Awesome-Youtube-Thumbnail-Generator

CUAI 4 기 모빌리티 B 팀

이나혁(소프트웨어), 장혜진(융합공학), 최원용(소프트웨어), 홍지호(기계공학)

[요약] 이 프로젝트의 목표는 사용자가 input으로 동영상을 넣으면 output으로 기준에 맞춰 썸네일을 제작해 주는 것이다. Instance Segmentation을 이용해 적절한 사람의 숫자를 뽑아내고, People Detection, Face Recognition을 이용해 적절한 사람의 수가 있는 프레임을 찾는다. 이후 사람의 크기가 가장 큰 프레임을 추천해주는 과정을 지니고 있다. 적절한 사람의 수가 있는 프레임의 대부분이 눈을 감은 사진, 혹은 카메라를 정면으로 응시하지 않는 프레임이었다. 사람의 크기가 가장 큰 프레임이 아니라, 눈이 가장 큰 프레임으로 기준을 바꾼다면 더 높은 확률로 원하는 프레임을 출력할 수 있을 것이다.

1. 서 론

디지털 광고 전문기업 인크로스가 2020 년 3 월 발표한 '동영상 플랫폼 이용 데이터'에 따르면 대표적인 온라인 동영상 플랫폼인 '유튜브'는 평균 체류시간에서 압도적인 1 위를 달리고 있으며, 이용률 또한 꾸준히 증가세를 보이고 있다.

이와 동시에 유튜버 시장에서의 생존 경쟁도 치열하다. 유튜브 영상에서 첫인상의 역할을 하는 것은 썸네일이다. 매력적인 썸네일은 시청자의 호기심을 불러일으키고, 채널의 콘텐츠 경쟁력을 높인다.

이에 본 참가팀은 머신러닝 기법을 적용하여 썸네일 이미지를 추천하는 모델을 설계하였다. 작동원리를 요약하자면 다음과 같다. MaskRCNN 기반 instance segmentation 을 통해 영상의 프레임 속 사람수를 추출하고 IQR(Interquartile range)으로 이상치를 제거한 뒤 프레임 면적과 사람의 얼굴 면적 비율을 통해 적절한 프레임을 선정하여 추천한다. 이후 OpenCV 의 Alpha Blending 을 통해 원하는 배경 이미지와 프레임 속 사람 이미지를 합쳐 하나의 썸네일 이미지를 생성한다.

2. 본 론

2.1 이론

1) Instance Segmentation with Mask R-CNN

썸네일 생성을 위해서 필요한 과정은 여러 단계가 필요하나 프로젝트 진행을 위해 가장 핵심적인 과정은 영상에서 인물 형상 추출하기라고 할 수 있다. 본 참가팀은 Mask R-CNN 기반의 instance segmentation 을 통해 썸네일에 삽입될 사람의 이미지를 추출하였다.

이미지 내에서 객체를 검출하는 것을 instance segmentation 이라 하고, semantic segmentation 과 달리 각각의 object 를 구분한다는 차이점이 있다. 또한 각각의 instance 를 해당하는 각각의 pixel 로 구분한다. 즉 객체를 검출하는 object detection 과 각 픽셀의 범주를 분류하는 semantic segmentation 이 결합된 task 이다.[1]

Mask R-CNN 은 Faster R-CNN 의 RPN 에서 얻은 RoI(Region of Interest)에 대하여 객체의 class 를 예측하는 classification branch, bbox regression 을 수행하는 bbox regression branch 와 평행으로 segmentation mask 를 예측하는 **mask branch** 를 추가한 구조를 가지고 있다. Mask branch 는 각각의 RoI 에 작은 크기의 FCN(Fully Convolutional Network)가 추가된 형태이다.[2]

먼저 영상을 다수의 프레임들로 분할한 후 Mask R-CNN 을 이용한 instance segmentation 으로 각 프레임당 들어있는 사람의 수를 추출하였다. 각 프레임이 썸네일로 활용되기에 적합한지 판단하는 가장 주된 기준으로 각 프레임에 몇 명의 사람이 들어가 있는지 확인하였다. 사람 수 데이터를 바탕으로 썸네일로 활용 불가능한 프레임들을 제거하는 과정이 이어진다.

2) IQR 을 이용한 outlier 제거

본 참가팀이 예시로 사용한 예능 프로그램 '유퀴즈 온 더 블럭'의 경우에는 대부분의 장면에서 진행자 2 명과 게스트 1~2 명을 합쳐 총 3~4 인이

출연하게 된다. 따라서 너무 많은 사람들이 들어간 프레임은 썸네일로 사용할 정도로 대표성을 띄기 힘들기 때문에 이러한 outlier 를 제거해 줄 필요성이 있다. outlier 를 탐지하는 방법은 매우 다양하지만, 그 중 가장 널리 사용되는 방식은 IQR Rule 이다. 본 참가팀은 outlier 를 제거하기 위한 기준으로 IQR 방식을 활용하였다.

IQR Rule 이란 이사분범위를 바탕으로 Q3 + 1.5*IQR 이상, Q1-1.5*IQR 이하의 값을 outlier 로 정의하는 것이다.[3]

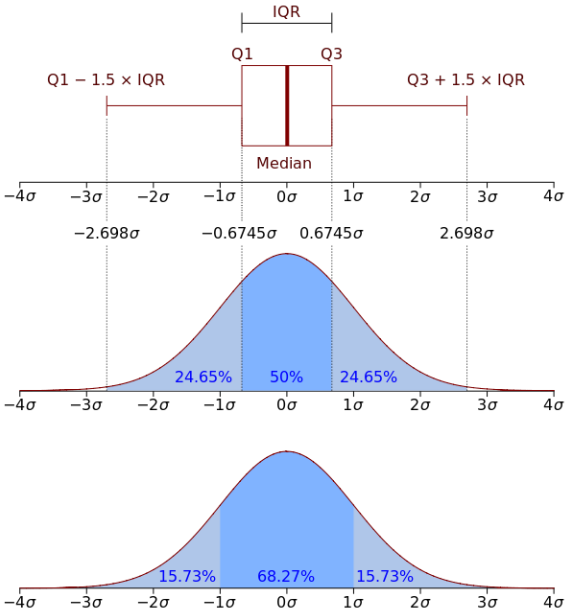

그림 1. IQR Box

3) People Detection / Face Recognition 을 사용한 프레임 추출

앞의 과정을 통해 각 프레임에 몇 명의 사람이 들어가야 하는지 판단했다면, 특정 프레임에 사람이 몇 명이 들어가 있는지 세어 보기 위해서 People Detection / Face Recognition 알고리즘을 사용하였다.

두 알고리즘 중 하나만 사용할 경우 사람의 얼굴만 나오거나 혹은 사람의 몸체만 나올 수 있기 때문에 people detection, face detection 을 모두 사용해 추천 프레임을 구하였다.

People Detection 은 Real-Time Object Detection 중 YOLO 라이브러리를 사용하였다. YOLO 라이브러리에서 사람이 인식되었을 때만 count 변수의 값을 증가하는 방식으로 People Detection 을 진행하였다.

face recognition 은 python 의 face_recognition 라이브러리를 사용하였다. face_recognition 라이브러리는 딥러닝으로 구축된 dlib 의 얼굴 인식 기능을 사용하여 구축되었으며 이미지 파일 혹은 영상 속 얼굴을 인식 및 조작 가능한 효율적인 라이브러리이다.

Instance Segmentation, People Detection, Face Recognition 이 인식한 사람의 수가 같은 프레임 중에서, 얼굴이 가장 크게 잡힌 프레임을 추천 프레임으로 결정했다.

4) Alpha blending 을 이용한 썸네일 배경 합성

Alpha blending 이란 **알파 블렌딩**이란 이미지 위에 또다른 이미지를 덮어 씌울 때 마치 투명하게 비치는 효과를 내기 위해 컴퓨터의 색상 표현 값 'RGB'에 'A'라는 새로운 값을 할당하여 배경 RGB 값과 그 위의 RGB 값을 혼합하는 표시하는 방법을 말하는데[4], 이 방법으로 최종적으로 선택된 frame 의 사람 mask 이외의 부분을 제거한 영역에 임의의 image file 을 이용해 thumbnail 의 배경을 합성하였다.

2.2 모델 구축 process

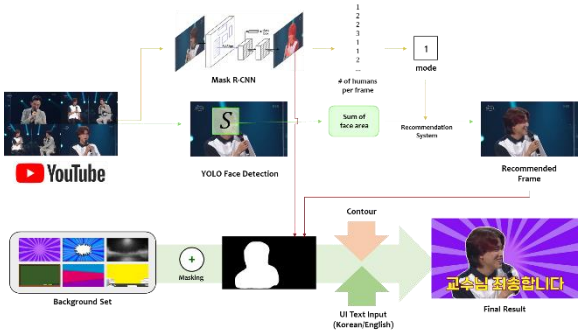

그림 2. the Overall Process of Thumbnail Generating Algorithm

위의 그림은 Youtube 영상으로부터 썸네일을 제작하기까지의 과정을 압축하여 보여준다.

위의 과정 첫 번째 단계에서, Instance Segmentation with Mask R-CNN 과 IQR 에 의한 outlier 가 제거된 최종적으로 추천 받은 프레임을 input 으로 준비하였다.

그림 3. Frame Example

두 번째로, Instance Segmentation Algorithm 을 이용하여 frame 내의 object(human)을 감지하고, 해당 영역의 mask 을 얻는다.

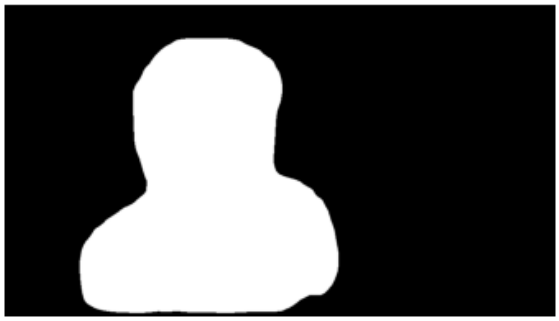

그림 4. Masking

세 번째로, 합성할 배경 이미지를 준비한다.

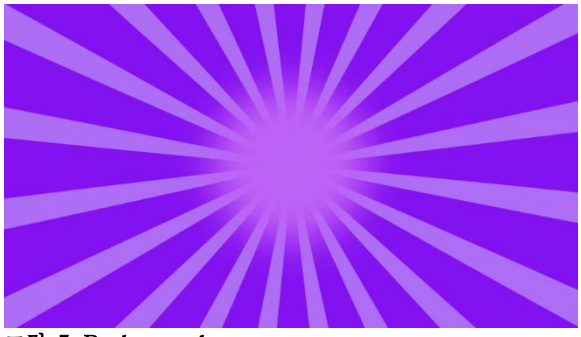

그림 5. Background

위의 이미지는 alpha blending 을 이용하여 배경을 합성하는 것에 사용된다. 이 이미지와 프레임을 합성한 뒤, 배경과 인물이 확실히 구분되는 느낌을 주기 위해 Segmentation 단계에서 얻은 mask 를 기반으로 Contour(윤곽선)을 그려낸다. 이 과정을 거친 후 모습은 다음 그림과 같다.

그림 6. 배경 합성이 완료된 프레임

마지막 단계로, 썸네일에 들어갈 자막을 합성해준다. 구상 단계에서 STT 를 이용한 subscript 추출로 자동 자막 추천 알고리즘을 설계할 계획이었으나, NLP 의 한계점으로 인하여 자막은 사용자가 직접 입력하는 방식으로 개발하였다. 사용자는 영상의 내용을 기반으로 적절한 자막을 구상하여 직접 삽입할 수 있다. 이를 바탕으로 완성한 최종적인 썸네일은 다음과 같다.

그림 7. 완성된 썸네일

3. 결 론

1) Output

위와 같은 과정을 통해 input 으로 '유퀴즈 온 더 블럭' 영상을 넣었을 때 다음과 같은 결과를 얻을 수 있었다.

2) 예상 활용 방안

본 연구는 Instance Segmentation 을 사용해 적정 인원수를 찾아내고, People Detection, Face Recognition 을 사용해 특정한 프레임을 추출하였다. 이후 Alpha Blending 으로 배경을 합성하여 자동으로 썸네일을 제작해주는 과정을 구현하였다. 위의 과정에서 사용된 여러 딥러닝 모델을 통해 많은 Youtuber 로 하여금 더욱 편리하게 Youtube 썸네일을 제작할 수 있는 중요한 역할을 할 것이라 예상된다.

3) 보완점

본 연구에서 적정한 인원수를 찾는 것에는 성공적인 결과를 얻었으나, output 으로 도출된 프레임의 사람이 눈을 감거나, 카메라를 정면으로 응시하지 않는 프레임이 추천되어 썸네일로 활용하기 어려운 프레임이 추출되는 문제가 발생하였다. 이를 보완하기 위해, 추천 기준을 Eye Detection 을 사용해 눈이 크게 나온 프레임으로

설정한다면, 앞 실험에서 나왔던 오류가 나올 가능성은 낮아질 것이다.

또한, 자막까지 자동으로 생성하는 완벽한 자동 생성 알고리즘을 구현하지 못하였기 때문에, 후속 연구에선 STT 와 KoNLPy(Korean NLP in Python)을 활용하여 자막을 자동으로 추천해주는 단계까지 설계할 예정이다.

4) Expectations

빠르게 증가하는 Youtube 이용자가 증가하는 요즘, Youtuber 가 시청자에게 양질의 컨텐츠를 제공할 수 있도록 하여 Youtube 이용 만족도를 높이고, 영상 제작에 필요한 비용과 시간을 효율적으로 관리하는 것에 기여할 수 있도록 본 연구가 활용되길 바란다.

참고 문헌

[1] Instance Segmentation with Mask R-CNN, https://towardsdatascience.com/instance-segmentation-with-mask-r-cnn-6e5c4132030b

[2] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, "Mask R-CNN", arXiv:1703.06870v3 [cs.CV] 24 Jan 2018

[3] Outlier 는 모두 제거해야 할까? — Outlier detection, IQR https://medium.com/@Aaron__Kim/outlier-%EB%AA%A8%EB%91%90-%EC%A0%9C%EA%B1%B0%ED%95%B4%EC%95%BC%ED%95%A0%EA%B9%8C-3aec52ef21b1

[4] Alpha compositing https://en.wikipedia.org/wiki/Alpha_compositing