프레임 누적과 UNET 네트워크를 이용한 동영상 내에서의 자막 영역 검출

김정환
서울대학교 컴퓨터공학부
jrheonghwan@naver.com

ABSTRACT

많은 동영상 내에서 사용자의 편의를 위해 자막이 제공된다. 영화 상에서는 자막을 통해 대사의 의미를 전달할 수 있고 예능 영상에서는 해당 장면의 분위기를 묘사해주거나 상황을 표현해 줄 수 있다. 이런 자막들을 추출해 낼 수 있다면 동영상에 대한 정보를 더욱 쉽게 얻을 수 있으며 이런 데이터를 이용하여 영상 내의 semantic analysis 학습 연구에 도움을 줄 수 있다. 현재 사진에서의 Text detection에 관한 연구는 진행이 많이 되어있는 상태이다. 동영상을 각 프레임으로 나누어 사진에서 위의 방법을 이용하여 자막 영역을 검출해낼 수 있지만 동영상 내에서 자막이 일정 시간은 유지된다는 특성을 이용하여 더 효과적으로 자막을 검출해내는 방법을 본 논문에서는 제안한다. 본 논문에서는 Text detection에서 좋은 능률을 보이는 UNET [12] 기반 네트워크를 사용함과 동시에 동영상의 시간적 정보를 이용하여 자막 영역을 검출할 것이다.

1 Introduction

최근 많은 사람들이 Youtube를 이용하면서 엄청나게 많은 양의 동영상 컨텐츠가 생성되고 있다. 물론 예전에도 많은 영상들이 만들어졌지만 최근에는 개인 제작자들이 컨텐츠를 많이 생성하기 시작하며 온라인 상의 동영상 컨텐츠 양이 대폭 증가하였다. 이런 동영상 내에서는 사용자의 편의를 위해 자막을 제공해준다. 이는 사용자들이 동영상 재생시에 음향을 듣지 않거나 외국어를 몰라도 내용을 이해하도록 도와주며 또한 영상 내에 상황을 설명해주기도 한다. 이처럼 자막은 동영상 내에서 영상의 정보를 제공해주는 데 큰 역할을 한다.
기존에 이런 자막을 검출하기 위해 색상이나 밝기 대비, 외곽선을 검출하는 방식등으로 연구가 진행되었다. [1] 이런 검출 방식는 영상 내에서의 시간적 특징을 이용하지 못하고 고정된 프레임을 이용하여 처리했다. 이를 개선하여 동영상의 시간적 특징을 이용하기 위해 프레임을 누적시켜 자막 영역을 찾아내는 방식도 진행되었다. [2]
본 논문에서는 기존 방식에서 사용되었던 시간적 특성을 활용한 자막 검출 방식에 더해 Text detection에서 좋은 성능을 보이고 있는 [3,4] UNET 기반 네트워크를 활용한 자막 영역 검출을 진행하고자 한다. 자막이 동영상 내에서 고정된 위치를 차지하고 있다는 정보를 이용하여 일차적으로 자막 가능 영역을 간추리고 이후에 UNET 기반 네트워크를 이용하여 이차적으로 영역을 탐지하였다.

2 Related Works

동영상 내에서의 자막 영역 탐지에 관한 연구는 고정된 하나의 프레임, 즉 사진에서 텍스트 영역을 탐지하는 연구와 동영상의 특징인 시간적 특성을 이용한 연구로 진행되었다. 그 중 사진에서 텍스트 영역을 탐지하는 연구는 딥러닝 모델을 활용한 Text detection 연구와 모폴로지 연산과 캐니 에지연산 [5] 등의 이미지 프로세싱 연구등이 있다.

2.1 Text detection

Text detection 과 관련된 연구는 OCR 연구의 일부로써 Text Recognition과 함께 많은 연구가 진행되었다. 맨 처음에는 SSD [9] 네트워크의 anchor를 텍스트에 맞게 수정해줌(Anchor-based methods)으로써 텍스트 영역을 바운딩 박스로 잡아내었다. [6] 그 후에 연구들에서는 기울어져있거나 휘어진 영역에 알맞게 영역을 잡아주거나

[3,4,7] 더 빠른 속도를 보여주는 모델 [8] 등 여러 방식으로 발전해왔다.

Figure 1: Text Detection 결과

2.1.1 Regression-based Text detection

기존의 YOLO [11] 모델이나 SSD [9] 모델을 통해 텍스트 영역 탐지를 진행하였으나 이러한 방식은 텍스트 디텍션에 잘 적용되지 못한 것이 anchor때문이었다. 일반적인 물체와 달리 텍스트는 가로로 혹은 세로로 엄청긴 비율을 가지게 되는데 이는 기존 모델에서 사람들이 정해준 anchor 비율에 맞지 않는 것이였다. 이에 대응하여 TextBoxes [6] 에서는 텍스트에 맞게 anchor 비율을 임의로 정해주어 문제를 해결하였다. 하지만 이러한 방식으로는 모든 텍스트 모양을 잡아주는데 한계가 있었다. Figure 1의 왼쪽 그림이 TextBoxes의 결과들 중 하나인데 사진에서 보면 알 수 있듯이 텍스트 탐지 영역이 특정한 모양으로만 잡힘을 알 수 있다. 이는 기존의 사용 모델들이 모두 바운딩 박스 형태로 물체를 디텍트하기 때문이다. 그 이후, 기울어진 바운딩 박스등을 그려주기 위해 여러 방식이 도입되었지만 이들도 엄청나게 발전된 성능을 보여주지는 못하였다.

2.1.2 Segmentation-based Text detection

앞서 말한 문제를 해결할 수 있는 방식이 pixel 단위로 영역을 잡아내는 것이다. TextSnake [7]나 TextCohesion [4] , CRAFT [3]등이 이와 같은 방식으로 영역을 잡아내었다. TextCohesion에서는 텍스트의 뼈대 부분과 각 주변 영역을 학습시켜 이를 통해 디텍트를 진행하였고 CRAFT에서는 Character 박스와 Affinity 박스 영역을 캐릭터 단위로 박스를 만들고 해당 영역을 gaussian value를 준 뒤 학습하고 디텍트를 진행하였다. 이처럼 TextCohesion이나 CRAFT의 경우 UNET 기반의 픽셀 단위 Segmentation을 진행함으로써 더욱 다양한 영태의 영역을 잡아내 줄 수 있었다. Figure 1의 오른쪽 그림이 CRAFT의 결과물들 중 하나인데 curved 된 글자들의 경우도 크게 잡지 않고 타이트하게 잡는 것을 확인할 수 있다.

2.2 Image Processing

자막 영역 검출을 위해 우리는 여러 방식의 이미지 처리를 할 수 있다. 모폴로지 연산과 캐니 에지 연산 등의 방식이 그 중 하나이다. 이런 연산등을 통하여 이미지 내 경계선을 찾아내어 우리가 원하는 영역을 검출해낸다. 모폴로지 연산에서는 팽창과 침식 연산을 이용하여 영상 내의 객체 영역을 넓혀가고 줄여가며 이를 통하여 외곽선 영역을 더 뚜렷하게 만들어낼 수 있다. 캐니 에지 디텍션 알고리즘은 크게 4단계로 나뉘게 된다. 가우시안 필터를 이용하여 노이즈를 제거한 이후 수평 수직 방향의 Gradient를 Sobel 커널을 이용하여 획득한다.

$$EdgeGradient(G) = \sqrt{G_x{}^2 + G_y{}^2}$$

$$Angle(\theta) = \tan^{-1}\frac{G_y}{G_x}$$

Figure 2: Canny Edge Detection 알고리즘에서 픽셀별 Edge Gradient 계산방법

이후, 이미지 전체를 스캔하면서 최대값의 gradient를 갖는 픽셀을 찾고 해당 영역이 실제 에지인지 판단하는 과정을 거쳐 에지를 판단해낸다. 이때, 픽셀별 gradient값을 계산하는 방법은 Figure 2에 나와 있는 식을 이용하여 계산한다.

3 자막 영역 검출 모델

본 논문에서 자막 후보군 검출에 대한 흐름은 figure 3와 같다.

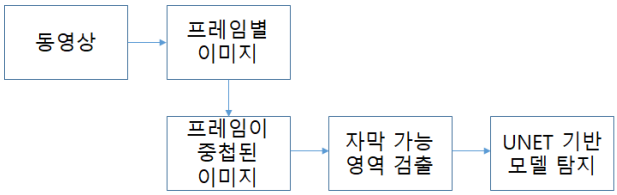

Figure 3: 프레임 누적과 UNET 기반 모델을 이용한 자막 검출 흐름

주어진 동영상을 동일한 시간 간격의 프레임으로 나눈 뒤, 1초간의 프레임들을 누적시켜 새로운 이미지를 만들

어낸다. 이렇게 여러 프레임들을 누적시킬 때, 영상 내에서 자막은 움직이지 않는다는 특성을 이용하여 해당 이미지에서 자막 가능한 영역을 우선적으로 검출해낸다. 그 후, 해당 영역들에 대해서 UNET 기반 모델을 이용하여 자막 영역을 검출해낸다.

3.1 프레임 누적을 통한 자막 영역 검출

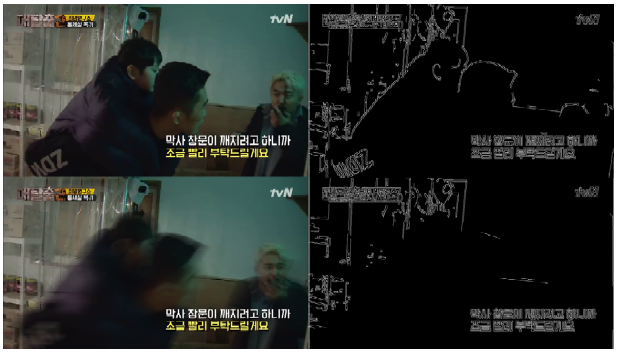

Figure 4: 한 프레임과 여러 프레임이 중첩된 이미지 원본과 캐니연산 모폴로지 연산 처리 이후 사진

기존 논문 [2] 에서는 프레임을 누적시킨 후 Morphology 연산 및 Canny Edge Detection을 통하여 자막 후보군을 줄여내고 영역을 찾아내었다. 본 논문에서는 자막의 지속시간은 최소 1초라는 가정하에 영상 내의 모든 프레임을 직후 1초간의 프레임을 누적시켜 새로운 영상을 만든 후 진행하였다. Figure 4에서 볼 수 있듯이 여러 프레임을 중첩시켜도 자막 영역의 값은 거의 변하지 않음을 확인할 수 있다. 이를 통해 여러 프레임을 중첩한 이미지를 이용하여 자막 후보군을 1차적으로 간추릴 수 있다. Figure 4에서 오른쪽 두 이미지는 원본 이미지와 중첩된 이미지를 각각 캐니 에지 디텍션과 모폴로지 연산을 통해서 경계선을 얻어낸 이미지들이다. 해당 이미지들에서 확인해볼 수 있듯이 여러 프레임이 중첩된 이미지에서는 고정되지 않은 사람의 움직임등은 경계선으로 찾아내지 못하고 여러 프레임에 걸쳐 값이 거의 그대로인 자막영역이나 배경 영역의 픽셀에서만 경계선을 잡아내는 것을 볼 수 있다. 본 논문에서도 이러한 동영상의 시간적 특성을 이용하여 1차적인 자막 후보군을 만들어낸다. 다만 기존 논문에서는 모폴로지 연산과 캐니 에지 연산을 통해 자막 영역을 검출했다면 본 논문에서는 여러 프레임을 중첩시켜 값이 오차 범위 내에서 변하는 픽셀들의 영역을 잡아내고 해당 영역을 1차적으로 자막 영역으로 잡아낼 것이다.

3.2 Unet 기반 텍스트 영역 검출 모델

3.2.1 구조

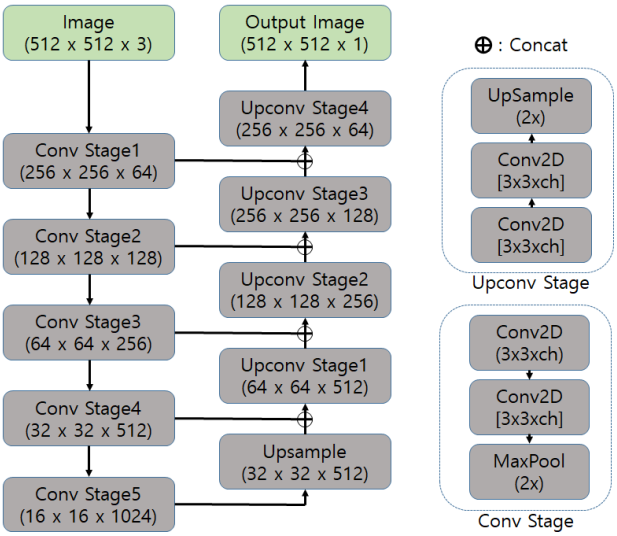

Figure 5: 모델 구조

본 논문에서 사용하는 모델의 구조는 Figure 5와 같다. 512x512x3 이미지를 받고 각 픽셀별로 텍스트 영역인지 알려주는 512x512x1 맵을 아웃풋으로 갖는 모델이다. Conv Stage에서 활성화 함수로는 relu가 사용되었고 각 스테이지별로 max pooling이 사용되어 이미지가 절반 사이즈로 down sampling되게 된다. Upconv Stage에서는 Conv Stage와 유사하게 진행되며 마지막에 max pooling 대신 up conv하여 이미지 크기를 2배로 만든다. 처음 이미지를 받고 Conv Stage를 거치며 이미지 크기를 점점 줄여나가고 이후 다시 Upconv Stage를 거치며 이미지를 크기를 키워 나가는데 이 때 Upconv Stage에서 이전 Conv Stage의 output을 concatenate 해줌으로써 context와 localization 모두 챙길 수 있다. 모델 구조에서 보면 알 수 있듯이 Fully Conected Layer가 없어 속도 측면에서 빠르다는 것을 확인할 수 있다.

3.2.2 학습

Figure 6: Totaltext와 예능에서 image와 groundtruth

학습은 원본 이미지와 해당 이미지에서 자막 영역을 표시해준 이미지 두 이미지로 이루어진다. Total text [10]에서 제공해주는 사진 1200장 가량에 더해 직접 예능에서 마스킹한 데이터 1000장 가량, 총 2200장 가량을 학습 데이터로 사용하였다. 2200가량의 사진이 Figure 6와 같이 원본 이미지와 각각 사진에 대응하는 groundtruth 이미지를 학습하는 데 사용하였다. 학습일 시킬때 loss function으로는 binary cross-entropy loss function을 사용하였다. 학습을 진행하면서 분류하는 class는 text영역과 나머지 영역으로만 나누었고, 이미지의 원본과 groundtruth 이미지들은 본 논문의 모델에 맞게 512x512로 바꿔주어 사용하였다.

4 실험 및 결과

Figure 7: 누적 프레임에서 달라진 영역 표시

영상 내에서 자막 영역은 프레임이 지나가도 rgb 값이 바뀌지 않을 것이라는 가정으로 우리가 비교하고자 하는 프레임과 그 뒤 10개 프레임을 각각 비교하여 달라진 영역에는 1의 값을 바뀌지 않은 영역에는 0의 값을 주었다. 그러나 생각과는 다르게 영상 내에서 같은 자막일지라도 프레임이 지나면서 값이 바뀌었다. 따라서 정확히 일치하는 값을 확인하기 보다는 r, g, b 각각 20의 오차 범위안에 들어가면 값이 바뀌지 않는다는 조건으로 범위 내에서 값이 바뀌었으면 0, 그 이상으로 바뀌었으면 1의 값을 결과로 준 결과물이 Figure 7과 같다. 물론 이전에 비해 더 나아진 결과물이지만 아직까지도 자막 영역의 값이 바뀐다고 표현되었다. 따라서 이런 부분을 없애주기 위해 전체 화면을 몇개의 패치로 나누어주고 각 패치에서 값이 어느 범위 안에 수준으로 바뀌지 않은 부분이 있는 경우 해당 패치 전부를 자막 가능 영역으로 분류해냈다. 하지만 이런 방식을 취한 경우 영상 내에서 화면의 이동이 많으면서 자막이 있는 경우를 제외한 나머지 모든 경우에서 거의 모든 영역을 자막 가능 영역으로 1차 분류하여 목적한 대로 결과가 나오지 않았다.

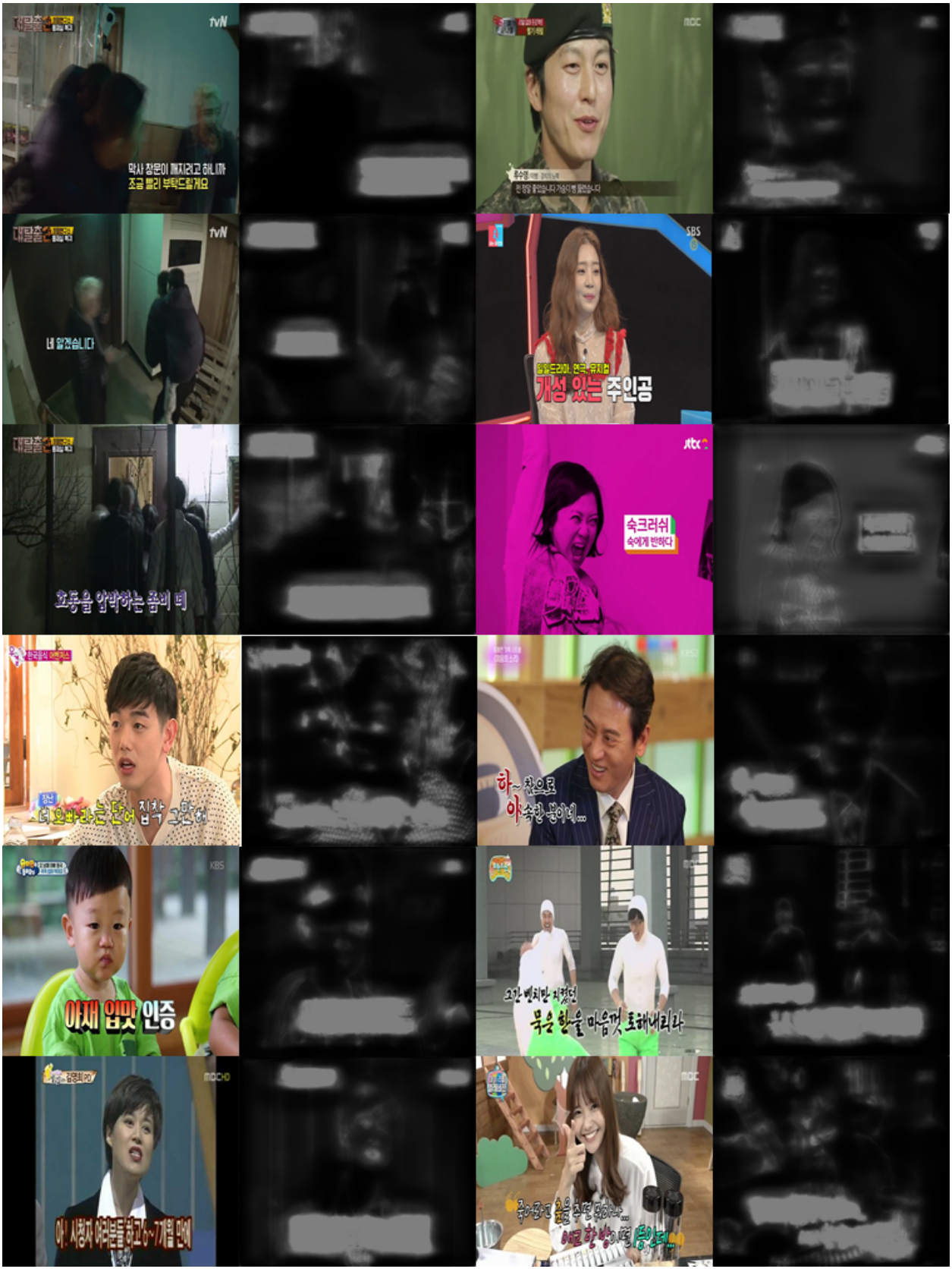

Figure 8: 자막 영역 검출 결과

본 논문에서 제안한 방식으로 자막에 대한 검출을 실시한 결과는 Figure 8과 같다. 위의 사진에서 왼쪽 사진이 디텍트를 하기위한 인풋 데이터이고 오른쪽 사진들이 그에 대한 자막을 찾은 영역이다. 결과에서 보면 알 수 있듯이 자막과 예능 로고영역등을 잘 탐지한 것을 확인할 수 있다. 아직까지 동영상에서 자막검출에 대한 evaluate dataset이 주어진게 없어 이 모델에 대한 정확한 성능 평가는 어려웠다. 예능 프로그램 하나에 대해 100개의 장면에서 자막 영역을 잡는지 확인해보았다. 이 경우 영상 내 100개의 장면에서 사람이 센 자막 영역의 수는 358개였고 그중 본 논문에서 제시한 모델이 잡아낸 자막 영역의 수는 316개로 높은 Recall을 보여주었다. 그러나 테스트 해본 영상 데이터셋의 수가 적다는 점, 그리고 본 논문에서 제시한 모델의 특성상 픽셀 단위로 값을 표현해주어 자막을 잡는 영역의 범위가 매우 넓어지거나 일부만 잡아내는 경우 등 정확하게 잡아내지 못한다는 문제점이 있다.

5 결론 및 향후 연구

본 논문에서는 동영상의 시간적 특성을 이용하여 자막의 영역을 찾아내려고 하는 동시에 Text detection 네트워크를 활용하여 영상 내에서 자막 영역을 검출하였다. 본 논문에서는 사진에서 정확도가 매우 높은 Text detection 네트워크를 동영상에 적용시킴으로써 높은 확률로 자막 영역을 검출할 수 있어 추후 Text Recognition 네트워크와 함께 사용한다면 자막을 이용한 동영상 분석등에 사용할 수 있을 것이라 생각된다. 본 논문에서는 Text detection 네트워크로 단순히 unet을 변형하여 쓰고 있지만 현재 Text detection에서 좋은 성능을 보이고 있는 CRAFT등의 네트워크를 연결한다면 더욱 좋은 결과를 낼 수 있을 것이라 생각된다.

References

[1] 김원준, 김창익 "해리스 코너 검출기를 이용한 비디오 자막 영역 추출," *정보과학회논문지 : 소프트웨어 및 응용*, pages 646–654. 2007.7

[2] 신광수, 남종호 "동영상 프레임의 시간적 누적을 이용한 효과적인 자막 영역 검출 방법" *한국정보과학회 2016년 동계학술대회 논문집*, pages 1313–1315. 2016.12

[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, Hwalsuk Lee, Clova AI Research, NAVER Corp "Character Region Awareness for Text Detection" *CVPR*, 2019.04

[4] Weijia Wu, Jici Xing, Hong Zhou "TextCohesion: Detecting Text for Arbitrary Shapes" 2019.04

[5] Ding, Lijun, and Ardeshir Goshtasby "On the Canny Edge Detector" *Pattern Recognition*, Vol. 34, No. 3, 2001

[6] Laio, Minghui and Shi, Baoguang and Bai, Xiang and Wang, Xinggang and Liu, Wenyu "TextBoxes: A Fast Detector with a Single Deep Neural Network" *AAAI*, 2017

[7] Long, Shangbang and Ruan, Jiaqiang and Zhang, Wenjie and He, Xin and Wu, Wenhao and Yao, Cong "TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes." *ECCV*, 2018

[8] Zhou, Xinyu and Yao, Cong and Wen, He and Wang, Yuzhi and Zhou, Shuchang and He, Weiran and Liang, Jiajun "EAST: An Efficient and Accurate Scene Text Detector" *CVPR*, 2017

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg "SSD: Single Shot MultiBox Detector" 2015

[10] C. K. Ch'ng and C. S. Chan "Total-text: A comprehensive dataset for scene Text detection and recognition" *ICDAR*, Vol. 1, pages 935–942. 2017

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection" *CVPR*, pages 30–43. 2016

[12] Olaf Ronneberger, Philipp Fischer, Thomas Brox "U-Net: Convolutional Networks for Biomedical Image Segmentation" 2015