사용된 데이터베이스(FER2013 data-set)를 보면 37,000여 개의 7가지 표정의 얼굴 영상으로 구성되어 있다[16]. 하지만 영상의 해상도가 48x48로 매우 낮고, 잘 못 레이블된 영상이 포함되어 있다. 낮은 해상도에 맞춰진 CNN 구조가 설계되면 학습 시 고해상도의 영상을 적용할 때는 구조에 맞게 강제적으로 영상 크기가 조절되어야 한다. 이 과정에서 영상의 비율이 변경되고 블러(blur) 현상이 발생하여 분류 성능을 저하시킨다. 또한 잘못된 레이블링은 학습과정에서 오류를 초래할 수 있다.

그림 2는 FER2013 data-set 중의 일부 영상이며 다른 표정으로 잘못 레이블링된 것을 찾을 수 있다. 또한 불필요한 워터마킹이 삽입된 영상도 포함되어 있는 것을 볼 수 있다. 이러한 문제점을 보완하기 위해서 FER2013 data-set을 사용하지 않고 각 대학 및 연구소에서 공개한 여러 data-set 중에서 감정을 잘 표현하는 얼굴 영상을 포함하는 아래의 9가지 data-set을 통합한다. 그리고 다른 표정으로 분류되어 있는 영상이나 구분하기에 애매한 얼굴표정 영상들을 재배치하기 위해서 과반수의 연구실 구성원들이 결정하는 표정

그림 2. FER2013 data-set에서 잘못된 표정으로 분류된 영상의 예
Fig. 2. Examples of images classified as incorrect faces in FER 2013 dataset

으로 분류하여 data-set을 정제하였다.

① 10k US Adult Faces Database: 2,222명을 대상으로 한 10,168장의 자연스러운 얼굴 영상을 포함한다. 대부분 '무표정'과 '행복' 두 가지 표정 영상으로 구성된다.

② Indian Movie Face database: 인도영화에 나오는 100명의 배우들의 34,512장의 얼굴 영상을 포함한다. '무표정', '행복', '슬픔', '화남', '놀람', '두려움', '역겨움' 등 7가지 표정의 얼굴 영상으로 구성된다.

③ Cohn-Kanade AU-Coded Facial Expression: 18세에서 30세 사이의 123명을 대상으로 593개의 비디오 시퀀스를 포함한다. 그 중 309개의 시퀀스에서 '행복', '슬픔', '화남', '놀람', '두려움', '역겨움', '경멸'의 감정을 표현하는 프레임을 찾아서 이용한다.

④ Chicago Face Database: 17-65세 사이의 597명을 대상으로 무표정의 얼굴 영상을 포함한다. 다양한 인종으로 구성되어 있으며 597명 중 158명에 대해서는 '행복', '화남', '두려움' 등의 표정 영상이 포함된다.

⑤ ESRC 3D Face Database: 45명의 남성과 54명의 여성에 대하여 카메라 4대를 이용하여 다양한 각도와 조명하에서 촬영된 영상을 포함한다. '행복', '슬픔', '화남', '놀람', '역겨움'의 표정 등으로 구성된다.

⑥ Amsterdam Dynamic Facial Expression Set: 북유럽과 지중해의 10명의 여성과 12명의 남성의 '행복', '슬픔', '화남', '놀람', '두려움', '역겨움', '경멸', '자신감', '당황스러움' 표정의 얼굴 영상으로 구성된다.

⑦ Karolinska Directed Emotional Faces: 20세에서 30세 사이의 35명의 여성과 35명의 남성에 대하여 -90, -40, 0, +45, +90의 다섯 각도에서 촬영된 4,900장의 영상을 포함한다. '무표정', '행복', '슬픔', '화남', '놀람', '두려움', '역겨움'의 7 가지 표정으로 이루어져 있다.

⑧ EU-Emotion Stimulus Set: 10-70세 사이의 19명의 배우들을 대상으로 '무표정', '행복', '슬픔', '화남', '놀람', '역겨움'의 표정 등으로 구성된다. 배우들은 영국에 있는 드라마 학교나 전문적인 연기 에이전시에서 채용되었다.

⑨ Warsaw Set of Emotional Facial Expression Pictures: 표정연기를 잘하는 30명의 배우들의 '무표정', '행복', '슬픔', '화남', '놀람', '두려움', '역겨움'에 해당하는 표정 영상으로 이루어져 있다.

통합된 data-set은 남녀노소 구분 없이 백인, 흑인, 황인을 대상으로 '무표정, 행복함, 슬픔, 화남, 놀람, 역겨움' 등의 여섯 가지 표정으로 구성한다. 표정의 종류는 각각의 data-set에 포함된 표정으로 정하고 표정의 개수가 너무 적은 경우('무서움')는 제외하였다. 표 1은 재구성한 data-set의 표정 별 영상의 수를 보여준다.

표 1. 수집한 data-set의 표정 별 영상의 개수
Table 1. Number of images per facial expression of collected data-set

Neutral [NE]	Happy [HA]	Sad [SA]	Angry [AN]	Surprise [SU]	Disgust [DI]	Total
1,000	1,008	465	553	569	501	4,096

2. 데이터 전처리 및 증대(augmentation) 과정

사람의 표정을 인식할 때 얼굴 영역의 데이터만을 이용하여 처리하므로 학습 영상의 얼굴영역만 검출해서 잘라내는 전처리 과정이 필요하다. 제안한 기법에서는 Haar 특징 기반의 얼굴 검출 기법을 활용하여 얼굴영역을 검출하고 잘라낸다[17]. 인간이 다른 사람의 얼굴을 보고 어떤 표정인지 판단할 때는 피부색은 고려하지 않고 눈, 눈썹, 코, 입 등의 모양이나 위치 정보를 이용한다는 점에 착안하여 수집한 영상들을 흑백 영상으로 변환한다. 그림 3은 원본 영상에서 얼굴영역을 검출하여 잘라내고 흑백영상으로 변환한 결과이다.

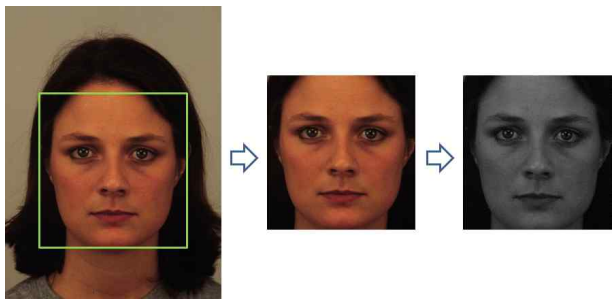

그림 3. 잘라낸 얼굴영역 영상을 흑백영상으로 변환한 결과
Fig. 3. The result of converting cut-out face region image into black and white image

학습 영상의 수가 CNN 구조에 비해 부족하면 분류 성능을 저해시키는 과적합 문제(over-fitting)가 발생하게 되는데 이를 해결하기 위해서 학습 영상의 수를 증가시키는 데이터 증대(data augmentation) 기법을 이용한다. Data-set의 얼굴영상은 대부분의 경우, 그림 3과 같이 목을 곧게 세워 촬영된 영상인데 실생활에서는 카메라의 각도나 사람의 자세에 따라 얼굴 각도가 변할 수 있기 때문에 이를 고려하여 증대시키도록 한다.

먼저 기준 영상에 대하여 시계 방향, 반시계 방향으로 각각 5°, 10°, 15° 만큼 회전 연산한 영상을 획득한다. 그리고 회전된 영상들과 기준 영상을 각각 수평 반전하여 하나의 기준 영상을 14 장의 영상으로 증가시킨다[5]. 그림 4는 데이터 증대 기법을 적용한 결과이다.

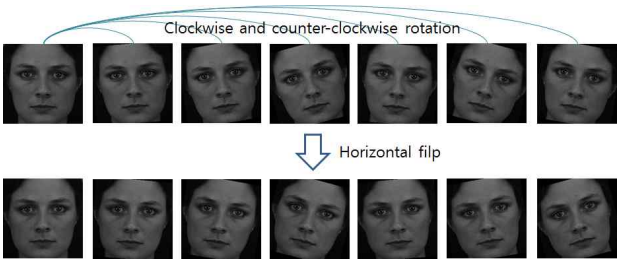

그림 4. 데이터 증대 기법을 적용한 결과
Fig. 4. The result of applying data augmentation technique

3. CNN 구조 최적화

본 논문에서는 초기 CNN 학습 모델을 선택할 때 절대적인 학습데이터도 많지 않고, 분류할 표정의 수도 여섯 가지 밖에 안되기 때문에 기존 CNN 학습 모델 중에서 비교적 얕은 구조인 AlexNet을 참고한다. AlexNet은 100만장 이상의 영상을 학습하고 1000가지의 부류로 분류하는 구조이다. 따라서 적은 학습 데이터를 이용하고 여섯 가지의 표정으로 분류하는 목적에 맞게 구조를 변경할 필요가 있다.

먼저 convolutional layer의 변경요소로는 각 layer의 마스크 필터 크기와 간격 그리고 추출하는 특징 지도의 수이다. Fully-connected layer에서의 변경요소는 layer를 구성하는 node의 수이다. ZFNet[17]의 경우 추출하는 특징 지도의 수는 AlexNet과 동일하게 구성하고 첫 번째 convolutional layer의 마스크 필터 크기를 11에서 5, 간격을 4에서 2로 변경함으로써 분류 성능을 향상시켰다. ZFNet은 AlexNet과 같은 ImageNet의 data-set을 사용하고 분류 성능향상만을 목적으로 하기 때문에 학습 파라미터 용량이나 수행시간은 고려하지 않았다[18].

본 논문에서는 적은 학습 데이터를 이용하여 각 표정을 잘 표현하는 특징 벡터를 추출하는 것과 동시에 영상이 입력되고 분류되기까지 수행 시간이 적게 걸리는 구조로 변경하는 것을 목표로 한다. CNN 구조에서 연산량의 대부분은 convolution 연산 과정에서 요구된다. 아래의 그림 5와 같이 임의의 l 레이어 마스크 필터 크기를 K, 특징 지도 채널수를 M, 연속되는 $l+1$ 레이어의 특징지도 채널수를 M이라고 하면 이 두 레이어 사이에서 발생하는 학습 파라미터 수와 연산량은 아래의 식에 비례한다.

$$(\text{학습파라미터 수, 연산량}) \propto (M \times N \times K \times K)$$

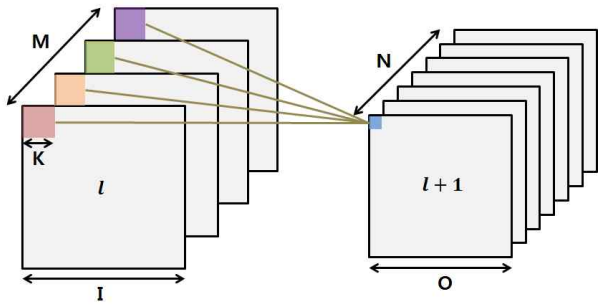

그림 5. 연속되는 convolutional layer 간의 연산 관계
Fig. 5. Computational relationship between consecutive convolutional layers

변경하는 구조는 AlexNet과 동일하게 레이어를 구성하고 마스크 필터의 크기와 간격은 유지하되 각 convolutional layer의 특징지도 수를 감소하여 학습 파라미터의 수와 연산량을 대폭 줄이고자 한다. 또한 fully-connected layer의 노드 개수도 변경하여 분류 성능을 향상시킨다. Convolutional layer의 채널 개수 및 fully-connected layer의 노드 개수를 변경하기 때문에 본 논문에서 제안하는 CNN 구조를 C1-C5 layer의 채널과 FC6-FC8 layer의 노드로 나열된 숫자열로 표현할 수있다. 예를 들어 AlexNet를 이 방법으로 표현하면 (96, 256, 384, 384, 256, 4096, 4096, 1000)이 된다.

각 채널과 노드의 개수는 실험적인 결과를 가지고 정한

다. 하지만 무작위로 정하기에는 경우의 수가 너무 많기 때문에 표 2와 같이 이미 정한 후보 모델에 따라 변경한다. 후보 모델은 AlexNet 구조와 AlexNet의 각 채널과 노드 개수를 1/2, 1/4로 줄인 세 가지의 구조로 한다. 연속되는 레이어 간의 연산이기 때문에 줄인 채널과 노드 개수의 제곱에 비례하여 연산량 및 학습 파라미터가 줄어들게 된다. 학습 파라미터의 급격한 감소는 분류 성능 저하를 일으키기 때문에 1/2, 1/4의 크기로 적당히 감소한 구조를 선택하였다.

동일한 분류 성능을 갖는 모델이 있으면 연산량이 적은 구조를 선택한다. 따라서 표 2의 후보 모델 중에서는 (24, 64, 96, 96, 128, 1024, 1024) 구조를 선택한다. 본 논문에서는 여섯 가지 표정만 분류하면 되기 때문에 적은 수의 노드로도 가능하다. 표 2의 결과에서 fully-connected의 노드 개수 감소로 인한 성능 저하는 발생하지 않는 것을 볼 수 있다.

표 2. 후보 모델의 구성 및 인식률
Table 2. Candidate model configuration and recognition rate

C1	C2	C3	C4	C5	FC6	FC7	인식률(%)
96	256	384	384	256	4096	4096	95.1
48	128	192	192	256	2048	2048	95.6
24	64	96	96	128	1024	1024	95.6

표 3. 첫 번째 기준 모델에서 채널과 노드 개수를 줄인 구조와 상응하는 인식률
Table 3. In the first reference model the structure that reduces the number of channels and nodes and the corresponding recognition rate

C1	C2	C3	C4	C5	FC6	FC7	인식률 (%)
24	64	96	96	128	1024	1024	95.6
12	64	96	96	128	1024	1024	94.3
24	32	96	96	128	1024	1024	95.1
24	64	48	96	128	1024	1024	94.3
24	64	96	48	128	1024	1024	94.5
24	64	96	96	64	1024	1024	94.8
24	64	96	96	128	512	512	94.0

표 2에서 선택한 기준 모델에 대하여 각 레이어의 채널과 노드 개수를 더 줄이기 위해 또 다시 1/2씩 감소하면서 높은 분류성능을 갖는 구조를 찾는 작업을 반복하였다. 그러나 표 3에서 보다시피 표 2에서 선택한 기준 모델보다 가벼운 구조에서는 분류 성능이 향상되는 모델을 찾을 수 없었다.

표 2의 또 다른 모델인 (48 128, 192, 192, 256, 4096, 4096)을 가지고 연산량이 적은 구조를 찾기 위해, 원래 기준 모델과의 중간단계인 (36, 96 ,144, 96, 128, 1024, 1024) 구조를 가지고 채널의 수를 2/3와 1/2로 줄여서 같은 과정을 반복하였다. 표 3에서 노드 FC6, 7을 512로 줄였을 때 분류 성능이 가장 낮았기 때문에 경우의 수도 줄일 겸 FC6, 7의 노드 개수는 1024로 고정한다. 표 4의 진행 과정에서 보면 C4 레이어의 채널 개수를 96으로 줄였을 때 분류 성능이 가장 향상되기 때문에 굵은 선의 (36, 96, 144, 96, 128, 1024, 1024) 구조를 다시 기준 모델로 선택한다.

표 4. 두 번째 기준 모델에서 채널과 노드 수를 줄인 구조와 상응하는 인식률
Table 4. In the second criterion model the structure that reduces the number of channels and nodes and the corresponding recognition rate

C1	C2	C3	C4	C5	FC6	FC7	인식률(%)
36	96	144	144	128	1024	1024	96.1
24	96	144	144	128	1024	1024	95.6
36	64	144	144	128	1024	1024	94.5
36	96	96	144	128	1024	1024	94.3
36	96	144	96	128	1024	1024	96.9
36	96	144	144	64	1024	1024	95.6

위의 과정을 모든 경우의 수에 대하여 반복하여 가장 최적의 구조를 찾는 것은 그 과정이 너무 길어지는 단점이 있다. 따라서 본 논문에서는 일부 채널과 노드의 개수만을 조정하면서 그중에 최적의 구조를 찾으려는 시도를 하였다. 마지막으로 표 5에서 찾은 기준모델의 C4,5, FC6,7은 고정하고 C1-C3만 변경하면서 더 최적의 구조가 있는지 찾아보았다. 기준 모델인 (36, 96, 144, 96, 128, 1024, 1024) 보다 더 성능이 좋은 구조가 없으므로 최종적으로 이 기준 모델을 최적의 구조로 결정한다.

표 5. 세 번째 기준 모델에서 채널과 노드 수를 줄인 구조와 상응하는 인식률
Table 5. In the third reference model the structure that reduces the number of channels and nodes and the corresponding recognition rate

C1	C2	C3	C4	C5	FC6	FC7	인식률(%)
36	96	144	96	128	1024	1024	96.9
24	96	144	96	128	1024	1024	94.8
36	64	144	96	128	1024	1024	95.3
36	96	96	96	128	1024	1024	96.1

그림 6은 제안하는 최적의 구조를 보여주고 있다. 그림 1과 같이 convolutional layer에서 추출하는 특징지도를 GPU 2대에서 반씩 할당하여 처리하는 구조와 달리 제안하는 구조에서는 전체 특징지도를 활용해서 연산하는 구조를 가지며 각 레이어의 채널과 노드 개수도 원래의 AlexNet과는 다르게 구성되었다.

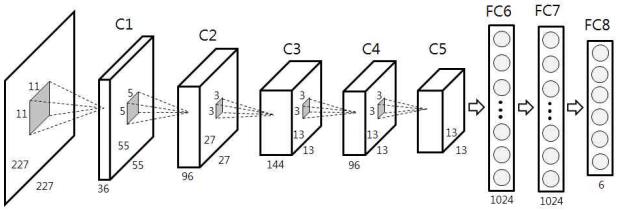

그림 6. 제안하는 최적의 구조
Fig. 6. The proposed optimal structure

4. 실험 결과

제안하는 구조는 Geforce GTX980 TI GPU 기반 Theano tool을 이용하여 설계하였다. 학습(training)과 테스트(test)는 앞에서 구성한 data-set을 9:1의 비율로 구분하여 사용한다. 128 개의 영상을 하나의 batch로 하고 stochastic gradient descent 방법을 이용하여 학습한다. 총 epoch는 60으로 하고 learning-rate는 초기값 0.01을 시작으로 epoch가 20, 40일 때 1/10 크기로 변경한다.

표정인식에 색상정보는 중요하지 않다는 경험적인 의견을 바탕으로 2장에서 설명한 바와 같이 얼굴영역 검출 후 잘라낸 3-채널 컬러영상을 1-채널의 흑백 영상으로 변환하는 전처리 과정을 거친다. 전처리 과정을 거친 흑백의 학습 영상은 여섯 가지 방향(-15°, -10°, -5°, +5°, +10°, +15°)으로 회전 연산을 적용하고 기준영상을 포함한 회전영상을 각각 수평반전하여 각각 열 네장의 영상으로 데이터를 증대시킨다.

3-채널 컬러영상도 마찬가지 방법으로 증대시킨다. 데이터 증대 기법 적용했을 경우와 안했을 경우의 3-채널 영상과 1-채널 영상을 가지고 얼굴 표정 인식률을 비교한다. 표. 6은 3장에서 결정한 최적의 구조 (36, 96, 144, 96, 128, 1024, 1024)에서 데이터 전처리 및 증대 기법이 인식률에 미치는 영향을 보여준다.

데이터 증대 기법을 적용하지 않았을 때 3-채널의 컬러영상을 이용하여 학습하면 1-채널 흑백영상을 가지고 학습한 것과 비교하여 미세하게 인식률이 높은 것을 확인할 수 있다. 하지만 세 개의 채널을 사용하게 되면 첫 번째 convolutional layer에서 연산량이 세 배 많아지게 된다.

데이터 증대 기법을 적용했을 때는 오히려 1-채널의 흑백영상을 가지고 학습할 경우에 인식률이 더 향상되는 것을 표 6에서 알 수 있다. 영상의 채널 수와 상관없이 데이터 증대 기법 적용 여부에 따라 인식률이 크게 차이가 나는 것을 볼 수 있고 영상의 채널 수는 연산량 대비 인식률에는 크게 영향을 미치지 않는 것도 확인할 수 있다.

표 6 데이터 전처리 및 증대 기법의 효과
Table 6 Effects of data preprocessing and augmentation techniques

Preprocessing Method	Accuracy (%)
1-channel gray image	88.80
3-channel color image	88.92
1-channel gray image + data augmentation	96.88
3-channel color image + data augmentation	95.33

다른 CNN 모델과의 성능 비교를 위하여 AlexNet, VGGNet(11-layer)[19], OverFeat(fast model)[20], inception 모듈을 이용한 GoogleNet[21]을 동일한 조건에서 학습시키고, 동일한 테스트 데이터를 적용하여 성능 평가를 수행하였다. 모든 비교 모델들을 여섯 가지 표정 분류 목적에 맞도록 기존의 1000개의 노드를 6개로 의 노드로 변경하였다.

VGGNet의 경우에 batch 크기를 128로 하면 할당 메모리 초과 문제가 발생하기 때문에 32로 줄여서 학습하였다.

먼저 각 모델을 통과하는 데 걸리는 시간을 측정한다. 학습시간 및 테스트 시간을 각각 측정하였다. VGGNet의 경우 batch 크기를 32로 하였기 때문에 동일한 기준을 적용하기 위하여 나머지 모델에 대해서도 batch 크기를 32로 하여 변경하였다. 표 7을 보면 제안하는 구조가 학습과 테스트 시간 모두에서 월등하게 적게 걸리는 것을 알 수 있다.

표 7. 각 모델들의 학습 및 테스트 시간 (batch : 32)
Table 7. Learning and testing time for each model (batch : 32)

Model	Training time (sec / batch)	Test time (sec / batch)
AlexNet	0.107	0.023
OverFeat	0.194	0.040
VGGNet	0.597	0.141
Inception Module	0.111	0.033
Proposed	0.031	0.008

학습데이터를 이용하여 충분히 학습시킨 구조에 대하여 테스트 데이터를 입력하여 분류 성능을 측정하였다. 통합 인식률은 테스트 데이터의 실측 부류 정보가 있기 때문에 실측 부류와 예측된 부류가 같은 경우의 수를 모두 세어서 총 테스트 데이터의 수로 나누어 표현한다. 표 8은 각 모델들의 통합 인식률을 보여준다. 본 논문에서 제안하는 구조의 분류 결과가 가장 뛰어남을 확인할 수 있다.

표 8. 각 모델들에 대한 인식률
Table 8. Recognition rate for each model

Model	인식률 (%)
AlexNet	95.05
OverFeat	95.83
VGGNet	96.35
Inception Module	95.83
Proposed	96.88

2장에서 구성한 data-set를 보면 '무표정'과 '행복' 표정에 대한 데이터가 다른 표정에 비해 많기 때문에 통합 인식률은 두 표정의 분류성능에 더 많은 영향을 받게 된다. 각 표정 별 인식률을 확인하기 위하여 표 9부터 표 13까지 각 모델들에 대한 confusion matrix를 산출하였다. Confusion matrix에서 행은 실측 표정, 열은 예측 표정을 나타내고 각 행렬 값은 실측 표정과 예측 표정이 같을 경우의 확률을 나타낸다. 이를 통해 테스트 데이터가 어떤 표정으로 분류되었는지에 대한 전체적인 분포를 확인할 수 있다.

성능 비교 결과를 보면 다른 모델의 경우 특정 표정에 대한 분류 성능이 떨어지는 경향을 보이는 반면 제안하는 구조에서는 모든 표정에 대하여 92.73% 이상의 좋은 분류 성능을 보이며, 특히 '행복', '놀람', '역겨움' 표정에 대해서는 오류 없는 분류 성능을 보이는 것을 알 수 있다.

표 9. 제안하는 구조의 confusion matrix (%)
Table 9. The confusion matrix of the proposed structure (%)

	NE	HA	SA	AN	SU	DI
NE	93.18	3.41	1.14	2.27	0.00	0.00
HA	0.00	100.00	0.00	0.00	0.00	0.00
SA	4.35	0.00	95.65	0.00	0.00	0.00
AN	0.00	1.82	1.82	92.73	0.00	3.64
SU	0.00	0.00	0.00	0.00	100.00	0.00
DI	0.00	0.00	0.00	0.00	0.00	100.00

표 10. AlexNet의 confusion matrix (%)
Table 10. The confusion matrix of the AlexNet

	NE	HA	SA	AN	SU	DI
NE	95.45	2.27	1.14	1.14	0.00	0.00
HA	0.00	100.00	0.00	0.00	0.00	0.00
SA	2.17	2.17	95.65	0.00	0.00	0.00
AN	1.82	3.64	1.82	83.64	0.00	9.09
SU	0.00	0.00	0.00	0.00	100.00	0.00
DI	0.00	2.00	2.00	4.00	0.00	92.00

표 11. 표 11 VGGNet의 confusion matrix (%)
Table 11. The confusion matrix of the VGGNet (%)

	NE	HA	SA	AN	SU	DI
NE	98.86	1.14	0.00	0.00	0.00	0.00
HA	1.14	97.73	0.00	0.00	1.14	0.00
SA	0.00	0.00	97.83	2.17	0.00	0.00
AN	1.82	1.82	1.82	92.73	0.00	1.82
SU	0.00	0.00	0.00	0.00	100.00	0.00
DI	0.00	4.00	6.00	2.00	0.00	88.00

표 12. OverFeat의 confusion matrix (%)
Table 12. The confusion matrix of the OverFeat (%)

	NE	HA	SA	AN	SU	DI
NE	93.18	0.00	2.27	4.55	0.00	0.00
HA	1.14	98.86	0.00	0.00	0.00	0.00
SA	0.00	2.17	95.65	2.17	0.00	0.00
AN	1.85	1.85	1.85	92.59	0.00	1.85
SU	0.00	0.00	0.00	1.75	98.25	0.00
DI	0.00	0.00	0.00	2.00	0.00	98.00

표 13. Inception module 구조의 confusion matrix (%)
Table 13. The confusion matrix of the Inception module (%)

	NE	HA	SA	AN	SU	DI
NE	94.32	3.41	1.14	1.14	0.00	0.00
HA	0.00	100.00	0.00	0.00	0.00	0.00
SA	6.52	0.00	93.48	0.00	0.00	0.00
AN	7.27	1.82	1.82	87.27	0.00	1.82
SU	0.00	0.00	0.00	0.00	100.00	0.00
DI	0.00	2.00	0.00	0.00	0.00	98.00

Ⅲ. 결 론

본 논문에서는 기존에 많이 활용하던 데이터베이스의 문제점을 보완하고자 고해상도 data-set을 수집하고 선별한다. 또한 불필요한 정보를 제거하기 위해서 얼굴 영역을 검출하여 자르고 1-채널의 흑백영상으로 변환한다. 제안하는 기법에서는 분류성능을 저하시키는 과적합 문제를 해결하기 위해서 학습영상을 증가하는 데이터 증대 기법을 적용하기도 했다.

기존의 CNN 구조에서 convolutional layer의 특징지도 채널 개수 와 fully-connected layer의 노드 개수를 조정하여 분류 처리 시간을 단축할 수있으며 동시에 분류 성능을 향상시키기 위한 최적의 구조를 실험적인 방법으로 결정하였다. 실험결과를 통해 데이터 전처리와 증대 기법의 효과를 확인하였고, 기존의 다른 CNN 모델과 비교하여 본 논문에서 제안하는 구조가 분류 수행시간과 분류 정확도의 성능 모두 뛰어남을 확인하였다.

참 고 문 헌 (References)

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.

[2] Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor, "Going deeper in facial expression recognition using deep neural networks." *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016.

[3] Jung, Heechul, et al. "Joint fine-tuning in deep neural networks for facial expression recognition." *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[4] Lopes, Andre Teixeira, Edilson de Aguiar, and Thiago Oliveira-Santos, "A facial expression recognition system using convolutional networks," Graphics, *Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*, IEEE, 2015.

[5] Hamester, Dennis, Pablo Barros, and Stefan Wermter. "Face expression recognition with a 2-channel convolutional neural network," Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, 2015.

[6] W. Bainbridge, P. Isola, and A. Oliva, “The intrinsic memorability of face photographs,” *Journal of Experimental Psychology: General*, 142(4):1323–1334, 2013.

[7] S. Setty and et al, “Indian Movie Face Database: A Benchmark for FaceRecognition Under Wide Variation,” *In NCVPRIPG*, 2013.

[8] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” *in Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[9] Ma, Correll, and Wittenbrink, *The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data*, Behavior Research Methods, 47, 1122-1135.

[10] ESRC 3D Face Database, http://pics.stir.ac.uk/ESRC/

[11] J. Van der Schalk, S. T. Hawk, A. H. Fischer, and B. J. Doosje, *Moving faces, looking places: The Amsterdam Dynamic Facial Expressions Set (ADFES)*, Emotion, 11, 907-920. DOI: 10.1037/a0023853, 2011.
[12] D. Lundqvist, A. Flykt, and A.Öhman (1998), T*he Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience,* Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.
[13] H. O'Reilly, D. Pigat, S. Fridenson, S. Berggren, S. Tal, O. Golan, S. B"olte, S. Baron-Cohen and D. Lundqvist, *The EU-Emotion Stimulus Set: A Validation Study,* Behavior Research Methods. DOI: 10.3758/s13428-015-0601-4, 2015.
[14] M. Olszanowski, G. Pochwatko, K. Kuklinski, M. Scibor-Rylski, P. Lewinski and RK. Ohme, W*arsaw Set of Emotional Facial Expression Pictures: A validation study of facial display photographs*, Front. Psychol, 5:1516. doi: 10.3389/fpsyg.2014.01516, 2015.
[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.
[16] Learn facial expressions from an image, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
[17] Viola and Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition*, 2001.
[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *In European Conference on Computer Vision* , Springer International Publishing, pp. 818-833, September 2014.
[19] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In Proc. International Conference on Learning Representations,* http://arxiv.org/abs/1409.1556 (2014).
[20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks." *In Proc. ICLR*, 2014.
[21] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki. Dexpression: "Deep convolutional neural network for expression recognition," .*CoRR*, abs/1509.05371, 2015.

저 자 소 개

최 인 규

- 2014년 2월 : 광운대학교 전자공학과 학사
- 2016년 2월 : 광운대학교 전자공학과 석사
- 2016년 3월 ~ 현재 : 광운대학교 전자공학과 박사
- ORCID : http://orcid.org/orcid.org/0000-0002-4239-1762
- 주관심분야 : 영상처리, 컴퓨터비전, 딥러닝

송 혁

- 1999년 2월 : 광운대학교 제어계측공학과 학사
- 2001년 2월 : 광운대학교 전자공학과 석사
- 2013년 2월 : 광운대학교 전자공학과 박사
- 2000년 ~ 현재 : 전자부품연구원 근무
- ORCID : http://orcid.org/0000-0003-0376-9467
- 주관심분야 : 영상 인식, 딥러닝, 영상 보안

이 상 용

- 1987년 : 서강대학교 전자공학과 학사
- 2001년 : 서강대학교 경영대학원 석사(MBA)
- 1987년 ~ 2001년 : ㈜데이콤 기술전략실
- 2001년 ~ 2008년 : ㈜브로드밴드솔루션스(BSI) 대표
- 2007년 ~ 2015년 : CJ HelloVision CTO & COO
- ORCID : http://orcid.org/0000-0003-0210-3591
- 주관심분야 : Digital Media Center, Smart Home, Cloud Broadcasting Platform

저 자 소 개

유 지 상

- 1985년 2월 : 서울대학교 전자공학과 학사
- 1987년 2월 : 서울대학교 전자공학과 석사
- 1993년 5월 : Purdue Univ. EE, ph.D
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : http://orcid.org/0000-0002-3766-9854
- 주관심분야 : 영상처리, 컴퓨터비전, 차세대 방송기술