AI OCR을 위한 오픈소스 기반 OCR의 사전 조사 및 실증 테스트에 관한 연구

김채윤, 안성열, 전은진, 차병래*

제노테크(주), *광주과학기술원 AI 대학원

chaeyunk0104@gmail.com, *brcha@smartx.kr

A Study on the Survey and Demonstration Test of OCR based on Open Source for AI OCR

Chae Yun Kim, Seong Yeol An, Eun Jin Jeon, *Byung Rae Cha.

요 약

본 논문은 기업이 비즈니스와 사회에 영향을 주는 새로운 기술에 빠르게 적응하기 위한 디지털 트래스포메이션(Digital Transformation, DX)인 AI OCR에 관한 내용이다. DX에 하이퍼-오토메이션(Hyper-Automation), RPA(Robot Process Automation), 인공 지능(Artificial Intelligence) 및 머신 러닝(Machine Learning)등 첨단 기술이 접목돼 인간의 노력과 능력을 강화한다. 이중 AI OCR은 AI 학습 데이터를 기반으로 디지털화 된 문서 이미지를 문자로 인식하는 규칙을 적극적으로 만들어 내고 있다. 본 연구에서 우리는 오픈소스 기반 OCR중 Tesseract와 EasyOCR을 이용하여 공공 서류의 문자 인식, 도로 전광판 표지와 표지판의 문자인식 인식률의 일부를 파악하였다.

Ⅰ. 서 론

본 논문에서는 새로운 디지털 기술의 혁신에 따른 프로세스 및 제품의 변화를 의미하는 디지털 트랜스포메이션(Digital Transformation, DX)[1]이 중요한 이유를 이야기 하고자 한다. 기업이 비즈니스와 사회에 영향을 주는 새로운 기술에 충분히 빠르게 적응하지 못하는 경우, 해당 기업의 제품 또는 기업 자체가 사장될 수 있기 때문이다. 더불어, 하이퍼-오토메이션(Hyper-Automation, HA)[2]은 자동화를 조직의 비즈니스 프로세스에 지속적으로 통합하는 프로세스이다. RPA(Robot Process Automation), 인공 지능(Artificial Intelligence) 및 머신 러닝(Machine Learning)과 같은 고급 기술을 결합하여 인간의 노력과 능력을 보강하고 있다.

본 논문의 2장 관련 연구에서는 RPA와 OCR의 기본 개념들을 서술하며, 3장에서는 AI OCR의 기능과 AI OCR을 활용한 비즈니스 모델을 제안한다. 4장에서는 오픈소스 OCR 중의 Tesseract와 EasyOCR을 이용한 성능을 테스트하고 마지막 결론을 맺고자 한다.

Ⅱ. 본론

2. 관련연구

본 논문에서는 AI OCR의 HA을 위한 RPA와 OCR에 관한 기본 개념들을 서술한다.

2.1 RPA

RPA(Robotic Process Automation)[3]은 디지털 시스템 및 소프트웨어와 사람 사이의 상호 작용을 에뮬레이션하는 소프트웨어 기술로 소프트웨어 로봇을 쉽게 빌드, 구현 및 관리할 수 있도록 한다. 소프트웨어 로봇은 사람과 마찬가지로 스크린에 표시된 내용을 이해 후 적절한 키를 입력하고, 시스템을 탐색하여 데이터를 식별 및 추출한다.

2.2 OCR

광학 문자 인식(OCR)[4]은 텍스트 이미지를 기계가 읽을 수 있는 텍스트 포맷으로 변환하는 과정을 의미한다. 예를 들어 양식 또는 영수증을 스캔하는 경우 컴퓨터는 스캔본을 이미지 파일로 저장하게 되며, 이미지 파일에서는 텍스트 편집기를 사용하여 단어를 편집 또는 검색하거나 단어 수를 계산할 수 없다. 그러나 OCR을 사용하면 이미지를 텍스트 문서로 변환하여 내용을 텍스트 데이터로 저장할 수 있게 된다.

3. AI OCR과 비즈니스 모델

AI OCR을 위한 오픈소스 기반 OCR의 사전 조사와 AI OCR 기술을 활용한 비즈니스 모델을 제안하고자 한다.

3.1 AI OCR

AI OCR은 기존의 OCR에서 AI가 결합한 기술로 AI 학습 데이터를 기반으로 디지털화 된 문서 이미지에서 문자를 인식하는 규칙을 능동적으로 만들어낸다. OCR 프로세스 전반에 머신러닝, 딥러닝(Deep Learning), 자연어처리(Natural Language Processing) 알고리즘을 활용하여 문자 탐지 및 문자 인식 성능을 높일 수 있다.

그러나 AI OCR은 기존의 AI에서도 겪고 있는 적대적 공격과 블랙박스 회피 공격 및 AI 연합학습 보안 취약점 공격 등에 대한 보안 허점이 있다. 한글의 경우 이전과 비교하면 향상은 되었지만 글자 구조를 분석해 특징을 뽑아내고 학습에 필요한 데이터셋(Dataset)을 만드는 작업은 여전히 어려운 상태이다.

3.2 오픈소스 기반의 OCR

AI OCR은 Google Cloud Vsion, Naver CLOVA, Nanonets, ABBYY FineReader PDF등 유료와 Tesseract, GOCR, CuneiForm, Kraken, A9T9, EasyOCR등 오픈소스가 있다[5].

Tesseract는 오픈 소스 OCR 엔진으로 2006년부터 Google이 후원하는 Apache 라이센스 하의 무료 소프트웨어이다. Tesseract OCR 엔진은 LSTM 기반 최신 안정 버전 4.1.1 버전에서 Tesseract는 이제 최대 116개 언어를 지원합니다. CIL(command-line interface)에서 실행되는 Tesseract는 자체 GUI(graphical user interface)가 없기 때문에 별도의 GUI(graphical user interface)가 필요하다. 정교한 이미지 전처리 파이프라인을 갖추고 있으며 신경망을 통해 새로운 정보를 학습할 수 있다.

또한, EasyOCR은 문자 영역 인식(Detection), 문자 인식(Recognition)을 손쉽게 수행 할 수 있도록 하는 Python 패키지이며, 구현이 간단하고 매우 직관적이다. 현재 80개 이상의 언어를 지원하고 있으며, 최근에는 손글씨 인식을 목표로 하고 있다. 본 연구에서 진행할 테스트는 Tesseract[6]와 EasyOCR[7] 도구를 이용하여 진행하였다.

3.3 AI OCR을 활용한 비즈니스 모델

정보 수집을 위해 다양한 포맷의 디지털 문서와 파일들을 프로세싱 하는 경우 과거에는 텍스트가 텍스트로서 인식되는 PDF 문서까지를 허용범위 안에 두었다. 그러나 AI OCR은 이런 과정에서 미인식 요소였던 스캔이나 촬영된 이미지 형태의 문서들을 읽어낼 뿐만 아니라 문서에 담긴 내용, 수치, 통계, 기준문서와의 대조정보, 서식, 날인의 진위 등 다양한 업무자동화 과정을 가능케 한다.

기업의 지자체 대상 계약과정이나 공공기관의 지원사업 제안서 접수과정을 대상으로 현재 기업과 기관 사업담당자의 세부적인 업무과정을 파악 후, 플랫폼 사용의 전체적인 흐름(UX - User Experience)과 세부 접근 방식(UI - User Interface) 구성을 위한 사전 조사과정을 진행하고 문서 송신, 제출과 수신, 접수과정을 AI OCR 기술을 기반으로 스마트화 하는 문서운용 플랫폼을 구축하는 것이 프로젝트 진행의 주축이다.

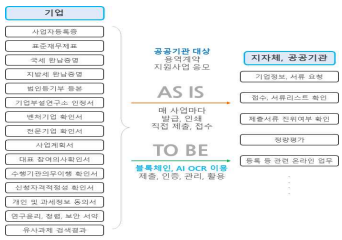

그림 1. AI OCR 기술기반에 스마트 문서운용 플랫폼

4. AI OCR의 문자 인식 테스트

AI OCR을 위한 문자 인식 테스트는 공공 서류의 문자인식, 도로전광표지와 도로표지판의 문자인식 테스트를 계획하였으며, 오픈 소스 Tesseract와 EasyOCR를 이용하여 테스트를 진행하였다.

4.1 공공 서류의 문자인식 테스트

AI OCR의 한글 문자 인식성능을 테스트하기 위해서 사업자 등록증(그림 2 참조)과 재무제표(그림 5 참조)를 각각 AI OCR 오픈소스인 Tesseract 및 EasyOCR를 사용하여 문자인식을 진행하였으며, 테스트 결과는 다음과 같다.

그림 2. 사업자등록증의 테스트문서

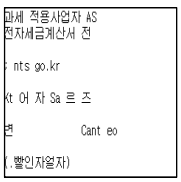

그림 3. Tesseract의 사업자등록증의 문자인식 결과

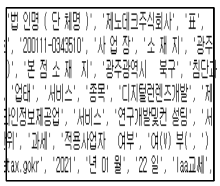

그림 4. EasyOCR의 사업자등록증의 문자인식 결과

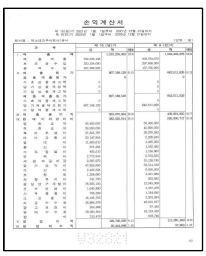

그림 5. 재무제표의 테스트 문서

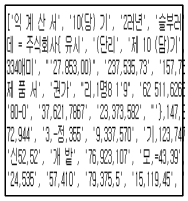

그림 6. EasyOCR의 재무제표의 문자인식 결과

테스트 결과 한국어의 문자 인식 성능은 저조한 것을 확인할 수 있었다. 무엇보다 사업자등록증의 경우에는 몇몇 제대로 인식하는 글자가 있는 반면 재무제표의 경우 전체적으로 문자 인식을 못하는 것을 확인할 수 있었다.

4.2 도로전광표지 문자인식 테스트

도로전광표지(그림 7 참조)의 문자인식 테스트를 각각 Tesseract 및 EasyOCR를 사용하여 문자인식을 진행하였으며, 테스트 결과는 다음과 같다.

그림 7. 도로전광표지의 테스트 사진

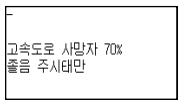

그림 8. Tesseract의 도로전광표지의 문자인식 결과

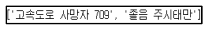

그림 9. EasyOCR의 도로전광표지의 문자인식 결과

테스트 결과 전체적으로 문자인식률의 차이는 미묘하였다. Tesseract의 경우 약 85% 이상 인식하였으며 EasyOCR은 문제없이 문자를 인식한 것을 확인할 수 있었다.

Ⅲ. 결론

최근 디지털 트랜스포메이션, 하이퍼-오토메이션, RPA, 인공 지능 및 머신 러닝 등과 같은 고급 기술을 결합하여 인간의 노력과 능력을 보강하고 있다. AI OCR은 기존의 OCR에서 AI가 결합한 기술로서 AI 학습 데이터를 기반으로 디지털화 된 문서 이미지에서 문자를 인식하는 규칙을 능동적으로 만들어낸다.

본 연구에서는 AI OCR을 활용한 비즈니스 모델을 수립하였으며, 오픈소스 기반 OCR의 사전 조사 및 실증 테스트를 진행하였다. 오픈소스 기반 OCR중의 Tesseract와 EasyOCR 도구를 이용하여 공공 서류의 문자인식, 도로전광표지의 문자인식을 통한 오픈소스 OCR 성능의 일부분을 확인하였다. EasyOCR이 Tesseract보다 인식률이 더 높음을 확인할 수 있었다.

ACKNOWLEDGMENT

본 과제는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구결과입니다.

참 고 문 헌

[1] 디지털 트랜스포메이션, “디지털 트랜스 포메이션이란?”, https://www.vmware.com/kr/topics/glossary/content/digital-transformation.html

[2] 하이퍼오토메이션, “하이퍼오토메이션이란 무엇입니까?”, https://www.tibco.com/ko/reference-center/what-is-hyperautomation

[3] RPA, “RPA(Robotic Process Automation)란?”, https://www.uipath.com/ko/rpa/robotic-process-automation

[4] OCR, “광학 문자 인식(OCR)이란 무엇인가”, https://aws.amazon.com/ko/what-is/ocr/

[5] Open Source OCR Tools, https://www.hitechnectar.com/blogs/open-source-ocr-tools/

[6] Tesseract, “Tesseract”, https://github.com/tesseract-ocr/tesseract

[7] EasyOCR, “EasyOCR”, https://github.com/JaidedAI/EasyOCR