Tesseract-OCR 의 한글 영어 혼용 인식률 향상

*전재민, *윤아영, *전형록, *궈우 ,*김정홍
*경북 대학교 컴퓨터학부
e-mail : {eexxqqtt,jaemin3663, yunay9710, guowu0105}@gmail.com

Korean/English mixed character Increasing recognition rate by tesseract

Jaemin JEON,Hyeongrok JEON, GUOWU, Ayeong YUN

Dept of Computer Science, Kyungpooknational University

요 약

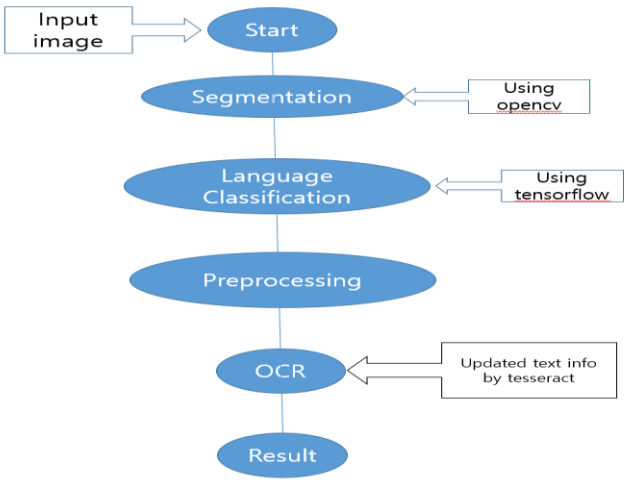
현재의 Tesseract-OCR 은 한글, 영어가 혼용되어 있는 경우 정확하게 인식하는 것에 한계가 있다. 이를 개선하기 위하여 기존에 학습되어 있는 한글 데이터에 영어를 더하여 학습시켜 Tesseract-OCR 의 인식률을 높인다. 향후 Tesseract 전처리 분류과정에서 영어 한글 구분 과정의 정확도를 높이고 이를 이용하여 기술이 향상된 Tesseract-OCR 모바일 어플리케이션은 개발할 예정이다.

1. 서론

광화 문자 인식 (Optical character recognition, OCR)은 사람이 쓰거나 기계로 인쇄한 문자의 영상을 이미지 스캐너로 획득하여 기계가 읽을 수 있는 문자로 변환하는 것이다. 초기 시스템은 특정한 서체를 읽기 위해 미리 샘플을 읽는 “트레이닝”이 필요했지만, 현재 대부분의 서체를 높은 확률로 변환이 가능하다.

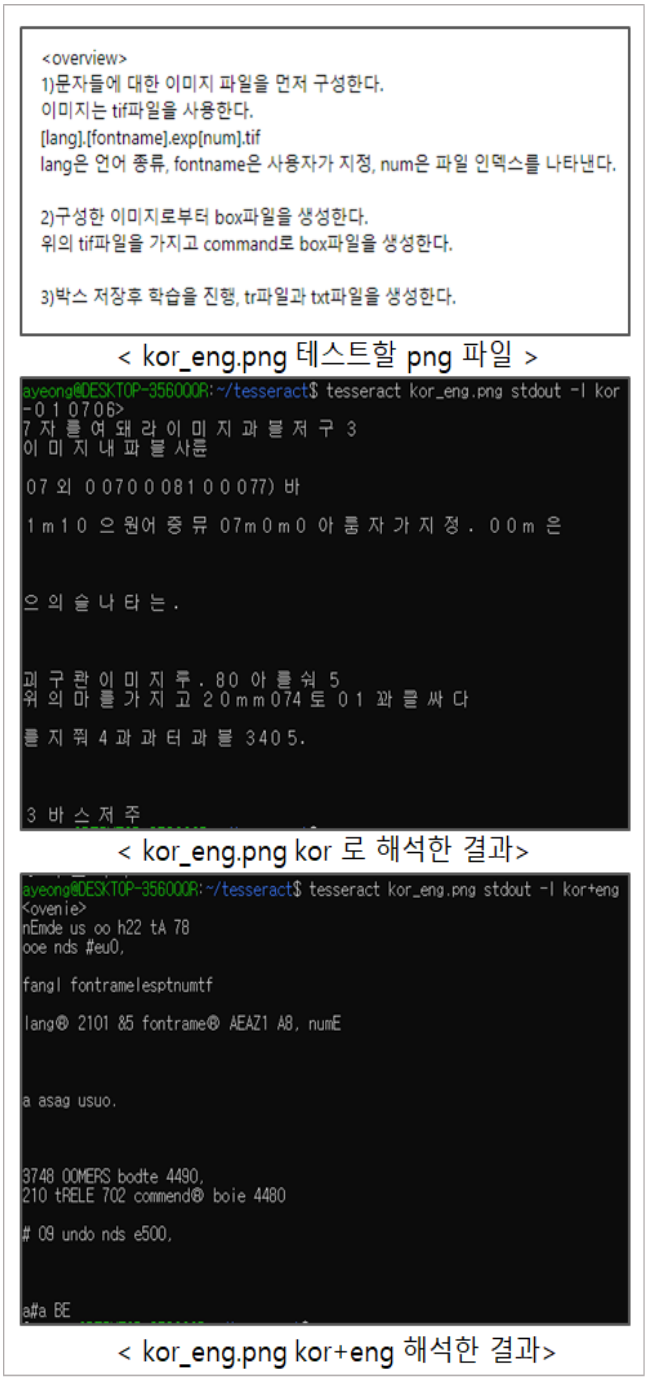
이를 제공해주는 라이브러리 “Tesseract”가 있다. 구글에서 제공하는 다양한 OS 를 지원하는 오프라인 OCR 엔진이자 오픈소스 라이브러리이다. 앞서 말한 것처럼 사진촬영을 해서 이미지를 넘겨주면 이를 분석하는 프로세스이다.

그림 1 Tesseract 프로세스



한국의 경우 업무에서나 일에서나 많은 이미지들이 한글과 영어를 혼용해서 쓰는 경우가 많다. 특히 논문이나 기사에서 그러한 경우를 많이 볼 수 있다. 이러한 경우 혼용되어있는 이미지를 오류 없이 인식할 수 있도록 하는 것이 중요하다.
현재 영어 뿐 아니라 한글을 포함한 다양한 언어들을 이미지를 통해 꽤 높은 적중률로 추출 할 수 있다. 하지만 문제는 한글과 영어 같이 섞여 있는 이미지의 경우 영어가 전부 깨져서 나오는 등 문제점을 파악할 수 있다. Tesseract 를 사용할 때 문제점은 인식해야 할 언어를 미리 알고 있어야한다. 그러나 컴퓨터가 이미지를 인식 할 때 그 이미지가 어떤 언어를 포함하고 있는지 모르는 경우가 대부분이다. 그렇기 때문에 인식률이 크게 떨어진다. 그림 1 에서 똑 같은 이미지를 한글 영어 혼합과 영어로만 교육된 tesseract 로 실험한 결과 한영 혼합한 학습데이터로도 인식률이 크게 떨어지는 것을 볼 수 있다. 이런 결과를 나타내는 이유는 컴퓨터가 미리 알지 못하고 이미지를 보고 한글 영어를 구분해야 하기 때문이다.
(1) 전처리과정을 모두 거친 후, OCR 단계에서 한글에 영어를 학습시키는 방법과 (2) 전처리과정에서 이미지의 텍스트가 한글인지 영어인지를 구분하는 방법을 계획중이다. 한글과 영어 혼용의 경우의 문제를 해결한 후에는 다른 언어와의 혼용했을 시에도 똑같은 방법을 적용하여 해결할 수 있을 것이다.

그림 2 한글 영어 혼용의 인식률 저하 현상



2. 관련연구

최근에 이미지 인식을 통하여 텍스트를 추출하는 방법은 점점 더 중요해지고 있다. 이런 프로그램 중에서 많은 사람들이 사용하는 것이 tesseract OCR 과 구글 비전이다. 구글 비전의 경우 한글 영어가 혼용되어 있는 이미지를 인식하여도 웹 크롤링을 통하여 단어 문단을 단위로 인식하며 스스로 학습을 더하여 더 정확한 결과를 보여준다. 거의 정확하게 한글 영어 혼용되어있는 이미지를 인식한다는 것을 볼 수 있다. 그에 반하여 tesseract 는 한글 영어 혼용되어있는 경우에서 인식률이 많이 낮음을 알 수 있다.

그러나 구글 비전의 경우 문단과 단어 단위로 학습을 하기 때문에 문맥상 말이 안 되는 경우의 단어가 혼용된 경우는 인식률이 크게 떨어진다. 이런 단점을 개선하고자 tesseract-OCR 을 이용하여 한글 영어 혼용되어있는 이미지 인식률 개선을 위한 방법을 제시한다.

표 1 Tesseract-OCR 구현 환경

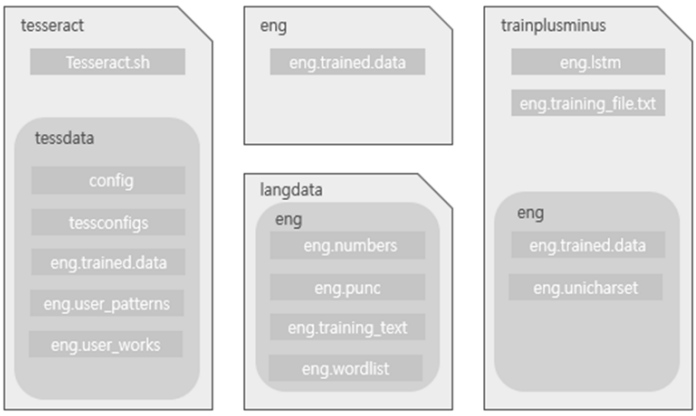
	환경
OS	Ubuntu 18.04
CPU	Intel Core i7 4790
RAM	8GB
SSD	Samsung Portable SSD T5 250GB
Tesseract	4.1.0-rc1
Leptonica	1.75.3

3. 설계

언어 학습 이전 "±"을 학습시켜보겠다. 우선 트레이닝 시키기위해 글자에 대한 정보(font, 글자크기, 가능한 문자 set 등등)를 langdata 폴더를 생성하여 안에 다운 받는다. 영어에 ±를 학습 시키기 위해 생성한 langdata 내 eng 폴더를 만들어 eng 학습데이터를 다운 받는다. 이제 학습에 필요한 데이터를 만든다. 결과물을 넣을 폴더를 미리 만들어놓고 tesstrain.sh 를 실행시킨다. 기존에 교육된 데이터 eng.traineddata 를 가져와서 새롭게 학습 시킬 data 를 만든다. 출력파일을 결합한다.

그리고 결과를 확인 할 수있게된다. 이 학습과정을 거치면 아래와 같은 데이터구조를 확인 할 수 있다.

그림 2 트레이닝 UML 다이어그램



4. 구현

본 연구에서는 tesseract 자체에서 제공하는 테스데이터를 이용하여 글에서 한글 영어 비율에 따른 오차율을 구하였다. 표 2 에서 우선 page 1 과 page 2 는 영어 한글이 각각 1:9 page 3 과 page 4 는 1:4 의 비율로 구성되어 있다. 본 연구의 테스트 데이터는 아무 의미 없는 데이터들을 모아서 각각의 비율에 맞추어 문장을 만들어 주는 파이썬 스크립트를 사용하여 만들었다.

표 1 에서 볼 수 있다시피 eng + kor 의 경우에는 비율이 커지면서 더 큰 오류가 생기는 것을 볼 수 있다. 게다가 영어의 경우는 영어로조차 인식을 하기 힘든 것을 볼 수 있었다. Hangul 의 경우에는 오차율이 그렇게 크지 않음을 볼 수 있다. 그래도 10 퍼센트가 넘는 오차율을 보이며 개선이 필요하다.

그림 2 트레이닝 UML 다이어그램

	Page1	Page2	Page3	Page4
Eng+kor error	40%	32 %	61%	62%
Hangul Eroor	17%	12%	16%	8%

본 연구에서는 한글 영어의 인식의 오차율을 줄이기 위하여 기존의 tesseract 의 한글만 학습되어 있는 데이터 kor.traineddata 에 영어를 학습시켜 보았다. 이를 통하여 앞에서 kor + eng 학습된 데이터로 실험한 결과에서는 영어와 한글을 전혀 구분하지 못했으나 새로 학습된 데이터에서는 한글과 영어를 정확하게 구분하였지만 오차률이 커 개선해야 한다.

5. 결론 및 향후 연구

본 연구는 한글과 영어가 혼용되어 있는 이미지를 Tesseract-OCR 의 인식률에 대해서 연구하고 Tesseract-OCR 의 인식률을 향상시키는 방안에 대하여 제안한다. 기존의 Tesseract-OCR 은 한글과 영어 각의 교육된 데이터에서만 적용이 가능하여 하나의언어에서의 인식률이 높았다. 그러나 한글과 영어의 혼용의 경우 인식이 거의 정확하지 않았다. 이를 개선하기 위하여 현재 Tesseract 한글 학습 데이터에 영어를 추가로 학습시켜서 한글과 영어가 혼용되어 있는 이미지를 인식하였으며 개선된 것을 볼 수 있었다.

향후에는 학습 데이터의 개선 만으로는 한계가 있는 한글 영어 혼용 이미지 인식률을 높이기 위하여 전처리 과정에서의 한글 영어 구분을 하는 방법을 도입할 예정이다. 이 후에 한글 영어 말고도 다양한 언어가 혼용되어 쓰는 경우를 위한 학습을 실시하고 연구할 예정이다. 그리고 Tesseract-OCR 기반으로 된 모바일 어플리케이션과 프로그램을 제작할 예정이다.

참고문헌

[1] 김원표, 이정근, 고영웅, 손철준, 진창규 (2018). Tesseract-OCR 의 인식률 향상을 위한 병렬 전처리 모델. 한국정보 과학회 학술발표논문집, 641-643.

[2] 이정원, 김수호, 김세현, 조영태, 조예은, 최유진, 주홍택 (2018). 사용자 정보 및 번역 통계 기반 영 · 한 단어 번역 최 적화 시스템. 한국정보과학회 학술발표논문집, 1765-1767

[3]김민기 (중앙대학교 컴퓨터공학과), 권영빈 (중앙대학교 컴퓨터공학과), 한상용 (중앙대학교 컴퓨터공학과)
The journal of the Korean institute of communication science v.22 no.3 ,pp. 410 - 422 , 1997 , 1226-4717 ,

[4] Huabin Zheng, Jingyu Wang, Zhengjie Huang, Yang Yang, Rong Pan Chinese/English mixed Character Segmentation as Semantic Segmentation Submitted on 7 Nov 2016 (v1), last revised 16 Nov 2016 (this version, v2

[5] B.A. Blesser, T.T. Kuklinski, R.J. Shillman, "Empirical Tests for Feature Selection Based on a Pscychological Theory of Character Recognition" in Pattern Recognition, New York:Elsevier, vol. 8, no. 2, 1976.

[6]M.Bokser,"Omnidocument Technologies", *Proc. IEEE*, vol. 80, no. 7, pp. 1066-1078, Jul 1992.

[7]Seung-Hun Lee, Jin-Ho Jeon, Hae-Sung Hong, Dong-Hyuk Kang and Mee-Hwa Park, 2017, "Korean Prescription Character Recognition System Using OCR Technology," 한국정보과학회 학술발표논문집, , pp. 362~364.

[8]Ji-yeon Kim, Young-jin Hur and Chulyun Kim, 2018, "InText : A Responsive Web Application to Translate Text in Image based on Deep Learning," 한국정보과학회 학술발표논문집, , pp. 2160~2162.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"(2015-0-00912)