OCR 기반의 의약품 성분 정보 검색 시스템

박진아O, 박승보*
O인하대학교 소프트웨어융합공학과,
*인하대학교 소프트웨어융합공학과
e-mail: pja9021@inha.eduO, molaal@inha.ac.kr*

OCR-Based Medicine Ingredient Information Retrieval System

Jina ParkO, Seungbo Park*
ODept. of Software Convergence Engineering, Inha University,
*Dept. of Software Convergence Engineering, Inha University

● 요 약 ●

본 논문에서는 의약품의 효율적인 구매와 안전한 복용, 또 의약품 성분에 대한 정보 전달을 위한 시스템을 제안한다. 이 시스템에서는 약품 후면을 촬영한 영상으로부터 이미지 프로세싱을 통해 이미지에서 관심영역을 설정한 뒤, OCR 엔진인 Tesseract-OCR을 사용하여 인식한 텍스트 데이터를 통해 약품 성분을 추출하며, 식품의약품안전처에서 제공하는 의약품 안전 사용 서비스(DUR) API와 네이버 의약품 사전 검색 결과를 이용해 관련 정보들을 읽어와 출력하도록 한다. 약품의 표준 서식을 따르는 이미지를 기준으로 백 개의 이미지를 이용해 테스트하여 65%의 검출 정확도를 보였다.

키워드: 이미지 프로세싱(Image Processing), OCR, 의약품, 성분 검색

I. Introduction

보건복지부는 2012년 11월 15일부터 등록 조건을 만족한다면 약국 의외의 장소에서 '안전상비의약품'을 판매할 수 있도록 약사법을 개정했다. 이로 인해 소비자는 의약품에 가볍게 접근하며, 약국이 문을 닫는 공휴일, 심야 시간대에도 의약품을 구입할 수 있게 되었다. 개인이 약품 각 성분마다의 주의점을 알 수 없기 때문에 전문가의 조언 없이 본인의 판단 하에 의약품을 선택해 복용하는 것은 소비자에게 의약품의 오남용, 그로 인한 부작용 등의 위험이 있을 수 있다. 그러나 주말이나 명절처럼 약국이 문을 닫았을 경우 부득이하게 편의점과 같은 곳에서 의약품 구매를 선택해야 하는 상황이 발생할 수 있다.

본 논문에서는 이런 경우를 대비할 수 있도록 의약품의 성분이 적힌 이미지를 업로드하면 전문가가 아닌 소비자도 의약품을 구성하는 성분들의 상세 정보를 한눈에 확인할 수 있도록 하는 시스템을 제안한다.

OCR(Optical Character Recognition, 광학식 문자 판독기)은 인쇄된 글자 및 글씨를 인식하여 텍스트 데이터로 치환하는 기술을 말한다. 삼성페이나 카카오페이 등과 같이 카드를 등록할 때 카드 이미지를 인식하여 정보를 저장하고, 신분증이나 자격증 또는 명함을 인식하여 데이터를 저장하는 기술 모두 OCR 기술을 응용하여 만들어진 시스템이다. 본 논문은 이런 OCR 기술을 응용해 약품 데이터와 접목시켜 새로운 시스템을 개발하는 것을 목적으로 한다.

II. Medicine Ingredient Information Retrieval System based on OCR

일반의약품의 경우 소비자의 안전한 의약품 사용을 위해 필요한 정보들을 작성해야 하며, 이 정보를 작성하기 위한 표준 서식을 가진다.

① 기본형(1)

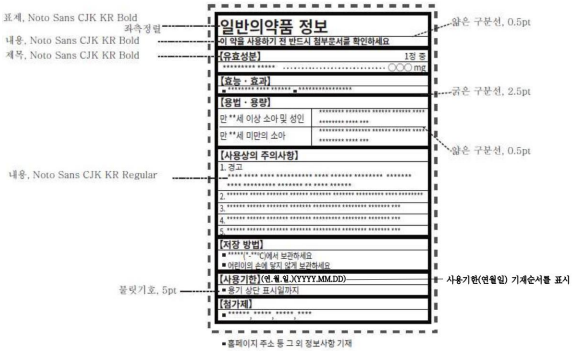

Fig. 1. Medicine Ingredient Standard Form Basic(1)

Figure 1과 같은 정보가 적힌 표시면을 '정보표시면'이라고 하며, 본 논문에서 제안하는 시스템의 경우 이 정보표시면의 [유효성분] 정보를 사용한다.

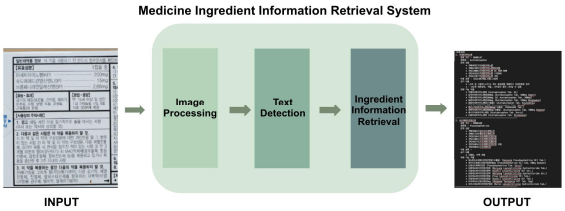

Fig. 2. System Architecture

본 시스템은 Figure 2와 같이 이미지 처리(Image Processing), 텍스트 검출(Text Detection), 성분 정보 검색(Ingredient Information Retrieval) 구조로 이루어져 동작한다.

1) Image Processing

Figure 1과 같이 굵은 테두리를 활용하여 이미지에서 Box Detection하며, 유효성분이 포함된 영역을 찾아 텍스트 인식을 진행할 관심영역으로 지정한다.

2) Text Detection by Tesseract-OCR

관심영역으로 지정된 영역에서 텍스트 추출을 위해 Tesseract-OCR을 이용한다. 2개의 OCR 엔진 중 LSTM 엔진에서 한글 사용을 위해 옵션 oem를 2로 지정하였다. 또한 페이지 세그멘테이션 모드를 지정하면서 한글 인식률을 높이기 위해 옵션 psm를 4로 지정하여 텍스트를 검출했다.

3) Ingredient Information Retrieval

검출된 텍스트인 약품 성분을 요청변수로 사용해 의약품 안전사용 서비스(DUR) API(식품의약품안전처 제공)에 성분 정보를 요청한다. 현재 API에 해당 성분 정보가 있을 경우 관련 정보를 저장하고, API에 없거나 검색이 되지 않는 성분은 대체하여 구매 가능한 약품 정보만 출력한다, 대체 약품의 경우 네이버 의약품 사전의 검색 결과 중 약품명만 저장하도록 하였다. 이 정보들은 Figure 3 (검출 결과 화면 일부)와 같은 형태의 텍스트 파일로 저장된다.

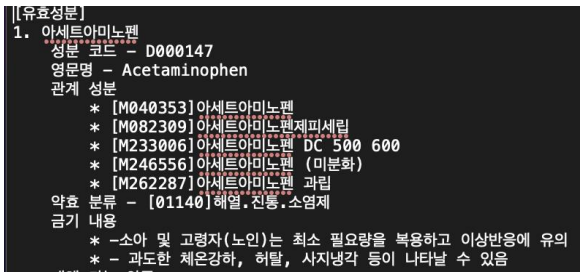

Fig. 3. Example of Ingredient Information Retrieval

III. Conclusions

본 논문에서는 이미지 프로세싱과 OCR 엔진을 이용하여 약 정보 이미지에서 성분들의 상세 정보를 검색해 출력하는 시스템을 제안하고 개발하였다. 이 시스템은 전문가가 아닌 소비자에게 의약품의 효율적인 구매와 안전한 복용, 또 의약품 성분에 대한 정보 전달이 가능하였다. 테스트 결과 약품 성분 표준 서식 기본형을 따를 때 텍스트 인식의 정확도가 약 65% 정도이나 그 외의 경우(단순형을 따를 경우) 각 성분의 구분이 명확하지 않아 텍스트 인식의 정확도가 약 50%로 떨어지게 된다. OCR의 한글 인식률과 특수문자 및 영문과 숫자가 섞어서 출력되는 경우의 인식률을 높인다면 기존보다 더 높은 성능 향상을 이룰 것으로 예상된다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단 4단계 두뇌한국(BK)21 사업 대학원 혁신 지원을 받아 수행된 연구임.

REFERENCES

[1] Gyu-Cheol Lee, and Jisang Yoo. “Development an Android Based OCR Application for Hangul Food Menu,” Journal of the Korea Institute of Information and Communication Engineering v.21 no.5 , pp.951 - 959, May 2017.

[2] Sun-Woo Park. "A Study on the OCR of Korean Sentence Using DeepLearning," IEEE Trans. on Parallel and Distributed Systems, Vol. 16, No. 3, pp. 219-232, March 2005.

[3] Ga-Hyeon Kang, and Ji-Hyun Ko, and Yong-Jun Kwon, and Na-Young Kwon, and Seok-Ju Koh. “A Study on Improvement of Korean OCR AccuracyUsing Deep Learning,” Journal of the Korea Institute of Information and Communication Engineering pp.693 -695, May 2018

[4] https://www.data.go.kr/data/15056780/openapi.do