온톨로지를 이용한 tesseract 기반의 OCR 모델 인식률 향상에 관한 연구

황치곤[1] · 윤대열[1] · 윤창표[2,*]

[1]광운대학교 · [2]경기과학기술대학교

A Study on the Improvement of Tesseract-based OCR Model Recognition Rate using Ontology

Chi-gon Hwang[1] · Yun Dai Yeol[1] · Chang-Pyo Yoon[2,*]

[1]Kwangwoon University · [2]GyeongGi University of Science and Technology

E-mail : duck1052@kw.ac.kr / hibig10@kw.ac.kr / cpyoon@gtec.ac.kr

요 약

기계학습의 발전에 따라 다양한 분야에 인공지능 기법이 적용되고 있다. 이 분야 중 이미지에 있는 문자를 텍스트로 변환하는 OCR 기법이 있다. HP에서 개발된 tesseract는 그 기법의 하나다. 그러나 이미지의 문자를 인식하는 인식률이 아직은 낮다. 이를 위해 본 연구에서는 온톨로지를 이용하여 문맥을 인지시키는 후처리 과정을 통해서 이미지의 문자 변환율에 향상을 기하고자 한다.

ABSTRACT

With the development of machine learning, artificial intelligence techniques are being applied in various fields. Among these fields, there is an OCR technique that converts characters in images into text. The tesseract developed by HP is one of those techniques. However, the recognition rate for recognizing characters in images is still low. To this end, we try to improve the conversion rate of the text of the image through the post-processing process that recognizes the context using the ontology.

키워드

OCR(Optical Character Recognition), Tesseract, Machine Learning, Ontology

Ⅰ. 서 론

광학 문자 인식(Optical Character Recognition, OCR)은 특정 문자의 인식이 컴퓨터 과학 분야에 적용되며, 이 기술은 텍스트의 디지털 이미지에서 인쇄된 문자와 손으로 쓴 문자를 구분하는 것으로 알려져 있다. OCR은 문서의 텍스트 인식하여 디지털 처리가 가능한 텍스트로 변환하는 것이다. 이 기술의 세 가지 기본 원칙은 무결성, 목적성 그리고 유연성으로 알려져 있다[1].

테서렉트(Tesseract)는 유명한 OCR 엔진으로서, 레이 스미스가 HP 연구소에서 1985년부터 1995년까지 개발했다[4]. 이 엔진은 60 이상의 언어와 서로 다른 영상 포맷을 지원한다. 그 후 문장의 인식률을 개선하기 위한 노력이 있었고, 2006년부터는 구글이 많은 부분을 개선했다. 현재는 정확한 오픈소스 광학 문자 인식 엔진 중 하나이다[2].

테서렉트 엔진은 UTF8을 지원하며 많은 언어를 인식할 수 있으며 더 많은 언어에 대한 지원이 계속 증가하고 있고, 학습할 수 있다. 즉, 일반적으로 지원되지 않는 새로운 언어나 글꼴을 학습하고 인식할 수 있다[3][4]. 그러나 한글은 자음과 모음의

* corresponding author

조합으로 구성되고, 수천 개 이상의 단어를 사용한다. 이로 인해 테서렉트는 문장부호와 문자의 오인되거나 누락되는 문제와 다른 언어, 기호, 숫자가 혼합된 문서일 경우 인식률이 떨어진다.

이에 본 논문에서는 문자 인식률을 올리기 위해 온톨로지[5]를 이용하여 문자 인식률을 높이는 방안을 제시한다. 적용하는 온톨로지는 단어사전의 역할을 수행하기 위한 기술로 문장에서 단어의 관계분석을 통해 정확도를 향상한다.

II. 테서렉트 아키텍처

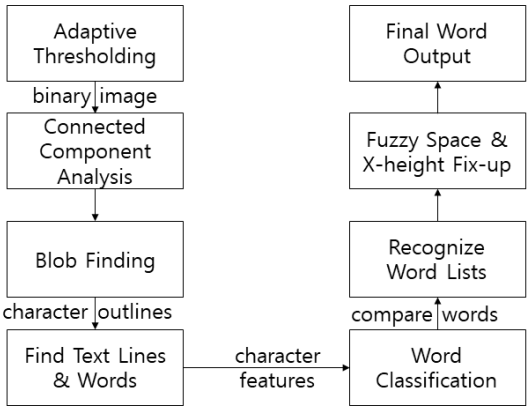

그림 1. Block Diagram of Tesseract OCR Engine Architecture

테서렉트는 그림 1과 같이 단계별 블록 다이어그램으로 표현된다. 첫 단계는 이미지 전처리로 적응적인 임계 값으로 수행된 이진 이미지를 생성하여 입력된다. 두 번째 단계는 연결된 요소들의 분석과 blob 찾기를 통하여 문자들의 윤곽선을 제공하고, 셋째 단계는 텍스트의 라인에 인접한 문자의 수직 중첩을 통해 라인을 감지하고, 문자 자르기와 문자 연결을 위한 단어들의 윤곽선을 통해 단어들을 구성하기 위한 문자의 특징을 파악한다. 넷째 단계는 단어 인식을 위해 클러스터링 및 분류 방법을 사용하여 단어를 인식하여 인식된 단어리스트를 산출하고, 마지막으로 인식된 단어리스트는 최종 퍼지 공간을 통해 단어를 확정하는 단계로 진행하는 구조를 가진다[3][4]. 이것은 학습을 통해 텍스트의 인식률을 향상할 수 있다.

III. 제안시스템

제안시스템은 OCR 기법의 인식률을 올리기 위한 연구로써 그림 2와 같은 구조를 가진다. 그림 2는 제안하는 시스템의 블록 다이어그램으로 테서렉트 아키텍처의 산출물에 대한 후처리 과정을 나타낸다. 산출물에 온톨로지를 제공하여 출력물의 정제과정을 통해 온톨로지를 갱신함으로써 출력물의 정확도를 높일 수 있다.

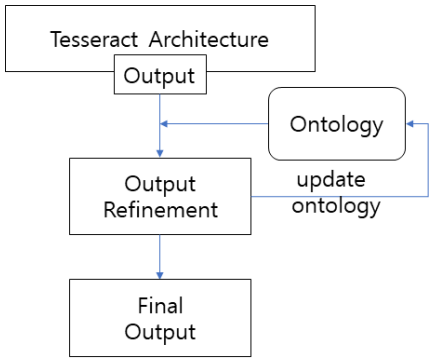

그림 2. Block Diagram of the Proposal System

이를 테스트하기 위한 실험으로 한글, 숫자, 영문자, 부호가 포함된 이미지를 이용하여 테스트한 결과 기존의 기법만으로 수행했을 경우 96.37%의 정확도가 나왔으며, 제안시스템을 적용한 경우 1.12% 정도의 향상을 볼 수 있었다. 이때 발생하는 문제는 다음과 같다.

- 하나의 단어인데 줄이 변경됨으로써 단어로 인식하지 못하고 단일문자로 인식돼 다른 문자로 인식되는 경우.
  예: “아”가 “oF”로 인식.
- 문자의 크기 차이에 의한 발생하는 경우.
- 단어에 숫자, 한글, 영문이 혼용된 경우.
  예: “15분”이 “15H”로 인식, “tion에”이 “tionOl|”로 인식.
- 한글 자체에서 발생하는 경우.

IV. 결 론

텍서렉트는 OCR로 이미지 형태의 문서에서 텍스트를 인식하여 문서로 생성하는 데 목적이 있다. 이것은 기본적으로 영어를 위해 만들어졌으며, 현재 다양한 언어들을 인식하기 위해서 확장되고 있다. 그러나, 아직 한글, 한자와 같은 문자의 조합이나 형성 문자와 같은 부분은 아직 부족하다. 이에 한글의 인식률 향상을 위해 온톨로지를 이용하는 방법론에 관해 연구하였다. 향후 이를 구현하여 정확도의 향상을 위한 연구할 예정이다.

References

[1] P. Divya, M. Varma, U. R. Mouli, Srinivas, Garima, Nikhil and Vishistha, “Web based optical character recognition application using flask and tesseract”, *Materials Today: Proceedings*, pp. 1-4, Jan. 2021.

[2] N. W. Kim and C. W. Hur, “Study on Performance Evaluation of Automatic license plate recognition”, Journal of the Korea Institute of Information and Communication Engineering, Vol. 20, No. 6 : 1209~1214 Jun. 2016.
[3] R. Smith, D. Antonova and D. S. Lee, “Adapting the Tesseract Open Source OCR Engine for Multilingual OCR.” *Proceedings of the International Workshop on Multilingual OCR*, No. 1, pp. 1-8, Jul. 2009.
[4] M. Gjoreski, G. Zajkovski, A. Bogatinov, G. Madjarov, D. Gjorgjevikj, and H. Gjoreski, “Optical character recognition applied on receipts printed in Macedonian Language”, *International Conference on Informatics and Information Technologies (CIIT)*, pp. 59-62, Apr. 2014.
[5] C. G. Hwang and C. P. Yoon, “Pre-processing Method of Raw Data Based on Ontology for Machine Learning”, Journal of the Korea Institute of Information and Communication Engineering, Vol. 24, No. 5: 600~608, May. 2020.