YOLO 기반 표정 인식기를 활용한 내담자의 감정 분석 및 상담 효율성 판단

윤경섭*, 김민지O
*인하공업전문대학 컴퓨터정보과,
O인하공업전문대학 컴퓨터정보과
e-mail: ksyoon@inhatc.ac.kr*, kei07001@naver.comO

Analyzing the client's emotions and judging the effectiveness of counseling using a YOLO-based facial expression recognizer

Kyung Seob Yoon*, Minji KimO
*Dept. of Computer Science, Inha Technical College,
ODept. of Computer Science, Inha Technical College

● 요 약 ●

본 논문에서는 딥러닝 기술을 활용한 객체 검출(object detection) 모델인 YOLO를 기반으로 하는 감정에 따른 표정 인식 시스템을 활용하여 상담 시 보조 도구로 사용하는 방법을 제공한다. 또한, 머신러닝 기술 기반의 툴킷인 dlib 라이브러리를 사용하여 마스크 착용자의 눈 형태 관측을 통한 표정 인식 및 감정 분석의 정확도 상승을 도모하였다. 이 기술은 코로나19의 장기화로 온라인 수업이나 화상회의를 지원하는 플랫폼들이 전성기를 누리고 있는 현시점에서 다양한 분야로 확장할 수 있을 것으로 기대한다.

키워드: 딥러닝(deep learning), 객체 검출(object detection), 욜로(YOLO), 표정인식(facial recognition)

I. Introduction

내담자의 표정은 상담 진행에 중요한 역할을 한다. 상담자는 내담자의 표정을 통해 단순히 감정을 유추하는 것뿐만 아니라 어떤 질문을 해야 할지, 또는 상담 도중 주제를 변경하여야 할지 유지하여도 괜찮을지에 대해 결정하기도 하며 즉석에서 질문하기도 한다. 상담 내용을 내담자의 동의를 구한 후 녹음하여 축어록을 작성하는 경우는 종종 있지만 이러한 경우 내담자의 표정에 대한 세세한 기록은 어려우므로 감정에 따른 표정 변화에 대한 데이터가 있다면 상담의 효율성을 높이는 데에 큰 도움이 될 것이다.

본 논문에서는 딥러닝 기술을 활용한 객체 검출(object detection) 모델인 YOLO를 기반으로 하는 감정에 따른 표정 인식 시스템을 활용하여 상담 시 보조 도구로 사용하는 방법을 제공하고자 한다.

II. Preliminaries

1. Related works

1.1 딥러닝 기반 객체 검출 방식

딥러닝 기반의 객체 검출 방식은 그 절차에 따라 2가지로 분류할 수 있는데, 그 중 먼저 등장한 방식은 2-stage 객체 검출 방식이다. 2-stage 방식은 후보 영역 검출(region proposal) 단계를 거친 후 후보 영역에 존재하는 객체를 분류(region classification)하는 방식으로, 대표적으로 R-CNN, Fast R-CNN, Faster R-CNN 등의 모델이 있다[1]. 이후에 등장한 1-stage 방식은 후보 영역 검출 단계와 객체 분류 단계를 한 번에 수행하는 방식으로, 검출 속도는 빠르지만 정확도가 낮은 방식이다. 2-stage 방식 모델의 경우 후보 영역 검출 방식으로 바운딩 박스(bounding box)를 여러 개 그리면, 그 수만큼 객체 분류 과정을 수행하여야 하는 반면, 1-stage 방식은 회귀 분석 방식을 사용하여 입력된 이미지를 한 번에 인식하므로 상대적으로 속도가 매우 빠르다[2].

1.2 1-stage 방식의 객체 검출 모델 YOLO

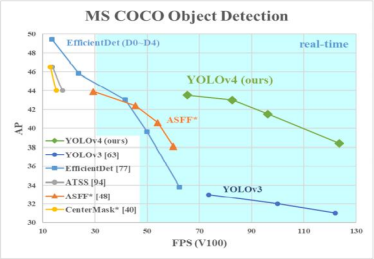


Fig. 1. Comparison of the proposed YOLOv4 and other state-of-the-art object detectors [3]

YOLO는 대표적인 1-stage 방식의 객체 검출 모델로, 현재 v1부터 v4까지의 논문이 공식적으로 게재되어 있다. 가장 최신 버전인 YOLO v4는 Fig. 1과 같이, 이전 버전인 YOLO v3과 비교했을 때 정밀도는 10%, 속도는 12%가 향상되었다. 비슷한 시기의 state-of-the-art 모델들과 비교하였을 때, 비슷한 정밀도에서 더 빠른 속도로 실행됨을 알 수 있다[3].

YOLO는 독립적인 객체 검출을 위한 여러 구성 요소들을 하나의 뉴럴 네트워크(Neural Network)에 통합한다. 또한, 전체를 한 번에 보는 형태를 가지고 있어 신경망의 입력 및 출력을 직접 고려하여 네트워크 가중치를 최적화하는 종단 간 학습(end-to-end training)이 가능하고, real-time에 준하는 속도와 높은 정밀도까지 유지한다[2]. 따라서 YOLO는 높은 처리 속도가 요구되는 동영상 처리 및 실시간 영상 처리에 적합한 모델이라고 할 수 있다[1].

2. Precautions

2.1 상담 영상 획득 시 주의사항

상담 영상을 분석하기 위해서는 내담자의 동의가 필요하다. 상담자는 한국 상담 심리 학회의 상담 심리사 윤리 규정에 따르면, 상담사는 영상을 녹화할 때 내담자에게 영상이 어떤 목적으로 촬영되는지를 명확하게 밝혀야 하고, '상담 종료 후 분석 과정을 거친 후 1시간 이내에 파기됨' 등 제한을 두고 사용하고 있음을 충분히 설명하여야 한다[4].

III. Design & Implementation

1. YOLO를 활용한 상담 영상 분석 및 시각화

1.1 상담 영상 분석 및 결과 집계와 시각화 프로세스

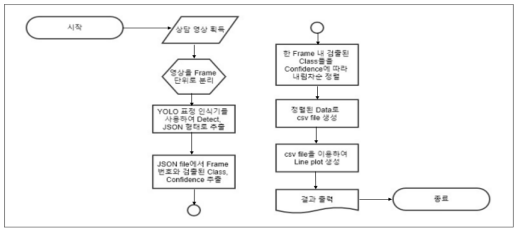


Fig. 2. Video analytics and results aggregation and visualization process

Fig. 2는 상담 영상을 분석한 후 그 결과를 집계하고 시각화하는 프로세스로, 상담 영상을 획득한 후부터 영상을 프레임 단위로 분석하고 그 결과를 실질적인 사용을 위해 가공하기까지의 과정을 담고 있다.

1.2 학습에 사용한 데이터셋 구축 과정

초기 학습 데이터셋은 구글의 FEC(Facial Expression Comparison) 데이터셋을 사용하였다. 이 데이터셋은 3개의 얼굴 이미지 중 유사한 표정의 두 이미지를 나타내며, 각 얼굴 이미지에 대해 이미지를 다운로드 받을 수 있는 URL과 이미지에서 얼굴이 존재하는 위치에 대한 바운딩 박스의 좌표를 제공한다[5]. 본 시스템에서는 학습을 위해 이미지에 존재하는 얼굴이 어떤 표정에 해당되는지에 대한 클래스 정보가 필요했으므로 직접 라벨링 하는 과정을 거쳐야 했다. 이 작업을 위해 YOLO v4 제작자의 Github에 업로드되어있는 YOLO 라벨링 툴인 Yolo Mark를 사용하였다[6].

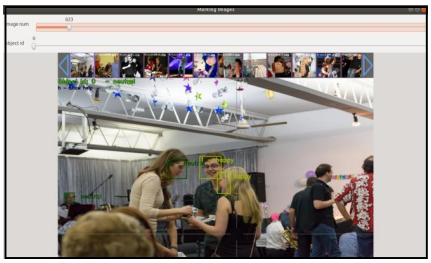


Fig. 3. Yolo Mark labeling results

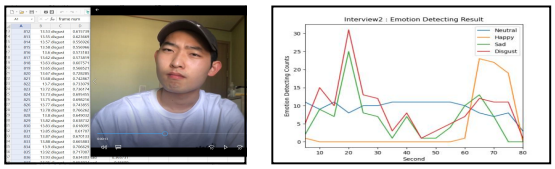
YOLO Mark는 이미지에서 오브젝트(object)가 존재하는 영역에 Fig. 3과 같이 바운딩 박스를 그리고 해당되는 클래스를 선택하면 클래스 번호와 바운딩 박스의 좌푯값, 바운딩 박스의 width, height 값이 기록된 텍스트 파일을 생성해준다.

FEC 데이터셋에서 제공하는 데이터 중 URL이 유효하지 않은 데이터를 제외하고 7779개의 이미지를 happy, sad, neutral, disgust의 4가지 감정 클래스로 구분하여 1000 iters, 2000 iters, 3000

iters 만큼 학습시킨 결과를 비교했을 때, Confidence, Accuracy, Recall 모두 2000 iters 만큼 학습시킨 경우 최적의 결과를 얻을 수 있었다.

1.3 상담 영상 분석 및 시각화

앞서 학습시킨 YOLO 모델을 사용하여, 상담 상황을 가정한 영상에 대해 Detection을 진행하였고, 그 결과를 JSON 형태로 저장하였다. 저장된 JSON 파일에서 프레임 번호와 검출된 클래스, 컨피던스를 추출하고, 한 프레임 내에 검출된 클래스들을 컨피던스를 기준으로 내림차순 정렬한 후 데이터의 실질적인 사용과 시각화를 위해 CSV 파일로 가공하였다.



(a) (b)

Fig. 4. Using and Visualizing object detection results

CSV 파일을 생성하는 과정에서 각 프레임이 실제 영상에서 어느 시점에 해당되는 지에 대한 컬럼을 추가하였으므로 시간대별로 기록된 Detection 결과를 Fig. 4의 (a)와 같이 영상 분석에 활용할 수 있다. 또한, Fig. 4의 (b)처럼 그래프를 생성하여 시간대별로 어떤 감정의 검출 횟수가 더 많은지를 시각화하여 나타낼 수도 있다.

2. dilb 라이브러리를 활용한 감정에 따른 눈 형태 변화 관측

2.1 YOLO를 활용한 상담 영상 분석의 한계점

FEC 데이터셋을 가지고 학습시킨 YOLO 모델은 마스크를 착용하지 않은 이미지의 경우에는 높은 정확도를 보였으나 마스크를 착용한 이미지가 입력되면 정확한 결과를 얻을 수 없었으므로 마스크 착용 이미지를 추가하여 새롭게 학습시키기로 결정하였다. 캐글의 Face Mask Detection 데이터셋[7]을 기존의 데이터셋에 추가하여 학습시켰으나, Detection 정확도가 크게 높아지지는 않았다. 따라서 마스크 착용 이미지를 추가하여 학습시키는 방식 대신, 마스크 착용 시에도 검출이 가능한 눈 형태의 변화에 대해 관측하는 방법을 고안하였다.

2.2 표정에 따른 눈 형태 변화에 대한 가설 설정

얼굴 인식 분야에서 가장 많이 사용되는 기술로는 OpenCV의 Harr Cascade 알고리즘과 dlib 라이브러리가 있는데, Harr Cascade 알고리즘은 비교적 작은 사이즈나 가변 사이즈에서 더 좋은 성능을 보여주고, dlib 라이브러리는 인식률과 사이즈가 일정 크기 이상이며 고정된 사이즈인 경우에 더 좋은 성능을 보여준다[8]. 내담자의 표정 분석과 같은 작업은 제한된 공간에서 고정된 크기의 오브젝트에 대한 검출을 진행하므로, dlib 라이브러리를 사용하는 것이 적합하다고 판단하였다. dlib 라이브러리는 이미지 처리를 비롯한 다양한 머신러닝 알고리즘을 활용할 수 있는 C++ 기반의 툴킷으로, HOG(Histogram Oriented Gradients)의 특성을 가지고 있는데, 이는 픽셀값의 변화로 파악할 수 있는 영상의 밝기 변화의 방향을 Gradient로 표현하여 객체의 형태를 찾아내는 방식이다[9].

눈 형태에 대한 분석을 진행하기 전, '무표정(neutral)인 경우와 비교하였을 때, 내담자의 표정이 부정적인 감정으로 분류되는 경우에는 눈꼬리가 처지고, 긍정적인 감정으로 분류되는 경우에는 무표정인 경우보다 눈꼬리가 올라가므로, 눈 앞머리와 눈꼬리가 각각 얼굴의 가로 선과 직교할 때의 거리를 이용하여 감정 클래스별 눈 모양의 특징을 찾을 수 있을 것'이라는 가설을 먼저 설정하였다.

2.3 표정에 따른 눈 형태 변화에 대한 가설 검증

Fig. 5. Creating landmarks using dlib

앞서 설정한 가정을 검증하기 위해 먼저 하나의 영상을 프레임 단위로 분리한 후 이전에 학습시킨 YOLO 모델을 사용하여 모든 프레임을 감정별로 분류하였다. 그 후 Fig. 5와 같이 dlib 라이브러리를 사용하여 얼굴 윤곽과 이목구비에 대한 랜드마크를 생성하였다. 이 과정에서 해상도가 낮은 영상의 경우 일부 프레임에서 실제 내담자의 얼굴과 관련이 없는 위치에 랜드마크가 생성되는 문제가 발생하였다. 이를 해결하기 위해 얼굴 포지션에 대한 정확도가 상대적으로 높은 YOLO Object Detection 결과를 이용하여, YOLO에서 예측한 바운딩 박스 내부에 랜드마크가 생성되었을 때만 얼굴 인식에 성공한 것으로 판단하였다.

Fig. 6. Set face horizontal and vertical lines

프레임 안에서 내담자의 얼굴 위치나 각도, 카메라와의 거리 등은 연속해서 변하므로, 눈의 형태를 바르게 분석하기 위해 각 프레임별로 생성된 랜드마크의 위치 정보를 통해 기준이 되는 가로, 세로 선을 설정하였다. Fig. 6과 같이, 가로 선은 얼굴 양 끝점을 연결한 선분으로 설정하였고, 세로 선은 코의 시작 지점과 턱 끝점을 연결한 선분으로 설정하였다.

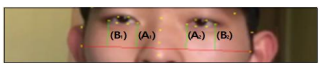


Fig. 7. Distance between landmarks

눈꼬리의 상승과 하강에 대한 판단은 Fig. 7에 나타나듯 얼굴 가로 선과 눈 앞머리의 랜드마크가 직교할 때의 거리를 A라고 하고 가로 선과 눈꼬리의 랜드마크가 직교할 때의 거리를 B라고 할 때, 두 값의 차인 B-A를 계산하여 진행하였다. 이는 눈꼬리가 상승하는 경우 B의 값이 증가하므로 B-A의 값이 커지고, 눈꼬리가 하강하는 경우 B의 값이 감소하므로 B-A의 값이 작아지기 때문이다. 이렇게 구한 값 또한 영상 속에서 내담자의 움직임에 따라 얼굴의 위치 정보나 각도 등이 변하면 다른 프레임에 대해 구한 값과 비교할 수 없으므로 얼굴 세로 선의 길이를 1이라고 했을 때 그 길이에 대한 B-A 값의 비를 구하였다.

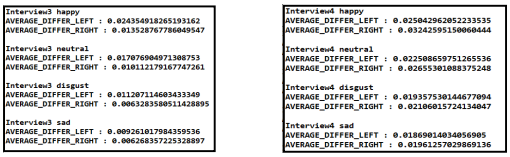


Fig. 8. Eye shape analysis result

Fig. 8은 각 프레임에 대한 계산 결과를 집계한 결과로, 4개의 클래스에 대한 평균 값이 기록되어 있다. 앞서 설정한 가설과 같이 무표정(neutral)으로 분류되는 프레임의 평균과 비교하였을 때, happy에 해당되는 경우 눈꼬리가 비교적 높은 곳에 위치하여 B-A의 값이 커졌고, disgust나 sad인 경우에는 눈꼬리가 비교적 낮은 곳에 위치하여 B-A의 값이 작아졌다.

dlib 라이브러리를 활용하여 감정에 따른 눈의 형태를 관측하는 실험 결과는 가설과 일치하여 neutral 클래스를 기준으로 부정적인 감정과 긍정적인 감정으로 분류하는 것은 가능하지만, disgust와 sad 클래스 간의 경계가 불분명하여 세부적인 감정 클래스로 분류하기에는 어려움이 있었다.

IV. Conclusions

본 논문에서 제시한 YOLO 기반의 표정 인식 시스템은 상담 시 보조 도구로써 상담의 효율성을 높이는 데에 도움을 줄 것이다. 또한, 코로나19의 장기화로 온라인 수업이나 화상 회의를 지원하는 플랫폼들이 전성기를 누리고 있는 현시점에서 영상 속 인물의 표정을 분류하는 기술은 응용할 수 있는 분야가 무궁무진할 것이다. 향후 연구에서는 dlib 라이브러리를 활용한 눈 형태 분석 방법으로도 완전히 극복하지 못한 마스크 착용자에 대한 Detection 정확도 문제를 보완하여 다양한 분야에 적용할 수 있도록 할 계획이다.

REFERENCES

[1] “Proposal for License Plate Recognition Using Synthetic Data and Vehicle Type Recognition System”, https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202029757728565&dbt=NART

[2] “You Only Look Once: Unified, Real-Time Object Detection”, https://arxiv.org/abs/1506.02640

[3] “YOLOv4: Optimal Speed and Accuracy of Object Detection”, https://arxiv.org/abs/2004.10934v1

[4] “Korean Counseling Psychological Association”, https://krcpa.or.kr/user/sub02_9.asp

[5] “Google facial expression comparison dataset”, https://research.google/tools/datasets/google-facial-expression/

[6] “AlexeyAB/Yolo_mark”, https://github.com/AlexeyAB/Yolo_mark

[7] “Face Mask Detection”, https://www.kaggle.com/andrewmvd/face-mask-detection?select=images

[8] “Design and Implementation of Visitor Access Control System using Deep learning Face Recognition”, https://www.koreascience.or.kr/article/JAKO202110650791289.pdf

[9] “Dlib C++ Library”, http://dlib.net/