딥러닝 객체 인식을 활용한 문서 내 표 텍스트 정보 추출

이혜민*, 김정수**, 문현석***
*한국건설기술연구원 UST KICT School 건설환경공학과
**한국건설기술연구원 구조연구본부
***한국건설기술연구원 미래스마트건설연구본부
e-mail: hyemin@kict.re.kr

Extraction of Table Text Information in Documents Using Deep Learning Object Detection

Hyemin Lee*, Jeongsoo Kim**, Hyounseok Moon***
*Dept of Civil & Environmental Engineering, UST KICT School
**Dept of Structural Engineering Research, KICT
***Dept of Future & Smart Construction Research, KICT

요 약

본 논문에서는 딥러닝 객체 인식을 기반으로 한글 문서 내 표 및 박스의 텍스트 정보를 추출하고 이를 재배열하여 문장화하는 방안을 제시하였다. 객체 감지 모델 Yolo를 통해 추출한 문서 내 표 및 박스의 좌표 정보를 토대로 OpenCV와 Google Cloud Platform의 Vision API를 적용하여 문서로부터 표 및 박스 이미지를 분리하고 텍스트를 감지하였다. 추출한 표 및 박스 내 텍스트를 문장화하기 위한 규칙 기반의 텍스트 재배열 방법을 제안하였으며 이를 통해 표의 텍스트 정보가 원래 의도된 문장으로 도출되는지 확인하였다.

1. 서론

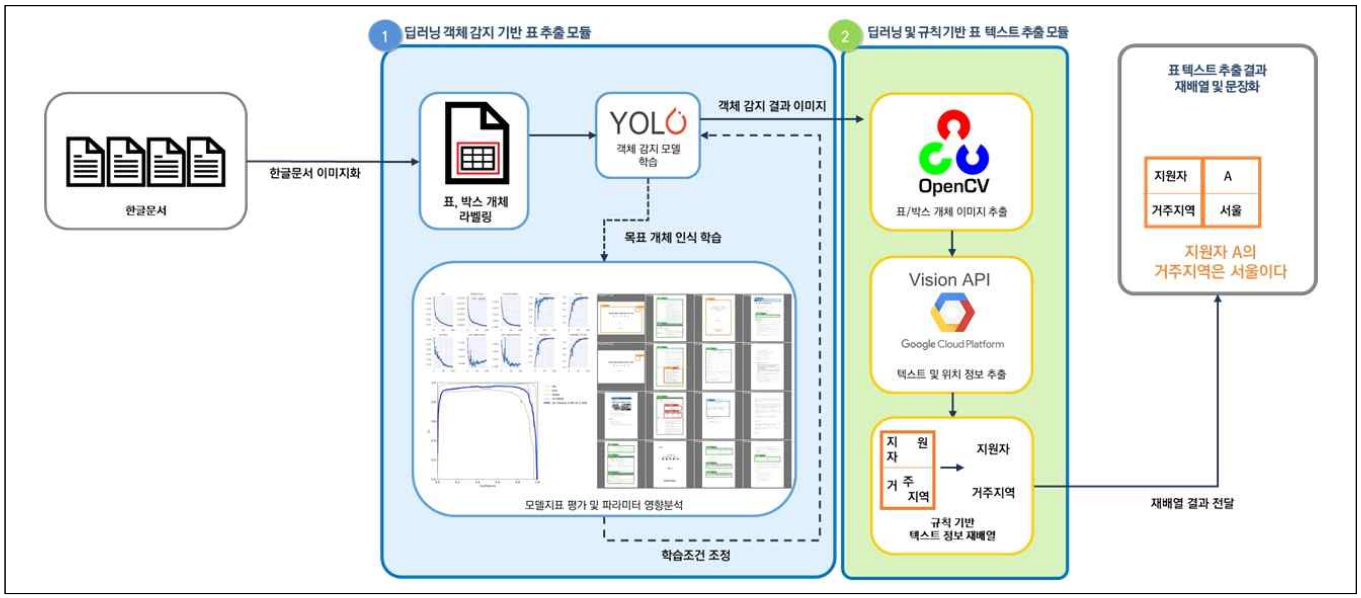
문서 내에는 본문의 텍스트 정보뿐만 아니라 다양한 형태의 그림이나 표 등의 자료를 포함하고 있는 경우가 많다[1]. 이러한 자료 중 표 내부의 텍스트를 항목 단위로 인식하는 것은 문서 처리 자동화에서 중요한 부분이다[2]. OCR의 활용으로 문서 내 텍스트 정보 인식은 가능하지만, 텍스트의 맥락적 의미를 제시하는 것은 아니다[3]. 따라서 문서 내 표에 담겨있는 정보들을 효과적으로 활용하기 위해서는 본문 내용과 표 부분의 내용을 구분하고 맥락적 텍스트 정보를 추출하는 과정이 중요하다.

최근 문서 분석과 관련하여 문서 이미지 내 표를 검출하거나 표 문자를 인식하는 몇몇 선행 연구들이 수행되고 있다. 관련 내용으로는 문서 내 표 항목 문자 인식률을 높이기 위해 표 항목을 분류하여 문자를 인식하거나[3] 문서 내 표 검출 정확도를 향상하기 위해 연구에서 제안한 모듈과 Faster R-CNN 모듈 간의 성능을 비교하는 연구가 수행되었다[1-2]. 하지만 문서 내 의미 있는 표 정보를 활용하기 위해 텍스트를 추출하고 이를 문장화하는 연구 사례는 미흡했다. 이에 본 연구에서는 문서 내 표 텍스트 추출 결과를 자동으로 문장화하기 위한 규칙 기반의 텍스트 재배열 방법을 제시하고자 한다.

2. 딥러닝 객체 감지 기반 표 이미지 및 텍스트 추출 방법

그림 1과 같이 한글 문서 내 표 텍스트 추출 결과를 규칙 기반으로 재배열하여 문장화하는 방법을 제안하기 위해, 먼저 딥러닝 객체 감지 기반 표 추출 모듈을 활용하여 문서 내 표와 박스 이미지를 기존 문서 이미지와 분리하여 저장하는 작업을 수행하였다. 다양한 표 형태를 포함한 건설 분야 입찰 문서(한글파일, hwp)로부터 2,000장의 이미지(jpg)를 추출하고, 객체 감지 모델인 Yolo가 각 이미지에 포함된 표 및 박스 객체를 인식하도록 학습시켰다. Yolo 모델이 표, 글 상자, 내부 표, 기타 형상을 구분할 수 있도록 학습 데이터를 생성하였으며, 모델의 정량적 성능지표 확인 및 표 객체 검출 결과에 대한 정성적 검토를 반복하여 최종적인 문서 내 표 검출모델을 선정하였다.

개발된 표 검출 딥러닝 모델은 표 객체의 종류와 좌표 정보를 제공한다. 문서 내 표 내부의 텍스트 정보만 추출하기 위해 표 검출 딥러닝 모델 결과를 Open CV로 전달해 표 영역만 구분하였다. 또한 표 영역 이미지에 대해서만 Google Cloud Platform의 Vision API를 적용하여 표 내부의 텍스트를 인식하도록 하였다.



[그림 1] 딥러닝 객체 인식을 활용한 문서 내 표 텍스트 정보 추출

3. 규칙 기반 표 텍스트 재배열 및 문장화 결과

딥러닝 및 규칙 기반 표 텍스트 모듈을 통해 추출한 표 및 박스 내 텍스트와 각각의 좌표 정보를 활용하여 규칙 기반의 텍스트 정보 재배열 작업을 수행하였다. 텍스트의 x, y 좌표를 기반으로 인접한 좌표 순서대로 텍스트를 연결하여 자동으로 문장화시키는 규칙을 작성하였다. 표 및 박스 이미지 단위로 추출된 텍스트를 규칙 기반으로 재배열 및 문장화 한 결과, 원문 문장과 동일한 형태로 문장이 생성됨을 확인하였다.

4. 결론

본 논문에서는 딥러닝 객체 감지 모델인 Yolo와 객체 이미지 추출을 위한 OpenCV를 활용하여 문서 내 표 및 박스 이미지를 추출하였다. 이를 기반으로 Vision API기반 텍스트 인식 및 위치 정보 도출을 통해 한글 문서에서 표 및 박스 부분을 구분하여 텍스트 정보를 추출하였다. 본 연구는 한글 문서 내 의미 있는 내용을 담은 표나 박스 정보를 문서의 본문 내용과 구분하여 텍스트를 추출하고 자동으로 문장을 생성함으로써 문서 내 표 정보를 활용할 수 있다는 점에서 의의가 있다. 향후 연구에서는 표의 셀 단위로 텍스트를 재배열하는 규칙을 통해 자동으로 문장을 생성하는 연구를 진행할 예정이다.

감사의 글

이 논문은 국토교통부 BIM 기반 인프라 발주-설계 프로세스 디지털 협업 체계 개발사업(R&D) 연구비 지원에 의한 결과로 수행되었습니다. (과제번호: RS-2022-00143371)

참고문헌

[1] M. Y. Kyoung, H. B. Lee, “Improving Accuracy of Table Detection in Document Image using Loss Compensation Faster R-CNN”, *Journal of the Institute of Electronics and Information Engineers*, Vol.58, No.6, pp.61-70, 2021.

[2] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, “TableBank: Table Benchmark for Image-based Table Detection and Recognition”, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp.1918 - 1925, May. 2020.

[3] D. S. Lee, S. K. Kwon, “Methods of Classification and Character Recognition for Table Items through Deep Learning”, *Journal of Korea Multimedia Society*, Vol.24, No.5, pp.651-658, 2021.