Journal of Korea Multimedia Society Vol. 23, No. 12, December 2020(pp. 1464-1475)
https://doi.org/10.9717/kmms.2020.23.12.1464

딥러닝 표정 인식을 활용한 실시간 온라인 강의 이해도 분석

이자연†, 정소현†, 신유원†, 이은혜†, 하유빈†, 최장환†

Analysis of Understanding Using Deep Learning Facial Expression Recognition for Real Time Online Lectures

Jaayeon Lee†, Sohyun Jeong†, You Won Shin†, Eunhye Lee†,
Yubin Ha†, Jang-Hwan Choi†

ABSTRACT

Due to the spread of COVID-19, the online lecture has become more prevalent. However, it was found that a lot of students and professors are experiencing lack of communication. This study is therefore designed to improve interactive communication between professors and students in real-time online lectures. To do so, we explore deep learning approaches for automatic recognition of students' facial expressions and classification of their understanding into 3 classes (Understand / Neutral / Not Understand). We use 'BlazeFace' model for face detection and 'ResNet-GRU' model for facial expression recognition (FER). We name this entire process 'Degree of Understanding (DoU)' algorithm. DoU algorithm can analyze a multitude of students collectively and present the result in visualized statistics. To our knowledge, this study has great significance in that this is the first study offers the statistics of understanding in lectures using FER. As a result, the algorithm achieved rapid speed of 0.098sec/frame with high accuracy of 94.3% in CPU environment, demonstrating the potential to be applied to real-time online lectures. DoU Algorithm can be extended to various fields where facial expressions play important roles in communications such as interactions with hearing impaired people.

Key words: Degree of Understanding, Real-time Analysis, Face Detection, Facial Expression Recognition, Deep Learning

1. 서 론

전 세계적으로 COVID-19가 확산됨에 따라 대부분 대학교가 강의를 온라인 비대면 형태로 전환했다. 이렇게 강의 형태가 변하면서, 오프라인 수업에 비해 교수와 학생 간의 소통이 부족하다는 문제가 제기되고 있다. 이에 대해 이화여자대학교 교수와 학생을 상대로 설문을 진행한 결과, 실제로 62.7%의 교수와

※ Corresponding Author : Jang-Hwan Choi, Address: Research Cooperation Building 209, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Republic of Korea, TEL : +82-2-3277-4759, FAX : +82-2-3277-3275, E-mail : choij@ewha.ac.kr
Receipt date : Aug. 28, 2020, Revision date : Oct. 20, 2020
Approval date : Nov. 11, 2020
† Division of Mechanical and Biomedical Engineering, Ewha Womans University
(E-mail : sally5200@ewhain.net)
(E-mail : jsh98021275@gmail.com)
(E-mail : swon617@gmail.com)
(E-mail : leeeh0921@gmail.com)
(E-mail : yubinha98@gmail.com)
※ This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (N0001111, Innovative Engineering Education) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2020R1F1A1073774).

86.3%의 학생이 '소통이 원활하지 않다'라고 응답했다. 또한, '학생의 웹캠 화면을 통해 표정을 살펴 이해도를 파악'하는 교수의 대부분이 수업 중 '학생들의 반응, 표정을 파악하기 어렵다'라고 응답하였으며, 학생들의 수업 이해도를 실시간으로 분석하여 보여주는 서비스가 필요한지를 묻는 질문에 '예'라고 응답한 교수의 비율이 76.5%였다. (Appendix 1, 2)

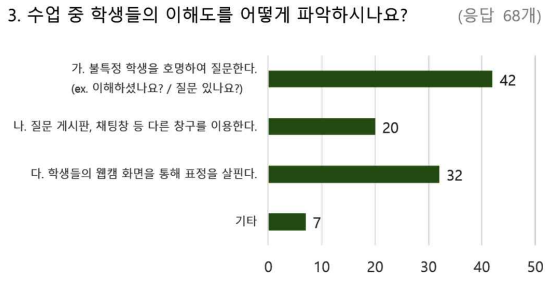
표정은 대면 수업 시 가장 빈번하게 사용하는 비언어적 소통 수단이며[1], 표정을 통해 알 수 있는 학생의 감정은 이해도와 서로 유의미한 관계가 있다[2]. 이는 표정을 바탕으로 학생의 이해도를 분석하는 것이 가능하며 그에 대한 시스템이 필요함을 뒷받침한다.

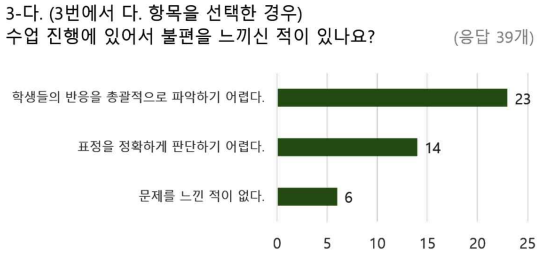
얼굴 표정 인식(Facial Expression Recognition, FER) 시스템이 개발된 이후로 이를 소셜 로봇, 의료, 운전자 모니터링 등 다양한 분야에 활용하는 연구가 활발하게 진행되어 왔다[3]. 그러나 이를 교육 분야에 적용하여 의미를 도출한 선행연구는 상대적으로 부족하다. 또한, 그에 관한 기존 연구들은 단순히 happy, surprise 등과 같은 감정 상태를 그대로 제시하여 실시간 피드백이 불가능했다[4-6].

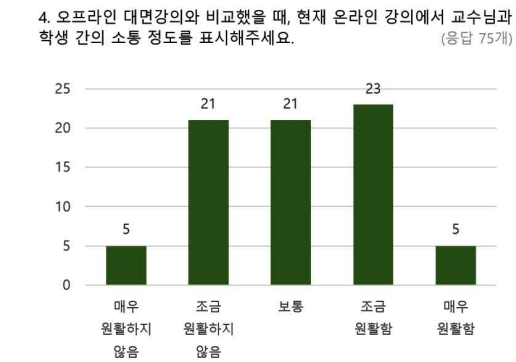
따라서 본 연구는 딥러닝 모델을 통해 표정 인식

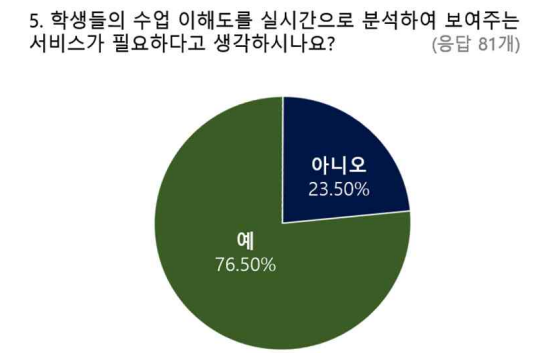
Appendix

1. 교수 설문 조사 결과

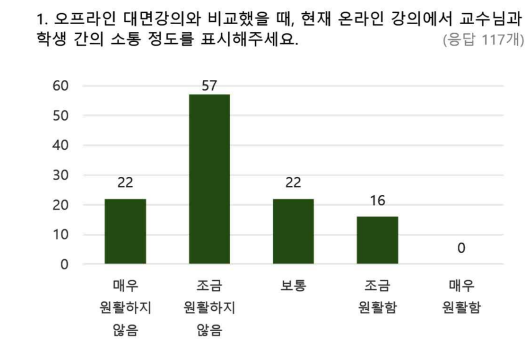








2. 학생 설문 조사 결과



3. 코드

본 연구에 사용한 코드는 다음 GitHub 링크에 기재되어 있습니다.
https://github.com/jaayeon/emotion_classification

의 정확도를 높이고, 실시간으로 다수 사용자의 표정을 이해도 클래스로 분류함으로써 기존 선행연구들의 단점을 보완하고자 한다. 더 나아가 분류 결과에 대한 통계를 실시간으로 제공하여 사용자 전체의 이해도를 한눈에 파악할 수 있게끔 시각화하는 시스템을 제공하고자 한다.

2. 이전 연구

기존에 이루어진 연구들은 주로 공공 데이터를 활용하여 감정을 happy, fear, disgust, angry, surprise, neutral, sad와 같이 7가지로 분류하였다[7-9].

학습 환경에서는 학습자의 감정이 중요한 요인으로 작용한다[1]. 그러나 기존의 대면 강의와 달리, e-learning 환경에서는 교수자와 학습자 간의 상호작용이 부족하여 교수자가 학습자의 감정 상태를 파악하기 어려워 학습 효율이 비교적 낮게 나타난다[23]. 이러한 e-learning 플랫폼의 문제를 해결하기 위해, FER 을 활용하여 표정으로 드러나는 학습자의 감정을 분석하는 연구들이 활발하게 진행되어 왔다. 이후, 감정 데이터를 활용하여 집중도와 참여도를 분석하는 연구로의 발전도 이루어졌다[24-26].

교육의 목적으로 FER 을 통해 표정을 분류한 선행연구들은 Table 1과 같이 진행되었다. 각 연구에서 사용한 face detection 알고리즘과 표정 분류를 위한 FER 알고리즘을 정리하였다.

기존 연구들은 Table 1에서와같이 face detection 모델로 Haar Cascade, Active Appearance Model (AAM), Active Shape Model (ASM), Face detection using an Artificial Neural Network(ANN) 등 간단한 모델을 사용하였다. FER 모델로는 Support Vevtor Machine (SVM)이 가장 많이 사용되었으며 그 외 Deep Belief Network(DBN), Convolutional Neural Network(CNN), Principal Component Analysis(PCA) 등이 사용되었다.

위 논문들의 경우 주로 개별적인 학습이 이루어지

Table 1. Related Works

Paper	Face Detection Algorithm	Facial Expression Recognition Algorithm
Sarrafzadeh *et al.* (2008) [10]	1. Face detection using an Artificial Neural Network (ANN) 2. Facial feature extraction and a fuzzy facial expression classifier	1. Linear model 2. Polynomial model 3. RBF Kernel with the Support Vector Machine (SVM)
C.-H. Wu (2016) [11]	FaceSdk	Support Vector Machine (SVM), Decision tree classifier
L. Chen *et al.* (2012) [12]	Active Shape Model (ASM)	Support Vector Machine (SVM) with Gabor wavelet
M.-P. Loh et al (2006) [13]	None	Convolutional Neural Network (CNN) with Gabor wavelet
J.-M. Sun *et al.* (2008) [14]	Active Appearance Model (AAM)	Support Vector Machine (SVM)
O. El Hammoumi *et al.* (2018)[6]	Haar Cascade	Modified Convolutional Neural Network (CNN)
O. K. Akputu e*t al.* (2018) [15]	Viola and Jones face detector [16]	Multiple Kernel Learning Decision Tree Weighted Feature Alignment (MKLDT-WFA) with Gabor Wavelet
U. Ayvaz *et al.* (2017) [17]	Facial landmarks detection algorithm proposed by Sagonas et al [18].	Support Vector Machine (SVM)
S. P. Deshmukh *et al.* (2018) [19]	Haar classifier	Kernel Support Vector Machine (SVM)
G. Tonguç *et al.* (2020) [20]	Facial Movements Coding System (FACS)	Microsoft Emotion Recognition API
M. A. A. Dewan *et al.* (2018) [21]	Viola and Jones face detector [16]	Deep belief network (DBN) proposed by Hinton et al [22]

는 e-learning에 적합한 시스템을 제안했다. 하지만 현재의 경우는 수업이 실시간 온라인 비대면 강의로 진행되는 상황이므로 위의 논문과는 달리 실시간으로 다수의 학습자를 일괄적으로 평가해야 한다는 점을 고려해야 한다. 그러므로 여러 장의 연속 이미지를 빠른 속도로 분석 및 판단하는 적합한 방법이 필요하다. 따라서 본 연구는 다양한 알고리즘을 비교, 분석하여 속도를 기준으로 연구의 목적에 적합한 최적의 알고리즘을 제안하고자 한다.

3. 제안한 방법

본 논문에서 제안하는 표정 인식 기반 온라인 실시간 강의 시스템의 전체적인 프로세스는 Fig. 1과 같다. 기존의 여러 공공 데이터셋을 본 연구의 취지에 맞게 재구성하여 연구를 진행하였다. 먼저 face detection 모델을 이용하여 데이터셋으로부터 정확한 얼굴의 위치를 얻고, 이를 FER 의 인풋으로 사용하였다. 최적의 결과를 도출하기 위해, face detection과 FER에 여러 모델을 적용하였고 학습 결과를 비교 분석하여 모델을 선정하였다.

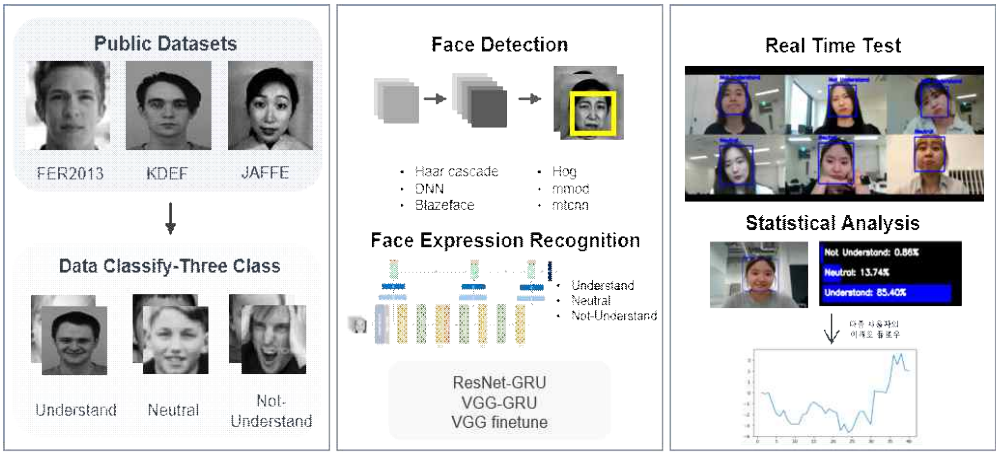
학습 데이터셋으로는 FER2013[27], JAFFE[28], KDEF[29] 공공 데이터셋을 본 연구의 취지에 맞게 재구성하여 연구를 진행하였다. 감정 상태와 이해도가 유의미한 관계가 있다는 연구에 기반하여 Neutral, Happy, Sad, Surprised, Angry, Disgust, Fear 총 7가지의 감정으로 분류된 데이터셋을 본 연구의 취지에 맞게 Not Understand, Neutral, Understand로 재구성하였다[2]. 표정 분류에는 총 6명의 학부생이 참여하였으며 Sad, Angry, Disgust, Fear는 Not Understand로, Happy는 Understand로 재분류하였다. Surprised는 표정에 따라 적절한 클래스로 재분류하였다. 입꼬리가 올라갔거나 미간이 찌푸려지지 않았다면 이해됨으로 분류하였으며 입꼬리가 내려가거나 미간이 찌푸려진 경우 이해되지 않음으로 분류하였다. 이에 해당하지 않는 Surprised 이미지는 분류에 참여한 사람들이 모두 동의할 시 데이터셋에서 제외하였다(Table 2).

Table 2. The number of classified images of each dataset

Dataset \ DoU	Understand	Neutral	Not Understand
FER2013	3,063	1,943	3,379
JAFFE	36	30	122
KDEF	279	203	826

본 연구는 실시간 표정 분석을 통해 학생의 이해도를 실시간으로 교수에게 보여준다. 다중 학생에 대한 분석일 경우, 해당 개인 수치의 평균 통계치를 활용해 학생들의 전체적인 이해도를 나타낸다.

모든 공공 데이터셋은 흑백 이미지이며 FER 의 인풋 이미지로 사용하기 위해 64×64로 크기 조정 후 수평, 수직 방향의 flip, 90° rotation의 image augmentation을 진행하였다.



Ⅰ. Data Processing Ⅱ. Modeling Ⅲ. Real-Time Test & Analysis

Fig. 1. Overall process of online real-time e-learning system based on FER.

3.1 Detection

FER 인풋으로 들어갈 얼굴 사진을 얻기 위한 전처리 단계로 각 프레임에서 face detection을 통해 얼굴 사진을 얻는다. 좋은 detection 성능을 보이는 State-of-the-art 알고리즘 및 전통적인 face detection 알고리즘을 선정하여 속도를 비교 분석하였다. 머신러닝 모델인 Haar Cascade, Histogram of Oriented Gradients(HOG)와 딥러닝 기반의 Deep Neural Network(DNN), Blazeface, Multi-Task cascaded Convolutional Neural Network (MTCNN)을 비교, 분석하였다[30, 32-35]. Haar cascade와 DNN은 OpenCV에서 제공되는 모듈을 사용하였으며, HOG는 DLib에서 제공되는 모듈을 사용하였다.

3.1.1 BlazeFace

BlazeFace[30]는 6개의 얼굴 특징점(눈, 귀, 입, 코)을 추출하여 얼굴의 회전각을 예측하고 그 결과를 사각형 영역으로 나타낸다. 이 모델은 MobileNet 기반의 구조를 가지며, 모바일 장치에 특화되어 있다. MobileNet[31]에서 Depth-wise separable convolution 연산과 Residual connection을 차용하여 연산량을 대폭 줄였다. 또한, layer 개수를 줄여 5x5 kernel을 사용하는 Bottleneck 구조를 새로 추가함으로써 이미지 수용 영역을 확대했다. 그뿐만 아니라 896개의 미리 정의된 크기의 bounding box (anchor)를 사용하여 객체의 위치를 융통성 있게 조정한다. 이때, box의 회귀 파라미터(regression parameter)는 예측값들의 가중 평균으로 계산한다.

3.1.2 DNN

DNN[32]은 여러 물체에 해당하는 영역을 예측하는 네트워크로, Convolutional Deep Neural Network를 기반으로 한다. DNN에 기반한 detection의 경우 물체의 마스크를 DNN을 기반으로 한 회귀를 통해 물체 분류(object classification) 및 기하학적 정보를 추출한다.

3.1.3 Haar Cascade

Haar Cascade[33]는 2001년에 제안된 알고리즘으로 파라미터의 영향을 많이 받아 정확도가 높지 않다. Haar feature를 이용하여 이미지에서 각 영역 간의 밝기 차를 반영한 얼굴 특징점을 뽑은 후, cascade 함수로 detection 한다.

3.1.4 MTCNN

MTCNN[34]은 joint learning 방식을 이용한 것으로 CNN인 P-net, R-net, O-net을 차례로 통과하는 cascade 모델이다. 각 네트워크는 face detection, bounding box regression, face alignment를 학습한다.

3.1.5 HOG

HOG[35]는 image descriptor와 Linear Support Vector Machine을 사용하는 모델이다. HOG는 대상 영역을 일정 크기로 나누고, 나눈 영역 내 edge orientation의 히스토그램 정보를 이용하여 얼굴 영역을 detection 한다. Edge 정보를 이용하기 때문에 이미지의 밝기 변화 등에 덜 민감하다.

위 다섯 개의 모델은 모두 학습된 파라미터를 가져와 성능을 평가하였으며, 비교 결과 BlazeFace가 가장 빠른 속도를 보여 최종 모델로 채택하였다(Table 3). Blazeface를 통해 얻은 얼굴 사진이 FER 알고리즘의 인풋으로 사용된 데이터셋의 얼굴 범위와 달라, 동일하게 맞추기 위하여 Blazeface의 아웃풋인 detection box의 수직 방향을 1.3배 늘려 얼굴 사진을 얻었다.

Table 3. Speed of each detection model

Model	Speed (CPU) [sec / 4 frames]
BlazeFace	0.089
DNN	0.209
Haar Cascade	0.244
MTCNN	0.648
Hog	0.149

3.2 Facial Expression Recognition(FER)

이해도를 분석하기 위한 Facial Expression Recognition 알고리즘은 실시간에 적합하도록 상대적으로 가벼우며 일정 수준 이상의 정확도를 보장해야 한다. 따라서 가볍고 빠른 VGG, ResNet 기반의 모델을 고안하였으며, 시계열 데이터 분석에 적합한 GRU 모델을 추가하였다.

3.2.1. VGG

Modified VGG 모델[36]은 3x3의 작은 receptive kernel만을 사용한 convolutional layer 5층과 fully

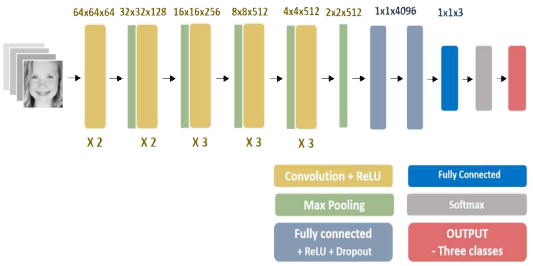


Fig. 2. Architecture of VGG.

connected layer 3층으로 구성된다. 3×3 kernel의 spatial padding은 1로 고정되어 있으며 spatial pooling은 5개의 max-pooling layer로 구현되었다. Convolutional layer 이후 Fully-connected(FC) layer가 이어진다. 앞의 두 FC layer는 4096 channels를 가지며 마지막 layer는 3 channels 아웃풋을 가진다(Fig. 2 [36]).

3.2.2 VGG-GRU

VGG-GRU 모델은 VGG 모델의 FC layer 중 일부를 삭제한 후, Gated Recurrent Unit (GRU) [37] 모델을 합친 네트워크이다. GRU는 [38]에서 제안된 Long Short-Term Memory (LSTM) unit과 달리 memory cell을 나누지 않은 gating unit을 가진다. Gating unit은 information flow를 저장하는 단위이다[37, 38]. 본 모델에서는 같은 클래스의 이미지들이 batch로 묶여 들어간다. (Fig. 3 [39]).

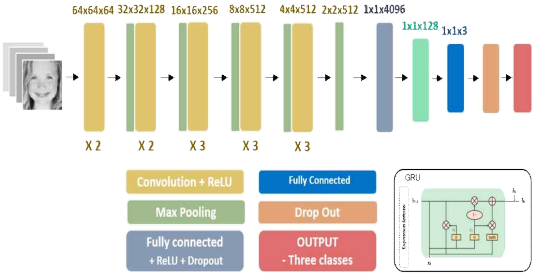


Fig. 3. Architecture of VGG-GRU.

3.2.3 ResNet-GRU

ResNet-GRU 모델은 gradient vanishing 문제를 해결한 ResNet[40]을 backbone으로 한다. 모델의 인풋으로 같은 클래스의 이미지가 batch로 묶여 들어간다. 3×3의 receptive kernel만 사용되며 40%의

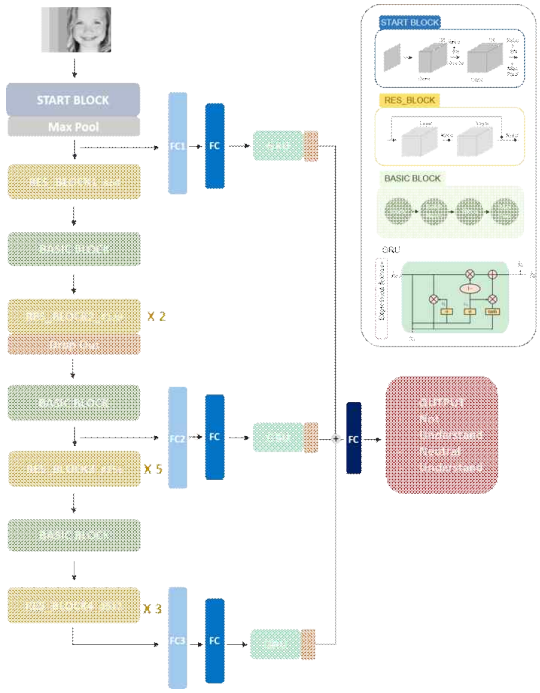


Fig. 4. Architecture of ResNet-GRU.

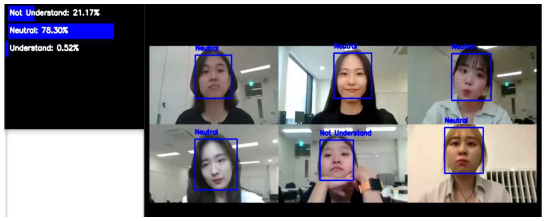
convolutional dropout, 80%의 FC layer dropout을 추가했다. ResNet 기반의 각기 다른 3개의 block에서 뽑은 1D features는 각각 FC layer를 통과한 후 sequence length는 1로 차원이 확장되어 GRU 모델의 인풋으로 들어간다. GRU 모델 아웃풋 합의 batch 간 평균이 마지막 FC layer의 인풋으로 들어가 최종적으로 3 channel의 모든 batch 이미지를 고려한 아웃풋을 얻게 된다 (Fig. 4 [39]).

VGG, ResNet-GRU는 Adam optimizer, VGG-GRU는 Stochastic Gradient Descent (SGD) optimizer을 사용하였으며, learning rate는 1e-05이다. VGG, VGG-GRU, ResNet-GRU의 Training batch size는 각각 128, 4, 4이다. FER 모델은 모두 PyTorch framework를 사용하여 구현되었다.

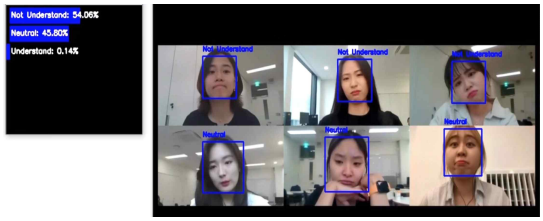
위 세 개의 모델을 비교해 본 결과, 'ResNet-GRU'가 가장 높은 정확도를 보여 최종 모델로 채택하였다 (Table 4).

4. 이해도 평가 모델 파일럿 테스트

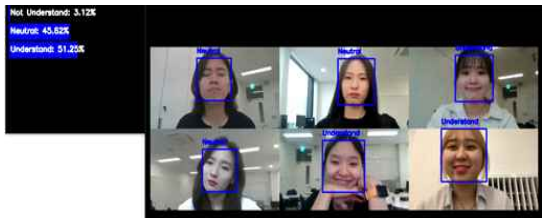
본 연구에서는 구현한 DoU 알고리즘이 실제 수업에서 어떠한 방식으로 활용이 가능한지 보이기 위해



(a)



(b)



(c)

Fig. 5. Prototype and statistics of understanding. (a) Neutral state, (b) Not-understanding state, (c) Understanding state.

Table 4. Accuracy and speed of each FER model

Model	Accuracy	Speed [sec / 4 frames]	
		CPU	GPU
VGG	45.93%	-	-
VGG-GRU	82.70%	0.213	0.024
ResNet-GRU	94.30%	0.303	0.032

Zoom 화상 회의 애플리케이션을 사용하여 실제 온라인 실시간 수업과 같은 환경을 조성하였다. 이 때, 회의에 참여한 6명의 학생이 모두 수업을 듣고 있다는 가정하에 각자 자유롭게 표정을 짓고 그 화면을 녹화하였다. 이 영상을 DoU 모델에 넣고 사용자들의 이해도를 실시간으로 Understand / Neutral / Not Understand의 세 개 클래스로 분류했다. 학생 모두의 클래스별 백분율 수치를 평균 내어 막대그래프로 나타냈다(Fig. 5). 속도 측정은 Windows 10 Home, Intel(R) Core(TM) i7-8565U CPU 환경에서 진행되었다.

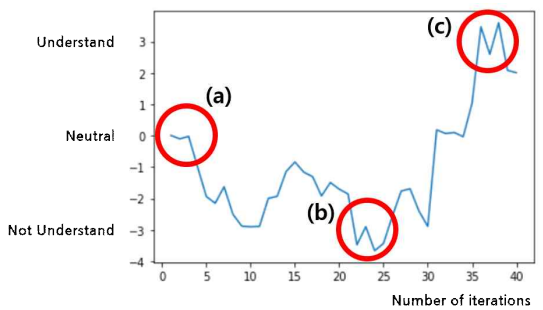


Fig. 6. DoU Graph per iteration.

나아가 수업 중 어느 시점에서 학생들의 이해도가 낮았는지를 직관적으로 확인하기 위해, 시간 프레임에 따른 이해도 분석 결과를 그래프로 나타냈다(Fig. 6). 한 사람에 대한 특정 시간의 이해도 값을 얻기 위하여 3개의 이해도 클래스에 대한 확률값 중 가장 높은 값 만을 그래프에 활용하였다. Understand에 속할 시 해당 클래스의 확률값을 양의 값으로, Neutral에 속할 시 확률값을 0으로 설정하였으며, Not Understand에 속할 시 해당 클래스의 확률값을 음수로 표기하여 한 사람의 시간에 따른 이해도 그래프를 얻었다. 모든 학생들의 이해도 그래프를 합하여 최종 그래프를 도출하였다. 따라서 이해도가 Neutral로 분류된 학생들이 많은 구간에서는 0에 가까운 값이(a), 이해하지 못한 학생들이 많은 구간에서는 음수 값이(b), 이해한 학생들이 많은 구간에서는 양수 값이 출력되었다(c). 여기서 구간 (a), (b), (c)의 값은 각각 Fig. 5의 프로토타입 장면에서 출력된 결과이다.

본 연구의 설계 요구 조건은 신뢰성을 줄 수 있는 정확도와 실시간에 적용 가능한 속도이다. Table 4와 Table 5에서 제시된 테스트 결과는 한 사람에 대하여 추출한 4개의 frame을 batch로 하여 측정되었다. 테스트 결과, DoU 알고리즘은 94.3%의 정확도(Table 4)와, CPU 환경에서 face detection부터 FER 까지 0.392초의 빠른 속도를 보였다(Table 5). 같은 알고리즘을 GPU 환경에 적용하였을 때는 0.121초로 속도가 더욱 빨라져 성능이 향상됨을 확인하였다.

Table 5. Speed of DoU algorithm

Model	Speed (CPU) [sec / 4 frames]
Detection (BlazeFace)	0.089
FER (ResNet-GRU)	0.303
Total	0.392

5. 고 찰

본 연구는 딥러닝 기반 표정 인식을 통해 온라인 실시간 강의에서 학생의 이해도를 실시간으로 분석하는 서비스를 제안한다. 신뢰성 있는 결과 도출을 위해, face detection과 FER 모델을 함께 사용했으며 여러 모델의 비교, 분석을 통해 최적의 결과를 도출하였다. BlazeFace와 ResNet-GRU 모델로 구성된 DoU 알고리즘을 제안했으며, 이는 안면인식 기술을 활용한 수업의 이해도 통계를 실시간으로 분석한 첫 번째 연구라는 점에 큰 의의가 있다.

현재 FER과 관련한 연구가 많이 진행되고 있다. 두 단계로 이루어진 전통적인 기법과는 달리 최근에는 end-to-end learning의 딥러닝 기법이 주로 사용되고 있다[3]. 이미지로부터 얼굴 특징점의 값을 추출하고 그 값들을 표정 인식에 활용하기보다, 이미지 자체를 학습해 얼굴 특징점 추출과 동시에 표정 인식을 하는 딥러닝 모델이 동영상의 각 프레임을 실시간으로 분석하는 본 연구에 적합하다고 여겼다. 또한, 얼굴 위치에 대한 오차를 줄이기 위해 정확한 얼굴을 딥러닝을 통해 찾고 이를 테스트에 적용해 정확도를 높이고자 하였다. 그뿐만 아니라, 실제 영상을 분석하기 위해 순차적인 데이터에 적합한 모델을 선정하고자 하였다. 따라서 Convolutional Neural Network와 Recurrent Neural Network가 결합한 모델을 비교 분석하였다. 그 결과 우리가 제안하는 최적의 모델 두 가지는 다음과 같다.

Face detection에서 가장 좋은 성능을 보인 Blaze Face는 Depth wise separable convolution 연산과 residual connection 기법을 활용하여 속도를 단축하였다. 또한, 영상의 이전 프레임에서 계산된 6개의 얼굴 특징점의 값을 다음 프레임에 반영함으로써 연산량을 대폭 줄였다. 그뿐만 아니라 다양한 크기의 anchor를 사용하여 얼굴의 위치를 유연하게 조절하였다. 이와 같은 Blazeface 모델의 장점들이 빠르고 정확한 결과의 도출로 이어졌다.

FER 에 가장 좋은 성능을 보인 ResNet-GRU의 residual learning은 모델의 비선형성을 높여 복잡한 바운더리로의 근사에 탁월한 성능을 보였다. 또한, 3단계의 field of view(FOV)가 GRU 모델의 인풋으로 주어지므로 표정의 중요한 정보를 더욱 효과적으로 뽑아낼 수 있었다. 적절히 추가된 convolutional dropout과 FC layer dropout은 모델의 파라미터 수를 대폭 줄여 실시간으로 적용 가능한 속도를 보장함에 큰 도움을 주었다.

최근 FER 관련 연구에서 비디오 데이터가 학습 데이터로 많이 사용된다[41, 42]. 하지만 비디오 데이터는 프레임 간 변동이 크고 잡신호가 많은 저품질의 이미지가 인풋으로 사용된다는 단점이 있다[43]. 따라서, 고품질의 이미지로 정확한 분류를 진행하기 위해 본 연구에서는 이미지 데이터를 학습 데이터로 사용하였다. 비디오 데이터가 사용될 경우 여러 개의 연속적인 프레임으로 모델에 들어간다. 이와 유사한 효과를 나타내기 위해, 동일 클래스에 해당하는 4개의 이미지를 하나의 배치로 모델에 넣어 학습하였다.

또한, 본 연구에서 사용된 트레이닝 데이터셋이 이해도를 측정하기 위해 구성된 데이터가 아닌 감정 데이터셋을 기반으로 재구축된 데이터셋임에 한계가 있다. 실제 수업에서는 Neutral한 표정도 상황에 따라 긍정과 부정의 의미를 모두 포함할 수 있으며, Neutral한 표정을 지어도 끄덕임, 고개를 젓는 행위 등 사용자의 행동 패턴에 따라 의미가 달라질 수 있기 때문이다. 따라서 이해도를 평가하는 비디오 기반의 데이터셋을 구축하여 얼굴의 생리학적 정보뿐만 아니라 사람이 이해했을 때 나타나는 동공, 근육의 움직임, 그리고 얼굴의 기울어진 방향 등의 추가적인 정보와 함께 표정 인식에 활용한다면 더욱더 세밀하고 정확한 분석 서비스를 제공할 수 있을 것이다[43-46].

본 시스템은 비대면 실시간 강의가 활발하게 진행되는 현시점에 가장 적합한 형태의 교육 시스템으로서, 다양한 학습 환경에서 매우 높은 활용도가 예상된다. 본 시스템이 제공하는 실시간 통계 수치를 통해 교수자는 학생들이 강의 중 이해하지 못한 부분을 직관적으로 파악할 수 있으며, 해당 내용에 추가적인 설명을 덧붙임으로써 적절하고 빠른 피드백을 할 수 있게 된다. 이는 학습자와 교수자 간의 소통 개선에 크게 이바지하여, 학습의 효율성을 더욱 높일 수 있

을 것으로 기대된다.

6. 결 론

앞에서 언급된 바와 같이, 본 연구에서 고안한 시스템은 인풋 영상을 BlazeFace를 통한 Facial Detection과, ResNet-GRU를 통한 FER 을 거쳐 이해도(Understand / Neutral / Not Understand)를 분석한 뒤 통계 결과를 도출하는 순서로 진행된다. 이 시스템은 결과적으로 CPU 환경에서 4개의 프레임당 0.392초의 빠른 속도와 94.3%의 높은 정확도를 나타냈다.

통제된 조건에서 시스템은 95% 정도의 훌륭한 정확도를 보이지만, 실제 영상에서 테스트한 경우에는 통제된 조건에서 만큼의 정확도가 나타나지 않는다는 한계점이 있다. 하지만 이해도에 따라 라벨링 된 데이터셋을 구축하여 추후 연구를 진행한다면, 실제 이해도 테스트에서 더 높은 정확도가 도출될 것이다.

이를 바탕으로 사용자의 이해도뿐만 아니라 집중도까지 분석할 수 있는 기능까지 결합한다면 교육환경 개선에 더욱 크게 이바지할 수 있을 것이다. 더 나아가 다양한 분야로의 적용 또한 기대해 볼 수 있다. 대표적으로 표정을 통한 표현이 소통에 있어 매우 중요한 청각 장애인들의 의사소통에 도움을 줄 수 있을 것이다.

REFERENCE

[1] M.H. Immordino-Yang and A. Damasio, “We Feel, Therefore We Learn: The Relevance of Affective and Social Neuroscience to Education,” *Mind, Brain, and Education,* Vol. 1, No. 1, pp. 3-10, 2007.

[2] M. Sathik and S.G. Jonathan, “Effect of Facial Expressions on Student’s Comprehension Recognition in Virtual Educational Environments,” *SpringerPlus,* Vol. 2, No. 1, pp. 455-463, 2013.

[3] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” *IEEE Transactions on Affective Computing*, 2020. (Accepted)

[4] J. Park, S. Jeong, W. Lee, and K. Song, “Analyzing Facial Expression of a Learner in E-Learning System,” *Proceedings of the Korea Contents Association Conference,* pp. 160-163, 2006.

[5] K. Otwell, *Facial Expression Recognition in Educational Learning Systems*, 10319249, US, 2019.

[6] O. El-Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El-Kafi, and A. El-Hore, “Emotion Recognition in E-learning Systems,” *Proceeding of International Conference on Multimedia Computing and Systems,* pp. 1-6, 2018.

[7] A. Sarrafzadeh, S. Alexander, F. Dadgostar, C. Fan, and A. Bigdeli, “How Do You Know that I Don’t Understand?” A Look at the Future of Intelligent Tutoring Systems,” *Computers in Human Behavior,* Vol. 24, No. 4, pp. 1342-1363, 2008.

[8] O.K. Akputu, K.P. Seng, Y. Lee, and L. Ang, “Emotion Recognition Using Multiple Kernel Learning toward E-Learning Applications,” *ACM Transactions on Multimedia Computing, Communications, and Applications,* Vol. 14, No. 1, pp. 1-20, 2018.

[9] P. Viola and M.J. Jones, “Robust Real-Time Face Detection,” *International Journal of Computer Vision,* Vol. 57, No. 2, pp. 137-154, 2004.

[10] C. Wu, “New Technology for Developing Facial Expression Recognition in E-Learning,” *Proceeding of Portland International Conference on Management of Engineering and Technology,* pp. 1719-1722, 2016.

[11] U. Ayvaz, H. Gürüler, and M.O. Devrim, “Use of Facial Emotion Recognition in E-Learning Systems,” *Information Technologies and Learning Tools,* Vol. 60, No. 4, pp. 95-104, 2017.

[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge,” *Proceedings of the IEEE*

International Conference on Computer Vision Workshops, pp. 397-403, 2013.

[13] L. Chen, C. Zhou, and L. Shen, "Facial Expression Recognition based on SVM in E-Learning," *Ieri Procedia,* Vol. 2, pp. 781-787, 2012.

[14] S.P. Deshmukh, M.S. Patwardhan, and A.R. Mahajan, "Feedback Based Real Time Facial and Head Gesture Recognition for E-Learning System," *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data,* pp. 360-363, 2018.

[15] M. Loh, Y. Wong, and C. Wong, "Facial Expression Recognition for E-Learning Systems Using Gabor Wavelet & Neural Network," *Proceeding of IEEE International Conference on Advanced Learning Technologies,* pp. 523-525, 2006.

[16] G. Tonguç and B.O. Ozkara, "Automatic Recognition of Student Emotions from Facial Expressions during a Lecture," *Computers & Education,* Vol. 148, pp. 103797, 2020.

[17] J. Sun, X. Pei, and S. Zhou, "Facial Emotion Recognition in Modern Distant Education System Using SVM," *Proceeding of International Conference on Machine Learning and Cybernetics,* pp. 3545-3548, 2008.

[18] M.A.A. Dewan, F. Lin, D. Wen, M. Murshed, and Z. Uddin, "A Deep Learning Approach to Detecting Engagement of Online Learners," *Proceeding of IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation,* pp. 1895-1902, 2018.

[19] G.E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* Vol. 18, No. 7, pp. 1527-1554, 2006.

[20] I.M. Revina and W.S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques," *Journal of King Saud University - Computer and Information Sciences*, Vol. 1, No. 5, pp. 1-9, 2018.

[21] F. Ghaffar, "Facial Emotions Recognition Using Convolutional Neural Net," *ArXiv Preprint ArXiv:2001.01456,* 2020.

[22] P. Tarnowski, M. Kolodziej, A. Majkowski, and R.J. Rak, "Emotion Recognition Using Facial Expressions," *Proceeding of International Conference on Computational Science,* pp. 1175-1184, 2017.

[23] L. Krithika and L.P. GG, "Student Emotion Recognition System (SERS) for E-Learning Improvement based on Learner Concentration Metric," *Procedia Computer Science,* Vol. 85, pp. 767-776, 2016.

[24] L. Linnenbrink-Garcia and R. Pekrun, "Students' Emotions and Academic Engagement: Introduction to the Special Issue," *Contemporary Educational Psychology,* Vol. 36, No. 1, pp. 1-3, 2011.

[25] M.A.A. Dewan, M. Murshed, and F. Lin, "Engagement Detection in Online Learning: A Review," *Smart Learning Environments,* Vol. 6, No. 1, pp. 1, 2019.

[26] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction," *Proceedings of the ACM International Conference on Multimodal Interaction,* pp. 653-656, 2018.

[27] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Proceeding of International Conference on Neural Information Processing,* pp. 117-124, 2013.

[28] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese Female Facial Expression (JAFFE) Database," *Pro-*

ceedings of International Conference on Automatic Face and Gesture Recognition, pp. 14-16, 1998.

[29] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The Karolinska Directed Emotional Faces: A Validation Study," *Cognition and Emotion,* Vol. 22, No. 6, pp. 1094-1118, 2008.

[30] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-Millisecond Neural Face Detection on Mobile Gpus," *ArXiv Preprint ArXiv:1907.05047*, 2019.

[31] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv Preprint ArXiv:1704.04861*, 2017.

[32] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for Object Detection," *Advances in Neural Information Processing Systems,* Vol. 26, pp. 2553-2561, 2013.

[33] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 511-518, 2001.

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters,* Vol. 23, No. 10, pp. 1499-1503, 2016.

[35] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 886-893, 2005.

[36] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *ArXiv Preprint ArXiv:1409.1556*, 2014.

[37] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *ArXiv Preprint ArXiv:1412.3555*, 2014.

[38] T.N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4580-4584, 2015.

[39] Emotion Classification, (2018). https://github.com/XiaoYee/emotion_classification/commits?author=XiaoYee (accessed April 12, 2020).

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 770-778, 2016.

[41] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks," *Proceedings of the ACM International Conference on Multimodal Interaction,* pp. 445-450, 2016.

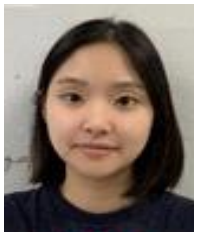
[42] Y. Zhai, and D. He, "Video-Based Face Recognition Based on Deep Convolutional Neural Network," *Proceedings of the International Conference on Image, Video and Signal Processing,* pp. 23-27, 2019.

[43] Z. Zhang, C. Wang, and Y. Wang, "Video-Based Face Recognition: State of the Art," *Proceeding of Chinese Conference on Biometric Recognition,* pp. 1-9, 2011.

[44] H. Wang, Y. Wang, and Y. Cao, "Video-based Face Recognition: A Survey," *World Academy of Science, Engineering and Technology,* Vol. 60, pp. 293-302, 2009.

[45] C. Shan, "Face Recognition and Retrieval in Video," *Video Search and Mining*, Vol. 287, pp. 235-260, 2010.

[46] Y.H. Jung, Y.M. Song, and Y.H. Ko, "Inclined Face Detection Using JointBoost Algorithm," *Journal of Korea Multimedia Society*, Vol. 15, No. 5, pp. 606-614, 2012.



이 자 연

2017년～현재 이화여자대학교 휴먼기계바이오공학부 학사 재학 중
관심분야 : 컴퓨터비전, 의료 인공지능, 딥러닝

이 은 혜

2017년～현재 이화여자대학교 휴먼기계바이오공학부 학사 재학 중
관심분야 : 딥러닝, 컴퓨터비전, 의료 인공지능, 의료용 로봇

정 소 현

2017년～현재 이화여자대학교 휴먼기계바이오공학부 학사 재학 중
관심분야 : 컴퓨터비전, 의료 인공지능, 딥러닝

하 유 빈

2017년～현재 이화여자대학교 휴먼기계바이오공학부 학사 재학 중
관심분야 : 컴퓨터비전, 딥러닝, 바이오센서

신 유 원

2017년～현재 이화여자대학교 휴먼기계바이오공학부 학사 재학 중
관심분야 : 데이터베이스, HCI, 인공지능, 컴퓨터비전

최 장 환

2015년 Stanford University, Mechanical Engineering, 공학 박사
2015년 Stanford University, Radiological Sciences Lab., Post-Doc.
2016년 한국전자통신연구원 선임연구원
2017년～현재 이화여자대학교 휴먼기계바이오공학부 조교수
관심분야 : 컴퓨터비전, 딥러닝, Bioinformatics, 의료 빅데이터