OCR 및 기하변환을 활용한 강건한 상품정보 인식방법

김동민[1], 전광명[2], 전찬준[3], 최우열[1,*]
[1]조선대학교 컴퓨터공학과
[2]인트플로우㈜ AI융합기술연구소
[3]한국건설기술연구원 차세대인프라연구센터
e-mail : 20164275@chosun.kr, kmjeon@int-flow.com, chanjunchun@kict.re.kr, wyc@chosun.ac.kr

Robust Product Information Recognition Method using OCR and Geometric Transformation

Dongmin Kim[1], Kwang Myung Jeon[2], Chanjun Chun[3], Wooyeol Choi[1,*]
[1]Department of Computer Engineering, Chosun University
[2]AI Convergence Technology Laboratory, IntFlow Co., Ltd.
[3]Future Infrastructure Research Center, KICT

Abstract

본 논문은 기하변환을 통해 OCR의 정확도를 높여 상품정보를 정확하게 인식하는 기술을 구현한다. 이미지에서 경계선을 추출한 뒤, 추출한 경계선에서 contour를 찾고 이를 근사화하여 사각형 영역을 구한다. 이 사각형 영역을 문자를 인식할 대상, 즉, 문서영역으로 간주하여 기하변환을 적용한 후, Tesseract를 사용하여 문자를 인식한다. 또한, 원본 이미지와 기하변환을 적용한 이미지의 인식률을 비교하여 분석하였다.

I. 서론

Optical character recognition (OCR)은 광학 문자인식 기술을 말하며, 사람이 쓰거나 기계가 인쇄한 문자 이미지를 기계가 읽을 수 있도록 문자로 변환하는 것이다. OCR 기술은 사용하기 쉽고 유용하기 때문에 다양한 분야에서 활용되고 있다. 하지만 이미지 내의 문자의 각도가 틀어져 있을 경우, 문자 인식 정확도는 현저히 떨어진다.

본 논문에서는 촬영된 문서에 기하변환을 적용하여 OCR의 인식률을 높이는 기술을 구현한다. OCR을 사용하는 과정에서 인식률이 많이 떨어지는 경우가 잦은데, 이러한 문제를 해결하기 위해서 기하변환을 하는 전처리 과정을 포함한다. 문서 영역을 찾고 해당 부분에 기하변환을 적용하는 것으로 상품 정보를 정확하게 인식하는 것을 목표로 한다.

II. 본론

2.1 Tesseract 개요

Tesseract는 Hewlett and Packard에서 개발한 OCR 엔진이다. 1995년 문자 정확도 측면에서 3대 ORC 엔진에 속할 정도로 성능이 좋다. Tesseract가 문자를 인식하는 방법은 그림 1과 같다. 먼저 이미지의 임계값을 찾고 이 임계값을 이용하여 이미지를 이진화한다. 이진화한 이미지의 연결된 구성 요소를 분석하여, 각 구성 요소의 외곽선을 추출하고 이진 데이터로 저장한다. 이후 찾은 문자간의 간격에 따라 단어 단위로 나누고 이렇게 나뉜 요소들을 단어 단위와 페이지 단위로 인식한다 [1]. 또한, Tesseract는 딥러닝 기법중 하나인 long-short term memory (LSTM)을 사용한다. 따라서, 추가로 지원하는 툴을 사용하면 손쉽게 사용자가 학습을 진행할 수 있다.

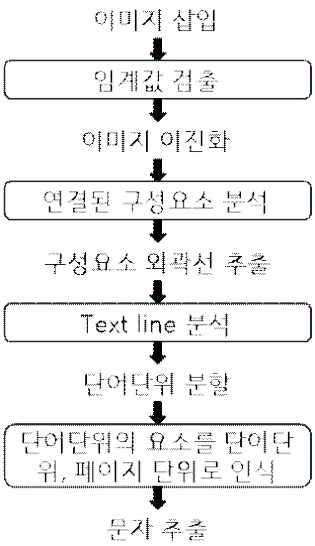

그림 1. Tesseract의 흐름도

2.2 텍스트 인식방법의 구현

상품정보를 인식하는 환경을 모사하기 위하, 문서에 특정 텍스트를 작성하여 성능 분석에 활용하였다. 상품정보 인식을 위한 프로그램에는 Python 3.7과 Tesseract 5.0.0-alpha, OpenCV 4.2.0.34를 사용하였다. 영상처리 및 텍스트 인식을 위해 사용한 라이브러리는 numpy, cv2(OpenCV), pytesseract(Tesseract)이며, 배열이나 리스트를 정렬하기 위해 operator 라이브러리의 itemgetter를 활용하였다.

우선, 입력된 촬영 이미지를 흑백으로 처리한 후, GaussianBlur를 이용하여 주변 픽셀간의 차이를 줄인다. GaussianBlur는 3*3크기의 필터, SigmaX는 0으로 설정한다. 이후 Canny edge detection기법으로 경계선을 추출한다. 이때 두 임계값은 각각 75와 200으로 설정하였다.

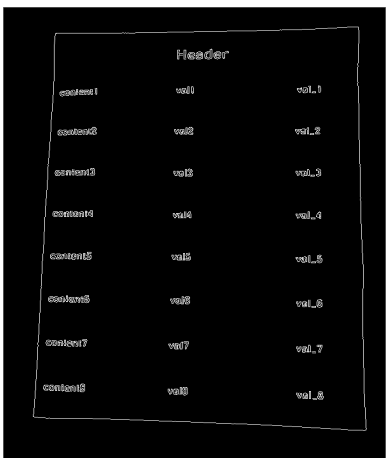

그림 2. 경계선을 추출한 이미지

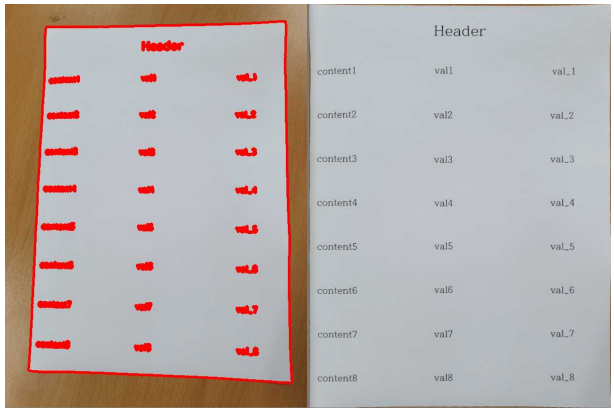

그림 3. Contour를 적용한 이미지와 기하변환을 적용한 이미지

GaussianBlur를 적용했기 때문에 노이즈가 줄어들어 좀더 명확한 경계선이 추출된다. 추출된 경계선을 바탕으로 contour를 찾고, 이를 근사화 시켜 사각형의 요소를 파악한다. 각 요소마다 contour의 길이를 구하고, 다각형 추출 기능을 활용하여 꼭짓점의 수를 파악한다. 꼭짓점의 수가 4개인 객체를 OCR의 대상으로 보고, 사각형의 각 꼭지점의 좌표를 서로 비교하여 기하변환을 수행한다. 이렇게 변환된 이미지는 문서를 스캔했을 때와 유사한 모양이 되고, 원본에서 인식하지 못한 문자들을 포함하여 대부분의 텍스트를 인식할 수 있다.

그림 4. 원본 이미지

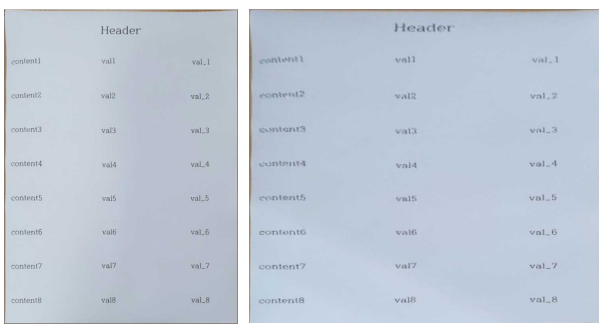

그림 5. 기하변환을 적용한 이미지

Ⅳ. 결론 및 향후 연구

본 논문에서는 상품정보를 인식하기 위해, 촬영된 이미지에 대해 기하변환을 수행하고 OCR 기술을 통해 이미지 내의 텍스트를 인식하는 방법에 대해 확인하였다. 실험에 사용된 모든 이미지에 대해, 기하변환을 적용하지 않은 경우 대부분의 텍스트를 인식하지 못하거나 잘못된 문자로 인식되었다. 반면, 기하변환 수행 후 텍스트를 인식하였을 때, 기존 경우에 비해 텍스트 인식율이 매우 향상되는 것을 확인하였다. 후속 연구에서는 이미지 내의 텍스트를 보다 명확하게 인식할 수 있도록 이미지 전처리 과정을 개선하고자 한다. 또한, 실제 상품정보를 인식하기 위해, 복잡한 형태의 이미지를 사용하여 인식할 수 있는 기술을 제안할 예정이다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업(2017-0-00137) 및 한국연구재단의 지원(NRF-2019R1C1C1011597)을 받아 수행된 연구임.

참고문헌

[1] 권순각, “광학 문자 인식을 통한 단어 정리 방법”, 동의대학교 2015

[2] https://github.com/UB-Mannheim/tesseract/wiki-Tesseract