Journal of Korea Society of Industrial Information Systems Vol.xx No.x, Xxx. 20xx :1-4
http://dx.doi.org/10.9723/jksiis.20xx.xx.x.xxx ISSN:1229-3741

특정 양식 문서 촬영 이미지에서 한글 텍스트를 추출하는 알고리즘 연구

(Algorithm for Extracting Hangeul Text from Pictures in a Specific Format)

민강현[1)]
(Kang Hyeon Min)

요 약 본 논문에서 특정 양식으로 구성된 문서를 촬영한 사진에서 한글 텍스트를 추출하는 알고리즘에 대한 연구를 진행한다. 촬영한 대상은 수기로 작성된 문서가 아닌, 디지털 문서라 가정한다. -- 논문 완성 후 작성 --

핵심주제어: 텍스트 추출 알고리즘, Hough Transform, PyTesseract, 특정 양식, 진단서

Abstract We

Keywords: Aut

1. 서 론

촬영한 대상에서 텍스트를 추출하는 알고리즘에 대한 연구는 이미 오래전부터 진행되어 왔다. 하지만 아직까지 어떠한 환경에도 구애받지 않고 필요한 텍스트를 전부 추출하는 완벽한 알고리즘은 완성되지 않았다. 실제로 대기업과 국가적 차원에서 사용하는 카드 인식 프로그램같은 경우에도, 사진을 찍을 때, 사용자 인터페이스에서 [ ]와 같은 특정 규격에 맞춰 대상을 촬영할 것을 요구하며, 텍스트를 제대로 인식하지 못해 사용자에게 직접 입력을 추가적으로 요구한다. 또한 기관에 문서를 제출하는 경우, 사람이 직접 파일 혹은 사진에서 키워드들을 일일이 옮겨 적어야 한다. 본 연구에서는 사용자와 관리자 모두에게 비효율적인 이러한 상황을 해결하고자 한다. 특히 이런 상황들 중 동물병원에서 받을 수 있는 진단서에서 '병명, 증상, 반려견 이름, etc'와 같은 각각의 키에 해당되는 값을 인식 및 추출하는 것을 목표로 한다.

기존의 문자 인식 연구의 대부분은 영어나 숫자 인식에 국한되어 있었다. 사진에서 한글을 인식하

* Corresponding Author: hoihoimkh@gmail.comr
+ 이 논문은 2024년 ...
1) 동국대학교 컴퓨터 공학과, 제1저자 민강현
2) OO대학교 OO학과, 제2저자

는 알고리즘과 딥러닝 연구는 꾸준히 진행되고 있지만, 현실적인 상황에서 한글을 인식조차 하지 못하는 경우도 빈번하며, 디지털 문서를 촬영하는 경우 글꼴에 따라 인식률이 큰 차이를 보이기도 한다. 또한 사진에서 특정 키와 매칭되는 값을 추출하는 연구도 이미 존재하지만 제한 요소가 존재한다. 대표적으로 TRIE (Text Reading and Information Extraction)와 KVPFormer (Key Value Pair Former)에서 확인할 수 있듯이, 두 연구에서 기본적으로 다룬 데이터들은 촬영한 대상이 아닌 문서 파일 그 자체를 대상으로 진행하였다.

따라서 본 연구는 Tesseract OCR 엔진에서 지원하는 한글 텍스트 인식에서 더 높은 정확도를 보일 수 있도록 이미지 전처리화 알고리즘과 사진에 존재하는 텍스트에서 key-value가 잘 맞도록 데이터를 추출하는 알고리즘을 소개한다. 본 알고리즘에서는 이미지 각도 조정, 텍스트가 존재하는 부분만을 Object화 및 그림자 무시를 지원하여 더 높은 텍스트 인식률을 제공한다.

2. 알고리즘 구성 및 실험

2.1 알고리즘 구성

본 연구의 메인 알고리즘을 Extract Data from Diagnosis라 명명하며, 'Fig. 1'에서 전체적인 시스템의 개략적인 상황을 보여주고 있다. Tesseract 엔진을 활용하므로 본 알고리즘은 파이썬을 기반으로 작성하였고 3개의 서브 알고리즘으로 확인된 문제를 보완하고자 한다. 이는

Table 1. Problems in Taking Pictures

Problems	Percentage (%)
Tilted Image	15.0
Background Objects	30.0
Shadowed Image	50.0
etc	5.0
Total	100.0

사용자가 촬영한 사진을 Tesseract를 사용하여 한글을 인식시킬 때 나타나는 문제를 세 가지 경우로 구분 가능하기 때문이며, 각각의 문제는 다음과 같다. 사진의 기운 정도로 인해 생기는 문제, 배경과 진단서의 구분 문제 그리고 그림자진 부분의 인식 저하로 발생하는 문제이다. 'Table 1.'은 텍스트 인식을 위해 실제 사진을 촬영하며 생기는 문제를 분류한 분포표이다. 진단서와 같은 문서를 촬영대상으로 삼을 때, 문서의 텍스트 부분이 잘 나오도록 빛을 등지고 촬영하는 경우가 많다. 그로 인해 특정 부분에 촬영자의 손으로 인한 그림자가 지는 경우가 많이 발생하는 것을 확인할 수 있으며, 책상 혹은 다른 물체에 대고 사진을 촬영하는 경우 다른 물체가 사진에 존재하는 경우가 생기며, 이는 Tesseract로 인한 텍스트 인식 퀄리티에 문제를 발생시킨다. 사진을 기울여 촬영하는 경우도 종종 발생하며 특히 한손으로 촬영하는 경우 특히 자주 발생하였다. 'Fig. 2'는 각 문제의 대표적인 경우를 보이며, 그 중 'Base'는 텍스트 인식 결과 가장 높은 텍스트 인식률을 보이는 경우 중 하나이며, 'Table 2.'은 원본 텍스트 데이터와 비교하여 추출한 텍스트가 얼마나 정확한지를

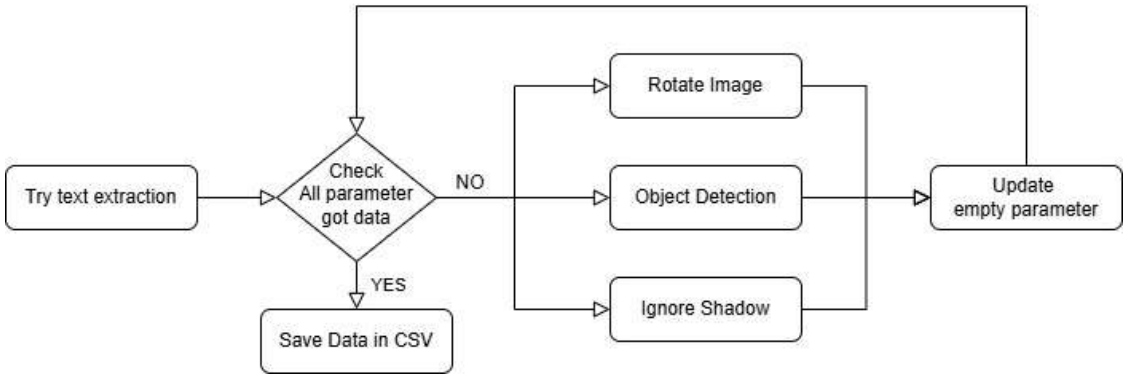


Fig. 1 Algorithm Flowchart

http://dx.doi.org/10.9723/jksiis.20xx.xx.x.xxx ISSN:1229-3741

보인다. 이미지에서 100% 완벽한 데이터가 추출되지 않는다는 분명한 한계를 고려해야 하며, 그 예로써 발생하는 다음과 같은 상황을 예시로 'Table 2.'의 각 패러미터들이 어떠한 메커니즘을 통해 산출되었는지 설명한다. 'key-value' 의 예시로 원본 데이터가 "성명": "한지원" 이고, 추출한 텍스트가 "성명": ". 한지원 인)" 인 경우, 원본과 완전 일치하지만 의미있는 데이터 '한지원'이 존재하므로 추출한 텍스트 데이터에 유의미해 보이는 데이터만을 추출하도록 한다.

값 낮은 이유 설명 한글 인식률과 Metric 패러미터 의미 설명 사용된 알고리즘 설명

Table 2. Initial Text Extraction Analysis

Metric	Value (%)
Weighted Similarity Accuracy	53.6
Weighted Similarity Precision	100.0
Weighted Similarity Recall	49.1
Weighted Similarity F1 Score	65.8

첫 번째, 서브 알고리즘은 rotate image로 명명하였으며, 촬영한 사진이 컴퓨터에서 인식하는 수평을 기준으로 어느 정도 기울어져 있는지 판단하고 최대한 수평에 가깝게 이미지를 재구성하는 알고리즘이다.

-세부 프로세스 설명 작성 예정입니다.-

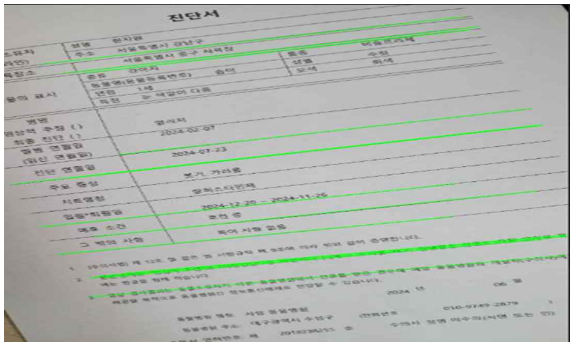


Fig. 3 Line Detection for Calculating Degree of Tilt

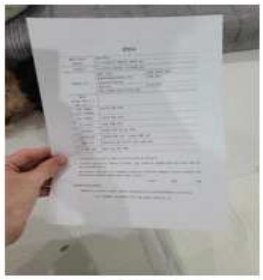
두 번째, 서브 알고리즘은 find objects로 명명하였으며, 촬영한 사진에서 배경과 다른 물체를 무시하고 진단서에 텍스트가 존재하는 부분만을 추출하는 알고리즘이다.

-세부 프로세스 설명 작성 예정입니다.-


Base


Tilted


Shadowed


Other Objects

Metric	Value (%)		
	Tilted	Shadowed	Other Objects
Weighted Similarity Accuracy	0.0	12.7	44.2
Weighted Similarity Precision	0.0	60.0	100.0
Weighted Similarity Recall	0.0	15.0	37.1
Weighted Similarity F1 Score	0.0	13.8	50.0

Fig. 2 Text Extraction Analysis for Each Cases

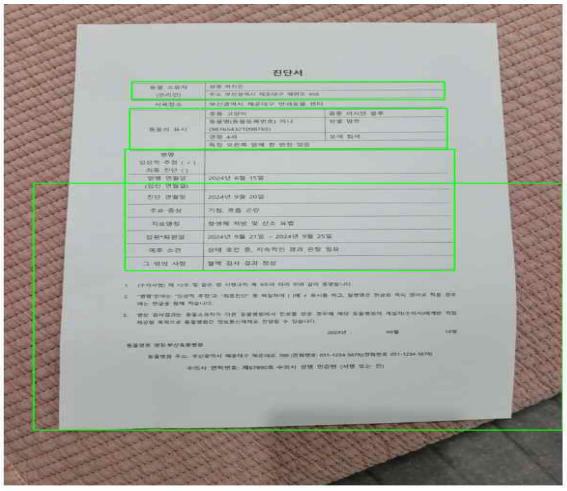


Fig. 4 Finding Tables

세 번째, 서브 알고리즘은 ignore shaodw로 명명하였으며, 촬영한 사진에서 음영진 부분에서 인식저하가 일어나는 Tesseract의 문제를 보완하는 알고리즘이다.

-세부 프로세스 설명 작성 예정입니다.-

메인 프로세스 설명 및 정리
추출한 텍스트를 어떻게 정확한 키-값 매치를 시킬 것인지 에 대한 알고리즘 설명 작성

3. 실험결과

3.1 각 상황별 텍스트 인식 결과

세부 프로세스 실행 결과 설명

Table 3. Text Extraction Analysis after Image Rotation

Metric	Value (%)
Weighted Similarity Accuracy	53.6
Weighted Similarity Precision	100.0
Weighted Similarity Recall	49.1
Weighted Similarity F1 Score	65.8

Table 4. Text Extraction Analysis after Object Detection

Metric	Value (%)
Weighted Similarity Accuracy	53.6
Weighted Similarity Precision	100.0
Weighted Similarity Recall	49.1
Weighted Similarity F1 Score	65.8

Table 5. Text Extraction Analysis after Shadow Ignore

Metric	Value (%)
Weighted Similarity Accuracy	53.6
Weighted Similarity Precision	100.0
Weighted Similarity Recall	49.1
Weighted Similarity F1 Score	65.8

3.2 결과분석

각 프로세스 실행 후 얼마나 향상되었는지 확인 및 분석

Table 6. Performance Improvement Analysis

Metric	Value (%)	
	Before	After
Weighted Similarity Accuracy	13.1	17.9
Weighted Similarity Precision	0.0	100.0
Weighted Similarity Recall	0.0	5.5
Weighted Similarity F1 Score	0.0	0.12

4. 결 론

기술적 한계 및 발전 사항 언급.

References

[국내논문] 딥 러닝 기법을 활용한 이미지 내 한글 텍스트 인식에 관한 연구

Research on Korea Text Recognition in Images Using Deep Learning, https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202018955008809

TRIE (Text Reading and Information Extraction) https://ar5iv.labs.arxiv.org/html/2005.13118

KVPFormer (Key Value Pair Former) https://ar5iv.labs.arxiv.org/html/2304.07957

* 인터넷 사이트의 기사내용은 가급적 인용하지 마시길 바랍니다.
* 본문에 인용된 모든 문헌들을 참고문헌 목록에 작성하시길 바랍니다.
* 참고문헌은 알파벳 순서대로 양식에 맞추어 작성하시길 바랍니다.
* 참고문헌은 모두 영문으로 작성하세요.

사 진
24 x 30

홍 길 동 (GilDong Hong)

- 정회원
- ○○대학교 ○○학과 ○○학사
- ○○대학교 ○○학과 ○○석사
- ○○대학교 ○○학과 ○○박사
- (현재) ○○대학교 ○○대학 ○○학과 ○교수

- 관심분야: 정보시스템 성과, e-Learning, 온라인 커뮤니티