오픈소스 기반 OCR의 한국어 인식성능 비교분석에 관한 연구

김정섭[1], 김현정[2], 유상현[3]
[1]경민대학교 컴퓨터소프트웨어과 학부생
[2]건국대학교 상허교양대학 교수
[3]경민대학교 컴퓨터소프트웨어과 교수
20234005@kyungmin.ac.kr, nygirl@konkuk.ac.kr, simonyoo@kyungmin.ac.kr

Comparative Analysis of Korean Language Recognition Performance in Open Source-Based OCR

Jeong-Seob Kim[1], Hyun-Jung Kim[2], Sang-Hyun Yoo[3]
[1]Dept. of Computer Software, Kyungmin University
[2]Department of Sang-Huh College, Konkuk University
[3]Dept. of Computer Software, Kyungmin University

요 약

문서 전자화 시스템의 도입에 따라 OCR에 관련된 많은 연구가 진행되고 있으며, 현재 넓은 분야에서 OCR을 활용 중이다. 그러나 OCR 라이브러리들의 한국어 인식성능에 어느 정도 차이가 있는지에 대한 의문이 생기고 있다. 본 논문에서는 현재 사용 중인 OCR 라이브러리의 한국어 인식성능을 비교, 분석하였고 Tesseract OCR이 더 인식성능이 좋다는 결과를 얻었다.

1. 서론

최근 서류 보관의 중요성이 늘어남에 따라 문서 전자화 시스템을 도입하는 기업들이 증가하고 있다. 이러한 상황에서 수작업에 의한 오류와 인력을 사용한 단순 작업을 줄이기 위한 기술로 OCR(Optical Character Recognition)이 부상하고 있다.
현재 OCR 기술은 전자문서 변환을 비롯한 자동차 번호판 인식, 영수증, 신분증 인식 등 넓은 분야에서 활용되고 있다.

본 논문에서는 현재 개발된 OCR 라이브러리 중 Paddle OCR과 Tesseract OCR의 한국어 인식성능을 비교분석 하고자 한다.

본 논문의 구성은 Paddle OCR의 소개로 시작하여 Paddle OCR과 Tesseract OCR 한국어 인식성능을 분석하였고 마지막으로 결론과 향후 과제에 관하여 서술하였다.

2. 관련 연구 소개

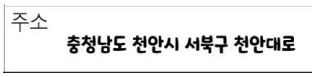
PaddlePaddle은 'PArallel Distributed Deep LEarning'의 약자로 C++기반의 파이썬 인터페이스를 가지고 있고, 수학 연산의 성능과 분산환경 그리고 순환신경망에 최적화를 한 프레임워크이다.[1]

Paddle OCR은 중국의 인터넷기업 바이두(Baidu)가 만든 딥러닝 프레임워크인 'PaddlePaddle'로 구현된 오픈소스 OCR로 중국어, 영어 이외에도 한국어를 포함한 80개 이상의 다양한 언어를 지원한다.[2]

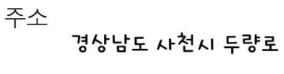
Paddle OCR의 경우 우선 글자 영역을 확인한 이후 보정 처리를 진행하고 글자를 인식하는 순서로 진행되는 반면에 Tesseract OCR은 전처리를 진행한 이후, 글자의 윤곽선을 통해서 인식을 진행하는 방식이다.

3. 실험 환경 및 결과

OCR의 한글 문자 인식성능을 테스트하기 위해서 두 라이브러리 공통으로 python 3.10 환경에서 CPU를 사용하였으며, Tesseract OCR은 5.0.0버전과 tessdata_best의 kor 모델을, Paddle OCR은 2.7.0.2 버전과 korean_pp-OCRv4 모델을 이용하였다. 샘플 데이터는 AI-Hub 사이트의 '다양한 형태 문자 OCR'의 샘플 데이터를 가공하여 제작한 고화질 2장(그림 1, 2 참조), 저화질 2장(그림 7, 8 참조) 두 개의 샘플 데이터를 이용하여 인식을 진행하였으며 결과는 다음과 같다.



(그림 1) 고화질 샘플 1

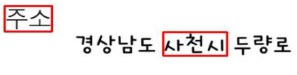


(그림 2) 고화질 샘플 2

(그림 3) Tesseract 인식결과 1

(그림 4) Tesseract 인식결과 1

(그림 5) Paddle 인식결과 1

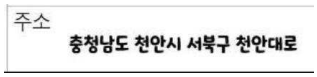


(그림 6) Paddle 인식결과 1

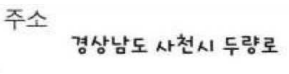
OCR명	샘플 1	샘플 2
Tesseract	100%	50%
Paddle	50%	42%

(표 1) 고화질 인식결과

테스트 결과 정확하게 인식한 글자 개수를 비교했을 때는 Tesseract OCR이 Paddle OCR보다 문자 인식 성능이 높은 것을 확인할 수 있었다.

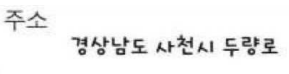


(그림 7) 저화질 샘플 1



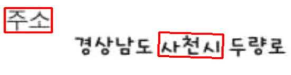
(그림 8) 저화질 샘플 2

(그림 9) Tesseract 인식결과 1



(그림 10) Tesseract 인식결과 2

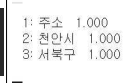
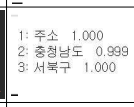
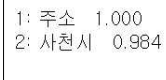
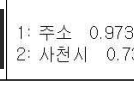
(그림 11) Paddle 인식결과 1



(그림 12) Paddle 인식결과 1

OCR명	샘플 1	샘플 2
Tesseract	88%	0%
Paddle	56%	42%

(표 2) 저화질 인식결과

OCR명	Tesseract	Paddle
고화질 1		
저화질 1		
고화질 2		
저화질 2		

(표 3) 인식결과 정리

화질 차이에 의한 성능 변동에 경우 Tesseract OCR은 화질이 낮거나 글꼴에 따라서 중복인식, 인식결과 깨짐 등의 현상을 보였고, Paddle OCR의 경우에는 큰 영향을 받지 않는 것을 확인할 수 있었다.

4. 결론 및 향후 계획

본 논문은 두 가지 OCR 라이브러리의 한국어 인식성능을 비교해 보았다. 실험결과 Paddle OCR의 경우 선명도에 의한 성능의 차이는 비교적 적었으나, Tesseract OCR보다 전반적으로 인식률이 떨어지는 결과를 볼 수 있었다. 이에 향후 Paddle OCR의 인식률 향상을 위한 연구를 진행할 예정이다.

참고문헌

[1] Ma, Yanjun, et al. "PaddlePaddle: An open-source deep learning platform from industrial practice." Frontiers of Data and Domputing 1.1 105-115. (2019)

[2] Li, Yunjie, and Dan Zhang. "Research and Application of Health Code Recognition Based on Paddle OCR under the Background of Epidemic Prevention and Control." Journal of Artificial Intelligence Practice 6.1. 9-16. (2023)

[3] Tesseract-OCR manual ver 4.0, Google International, 2017