Haar Cascade와 DNN 기반의 실시간 얼굴 표정 및 음성 감정 분석기 구현

유찬영[O], 서덕규[*], 정유철(교신저자)[*]
[O]금오공과대학교 컴퓨터공학과,
[*]금오공과대학교 컴퓨터공학과
e-mail: 20181397@kumoh.ac.kr[O], ejrrb419@kumoh.ac.kr[*], jyc@kumoh.ac.kr[*]

Implementation of Real Time Facial Expression and Speech Emotion Analyzer based on Haar Cascade and DNN

Chan-Young Yu[O], Duck-Kyu Seo[*], Yuchul Jung[*]
[O]Dept. of Computer Engineering, Kumoh National Institute of Technology,
[*]Dept. of Computer Engineering, Kumoh National Institute of Technology

● 요 약 ●

본 논문에서는 인간의 표정과 목소리를 기반으로 한 감정 분석기를 제안한다. 제안하는 분석기들은 수많은 인간의 표정 중 뚜렷한 특징을 가진 표정 7가지를 별도의 클래스로 구성하며, DNN 모델을 수정하여 사용하였다. 또한, 음성 데이터는 학습 데이터 증식을 위한 Data Augmentation을 하였으며, 학습 도중 과적합을 방지하기 위해 콜백 함수를 사용하여 가장 최적의 성능에 도달했을 때, Early-stop 되도록 설정했다. 제안하는 표정 감정 분석 모델의 학습 결과는 val loss값이 0.94, val accuracy 값은 0.66이고, 음성 감정 분석 모델의 학습 결과는 val loss 결과값이 0.89, val accuracy 값은 0.65로, OpenCV 라이브러리를 사용한 모델 테스트는 안정적인 결과를 도출하였다.

키워드: 인공지능, 객체탐지, 얼굴 인식, 음성 인식, 딥러닝, DNN, Haar Cascade

I. Introduction

최근 사람에게 도움되는 인공지능 연구가 활발하게 이루어지고 있다. 이는 자동차, 스마트폰 등 다양한 플랫폼에 인공지능 기술을 융합하여 사람에게 안전성 및 편의성을 제공해 준다. 특히, 사람의 표정과 목소리를 이용하여 어떤 감정인지를 자동으로 인식해주는 기술은 차량 네비게이션 시스템, 스마트폰 등에 활용될 수 있다[1].

따라서, 제안하는 분석기는 사람의 감정을 표현하는 얼굴 표정을 자동으로 인식하여 그 사람의 감정을 분석하고 판단하는 것과 음성을 캡처하여 그 사람의 감정을 분석하고 판단하는 동작을 한다. 제안하는 분석기들은 파이썬을 사용하여 프로그래밍 되고, 케라스 환경에서 동작한다.

II. Related Work

일반적으로, 얼굴 표정 인식은 첫 번째로 얼굴 영역 검출, 두 번째로 얼굴 특징 추출, 마지막으로 표정 분류의 세 단계로 구분된다[2]. 얼굴 영역 검출은 2001년 P. Viola와 M. Jones가 제안한 Haar 기반 다단계 분류기 모델을 사용하여 자동으로 얼굴 영역을 검출하였다[3]. 검출된 얼굴 영역에서 얼굴 특징을 추출하는 방법은 외형 기반 방법(Appearance- based Methods)과 각각의 표정들의 동작 단위를 통해 얼굴의 움직임을 파악하는 방법으로 분류된다[4]. 다음으로, 앞서 추출한 얼굴 특징을 사용하여 표정을 분류한다. 표정 분류에는 일반적으로 KNN (K- Nearest Neighbor), SVM(Support Vector Machine) 등이 사용된다. 또한, 최근에는 대량의 학습 데이터를 이용하여 DNN(Deep Neural Network), CNN(Convolution Neural Network)등과 같은 딥러닝 기반의 얼굴 분류 알고리즘이 연구되었다[5][6].

음성 감정 인식은 음성 데이터 로드, 음성 데이터 추가, 음성 특징 추출, 모델 학습, 감정 분류의 순서대로 진행한다[7]. 음성 데이터의 전처리에는 음성 처리 라이브러리인 librosa[8]가 많이 사용된다. 음성 데이터 확장(Augmentation)을 위해서는 기존에 제안된 음성 데이터셋에서 잡음 추가, 파장 늘이기, 파형이동, 진폭 증가 등이 사용되었다. 음성 특징 추출에서는 librosa의 zero_crossing_rate,

Chroma_stft, mfcc, rms, 그리고 melspectrum을 사용할 수 있다. 모델 학습에는 기존에 제안된 CNN이나 ResNet-18, ResNet-20, 그리고 TC-ResNet 등이 사용될 수 있다.

III. The Proposed Scheme

1. 데이터 셋

본 논문에서는 얼굴 감정 학습을 위한 데이터셋으로 FER 2013(Facial Emotion Recognition 2013)을 사용하였다[9]. 해당 데이터셋은 사람의 뚜렷한 특징이 나타나는 화남, 경멸, 두려움, 행복, 슬픔, 놀람, 무표정의 7가지 감정 상태를 클래스로 분류해놓은 것으로, 총 28,821장의 얼굴 데이터로 구성되어있다. 또한, 음성 감정 학습을 위한 데이터셋으로 Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D), Toronto emotional speech set (TESS), Surrey Audio-Visual Expressed Emotion (SAVEE), 그리고 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)의 일부 파일을 사용한다.

2. 제안하는 얼굴 감정 분석기

제안하는 분석기는 사람의 7가지 감정 데이터셋인 FER 2013 데이터를 이용하여 모델을 학습시키고, 학습된 모델을 바탕으로 입력된 영상을 Python의 OpenCV 라이브러리를 이용하여 사람의 얼굴을 실시간으로 캡쳐하여 감정을 분석한다.

2.1 얼굴 검출

입력된 영상에서 정확한 얼굴 위치를 검출하기 위해 P. Viola와 M. Jones가 제안한 에이다부스트(Adaboost) 기반의 얼굴 검출 알고리즘인 Haar Cascade를 사용하였다. 이는 Positive Image (얼굴이 있는 이미지)와 Negative Image(얼굴이 없는 이미지)를 최대한 많이 사용한다. Haar Cascade의 특징으로는 이미지를 스캔하며 위치를 이동시키는 인접한 직사각형들의 영역 내부에 있는 픽셀들의 밝기 합의 차이를 이용하여 얼굴 영역을 검출해낸다[10]. 검출된 얼굴 영역은 정사각형의 박스 형태로 나타난다.

2.2 학습에 사용된 모델

본 논문에서는, 2.1절에서 검출한 얼굴의 표정을 인식하기 위해 DNN 모델을 사용하여 학습하였다. DNN 모델은 입력층(Input Layer)과 출력층(Output Layer) 사이에 수많은 은닉층(Hidden Layer)들로 이루어진 인공신경망(Artificial Neural Network)이다. 본 논문에서는 48x48 크기의 얼굴 이미지를 사용했기에 이미지 크기를 줄이는 작업을 생략하였다.

2.3 모델 학습

그림 1은 FER 2013 데이터셋 분포도를 나타낸다. 이 데이터셋은 총 35,887장의 48x48 흑백 이미지로 구성되어있다. 또한, 학습 데이터 28,709개, 검증 데이터 3,589개, 그리고 테스트 데이터 3,589개로 이루어져 있다.

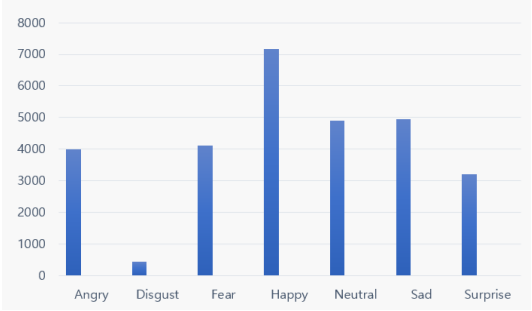


Fig. 1. Number of classes by facial expression in the FER 2013 datasets.

3. 제안하는 음성 감정 분석기

음성 감정 분석기 또한 7가지 감정 데이터를 사용한다. 모델을 학습시키고, 학습된 모델을 바탕으로 녹음 된 음성을 Python의 OpenCV 라이브러리를 이용하여 감정을 분석한다.

3.1 음성 데이터 입력

음성 데이터를 입력하는데 사용되는 데이터셋은 크게 CREMA-D, TESS, SAVEE, 그리고 RAVDESS의 일부 음성이다. CREMA-D는 특정 문장을 감정과 다양한 톤에 맞춰 말한다. TESS는 4명의 남녀가 감정을 담아 특정 단어를 말하고, SAVEE는 오직 감정만을 구분하여 30개의 문장을 말한다. 마지막으로, RAVDESS는 8개의 감정을 연설/노래, 감정 세기, 그리고 문장에 따라 구분하였는데, 본 논문에서는 통일성을 맞추기 위해 '고요함'에 해당하는 데이터셋은 제외하였다.

3.2 음성 데이터 증가(Data Augmentation)

본 논문에서는 음성 모델의 정확도를 증가시키기 위해 Augmentation을 사용한다. Augmentation에 사용된 기법으로는 librosa 라이브러리 함수인 잡음 추가(noise), 파장 늘이기(stretch), 파형이동(shift), 진폭 증가(pitch) 등을 사용하여 음성 학습 데이터를 추가한다.

3.3 음성 특징 추출

음성 데이터의 특징을 추출하여 배열로 표시하기 위해 본 논문에서는 librosa 라이브러리의 5개의 특징 : zero_crossing_rate(ZCR), Chroma_stft, mfcc, rms(Root Mean Square Value), 그리고 melspectrum을 추출한다.

3.4 모델 학습

그림 2는 음성 데이터셋 분포도를 나타낸다. 이 데이터셋은 총 11,970개의 WAV파일로 구성되어 있다. 또한, 학습 데이터 35,910개, 검증 데이터 26,932개, 그리고 테스트 데이터 8,978개로 이루어져

있다.

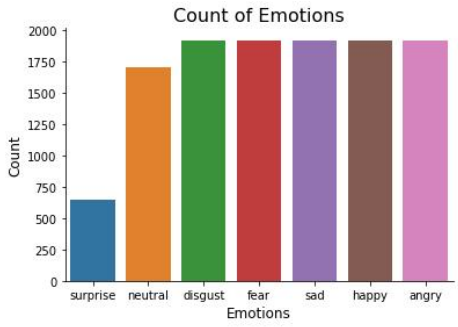


Fig. 2. Number of classes by speech emotion in datasets

IV. Experimental Results

그림 3과 4는 DNN 모델로 얼굴 감정 인식을 학습한 결과를 나타내며, Kaggle Notebook을 이용하여 학습하였다. Batch 사이즈는 32이며, 총 Epoch는 100을 할당하였다. 학습 결과 val loss는 0.9382, val accuracy로는 0.6897이 도출되었다.

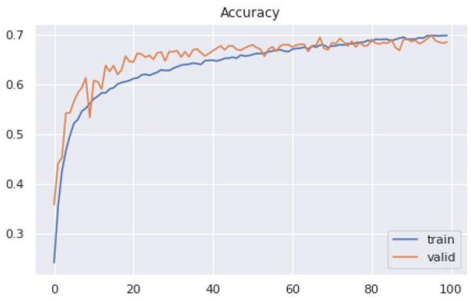


Fig. 3. DNN Model Learning Accuracy Rate

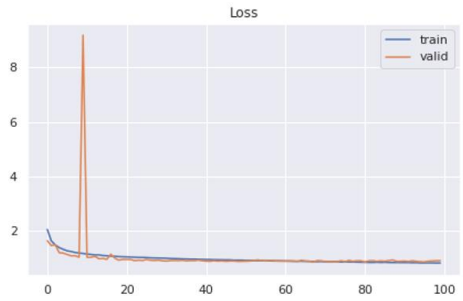


Fig. 4. DNN Model Learning Loss Rate

그림 5와 6은 CNN 모델로 음성 감정 인식을 학습한 결과를 나타내며, Kaggle Notebook을 이용하여 학습하였다. Batch 사이즈는 64이며, 총 Epoch는 146을 할당하였다. 학습 결과 val loss는 0.8859, val accuracy로는 0.6544가 도출되었다.

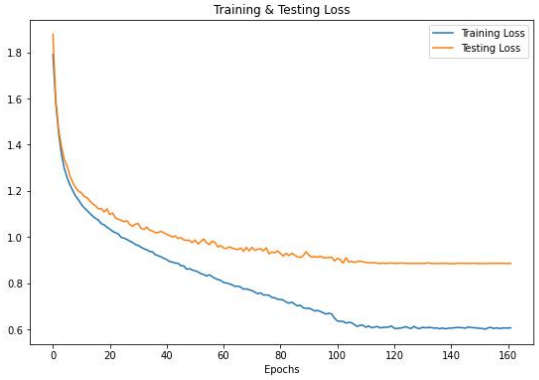


Fig. 5. CNN Model Learning Accuracy Rate

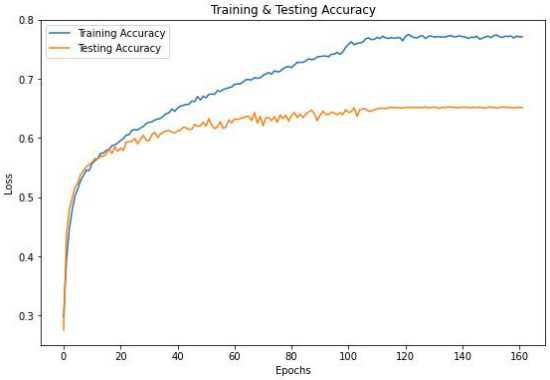
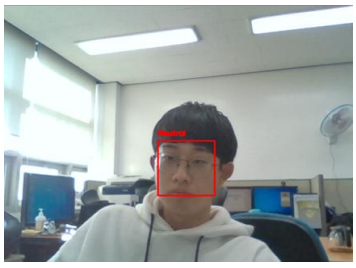
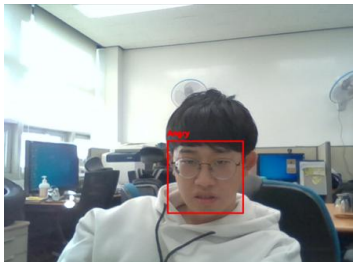


Fig. 6. CNN Model Learning Loss Rate

그림 7은 학습된 모델을 바탕으로, Python의 OpenCV 라이브러리를 이용하여 웹캠을 연결하여 실시간 기반의 얼굴인식 테스트를 하였다. 그림 8 또한 학습된 모델을 바탕으로 음성인식 테스트를 하였다.



(a) Neutral



(b) Angry

Fig. 7. Test Learning Models Using Webcam

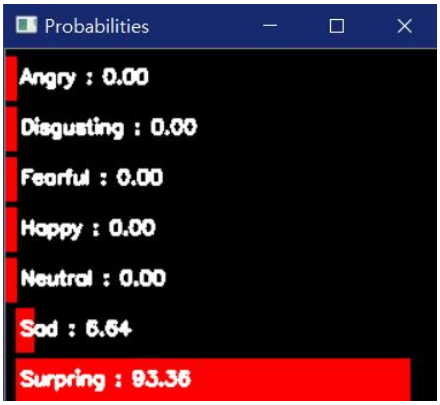


Fig. 8. Test Learning Models Using Librosa

V. Conclusions

본 논문에서는 사람의 수많은 감정 중 뚜렷한 특징을 가진 7가지 감정을 얼굴표정 및 목소리 톤을 이용하여 판별하는 감정 분석기를 제안하였다. 얼굴 표정을 학습하기 위한 데이터로는 Kaggle의 FER 2013 데이터셋을 이용하였으며, 목소리를 학습하기 위한 데이터로는 CREMA-D, TESS, SAVEE, 그리고 RAVDESS의 네 가지 데이터를 이용하였다. FER 2013 데이터를 사용하여 모델을 학습한 결과 각각의 클래스에 대한 충분한 데이터 확보 및 인식 오류 등 개선해야 할 부분이 많지만, 뚜렷한 특징을 가지고 있는 표정들에 대해서는 인식률이 만족할 만한 결과를 보였다.

추후 연구로는, 입력된 영상에서 다수의 사람들이 존재하는 경우, 그 사람들의 얼굴을 검출하여, 한 명만 인식하는 것이 아닌, 영상에 나오는 사람들 전체의 얼굴표정을 인식하여 감정을 분석할 수 있게끔 할 것이며, 정면의 얼굴만이 아닌 다양한 각도에서 바라본 얼굴 데이터들을 추가하여 분류 정확성을 높이고자 한다. 또한, 얼굴 표정과 음성 데이터를 합쳐 실시간으로 각 얼굴 표정에서 도출된 감정과 목소리에서 도출된 감정을 동시에 분석하여, 이 사람이 어떤 감정인지를 판별해주는 분석기를 설계하고자 한다.

REFERENCES

[1] 현대모비스, “자동차가 드라이버의 감정을 읽는다면?”, HMG JOURNAL, 2018.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," Pattern recognition, Vol. 36, No. 1, pp. 259-275, 2003.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," In Proceedings of the IEEE Computer Society Conference, Vol. 1, pp. 511-518, 2001.

[4] C. Shan, S. Gong and P. W McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study, " Image and Vision Computing, pp. 803-816, 2009.

[5] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 503-510, 2015.

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

[7] Speech Emotion Recognition Project, *https://www.kaggle.com/shivamburnwal/speech-emotion-recognition*

[8] librosa, *https://librosa.org/doc/latest/index.html*

[9] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," International Conference on Neural Information Processing, pp. 117-124, 2013.

[10] J. Whitehill and C. W. Omlin, "Haar features for faces au recognition," 7th International Conference on Automatic Face and Gesture Recognition, 2006.