Journal of the Korea Academia-Industrial cooperation Society
Vol. 21, No. 5 pp. 451-457, 2020

https://doi.org/10.5762/KAIS.2020.21.5.451
ISSN 1975-4701 / eISSN 2288-4688

전자문서에서 서식인식과 광학문자인식을 이용한 개인정보 탐지 및 보호 시스템

백종경[1], 지윤석[2], 박재표[3*]
[1]숭실대학교 대학원 컴퓨터학과, [2]숭실대학교 대학원 IT정책경영학과, [3]숭실대학교 정보과학대학원

A Personal Information Security System using Form Recognition and Optical Character Recognition in Electronic Documents

Jong-Kyung Baek[1], Yoon-Seok Jee[2], Jae-Pyo Park[3*]
[1]Division of Computer, Graduate school of Soongsil University
[2]Department of IT Policy Management, Graduate school of Soongsil University
[3]Graduate School of Information Science, Soongsil University

요 약 전자문서에서 개인정보를 보호하기 위한 방법으로 서식 인식과 광학 문자 인식 기법이 많이 이용되고 있으나 OCR 엔진의 저조한 인식률로 인해서 개인정보를 탐지하지 못하거나 오탐이 많이 발생하고 있고 또한 대량의 전자문서를 분석하는데도 오랜 시간이 걸린다. 본 논문에서는 기존의 방법을 개선하여 전자문서의 이미지 분석 속도와 OCR엔진의 글자 인식률, 그리고 개인정보의 탐지율을 향상할 수 있는 방안을 제시한다. 서식 인식 방법을 이용하여 분석 속도를 높이고, 이미지 보정을 통해 OCR 엔진 분석 속도 및 글자 인식률을 향상한다. 이미지에서의 개인정보 분석 알고리즘을 제안하여 개인정보의 탐지율을 높였다. 실험을 통하여 이미지 서식 인식 시료 1755개를 분석하여 평균 0.24초가 소요되어 기존의 PAID 시스템 서식 인식 방안보다 0.5초 향상되었으며 이미지 서식 인식률은 평균 99%를 기록하였다. 본 논문에서 제안한 방법은 전자문서에서 개인정보를 보호할 수 있는 시스템으로서 공공, 통신사, 금융, 관광, 보안 등 여러 분야에서 활용할 수 있을 것이다.

Abstract Format recognition and OCR techniques are widely used as methods for detecting and protecting personal information from electronic documents. However, due to the poor recognition rate of the OCR engine, personal information cannot be detected or false positives commonly occur. It also takes a long time to analyze a large amount of electronic documents. In this paper, we propose a method to improve the speed of image analysis of electronic documents, character recognition rate of the OCR engine, and detection rate of personal information by improving the existing method. The analysis speed was increased using the format recognition method while the analysis speed and character recognition rate of the OCR engine was improved by image correction. An algorithm for analyzing personal information from images was proposed to increase the reconnaissance rate of personal information. Through the experiments, 1755 image format recognition samples were analyzed in an average time of 0.24 seconds, which was 0.5 seconds higher than the conventional PAID system format recognition method, and the image recognition rate was 99%. The proposed method in this paper can be used in various fields such as public, telecommunications, finance, tourism, and security as a system to protect personal information in electronic documents.

Keywords : Classification, OCR, Image Correction, Personal Information, Security

*Corresponding Author : Jae-Pyo Park(Soongsil Univ.)
email: pjerry@ssu.ac.kr
Received February 4, 2020 Revised March 20, 2020
Accepted May 8, 2020 Published May 31, 2020

1. 서론

스마트워크와 클라우드의 기술 발달에 따라 언제, 어디서나 전자문서를 작성하고 편집할 수 있는 환경이 되었다. 이로 인해서 개인정보들이 문서에 포함되어 서버에 저장되어 해킹의 위험에 노출되어 있다. 기업에서는 이를 방지하기 위하여 개인정보보호 솔루션과 같은 보안 솔루션을 구축하여 운영하고 있으며, 한국인터넷진흥원에서는 공공사이트나 민간사이트에 대해서 개인정보를 검출하여 삭제하는 등의 조치를 취하고 있다. 하지만 현행 보안 솔루션들은 정해진 문서 포맷에 대해서만 식별하기 때문에 다른 형태로 기밀정보와 개인정보가 존재한다면 외부로 유출될 가능성이 크다. 특히, 이미지 파일 포맷 같은 경우 텍스트 형태로 존재하지 않기 때문에 보안 솔루션에서 원천적으로 차단하지 않은 이상 식별하기가 어렵다. 이미지를 검출하기 위해서는 OCR(Optical Character Reader) 기술을 이용하여 글자를 추출하는데 한글의 OCR 인식률 문제, 분석속도 문제, OCR 인식률에 따른 개인정보 오류 탐지 문제가 있어서 적용이 어려운 실정이다.

본 논문에서는 이미지 보정을 통한 OCR 인식률 향상을 통한 개인정보 검출률 향상, 이미지 서식인식을 활용한 개인정보 검출 방안을 제시한다.

2장에서는 이미지에 서식인식과 OCR기술, 그리고 개인정보의 검출방법을 서술하고 3장에서는 개인정보 검출 및 보호 시스템을 설계하고 4장에서는 검증을 위하여 구현 및 성능 평가를 수행하고 5장에서 결론 및 향후 연구 과제에 대하여 서술한다.

2. 관련연구

2.1 이미지 서식인식

이미지 서식인식은 서식의 종류에 따라 문서를 분류하는데 분류된 각각의 문서에서 필요한 영역의 데이터에 대해서만 연산하기 때문에 분석 및 처리가 빠르고, 서식의 특징이 되는 몇 가지 형태를 보고 판단함으로써 원본 데이터의 품질이 좋지 않더라도 우수한 판단결과를 보이며, 필요한 영역에 대해서만 고수준/고품질의 이미지 보정작업이 적용되어 처리시간 대비 검출결과가 우수하다. 또한, 새로운 서식에 대해서도 간단한 학습 과정을 통해 생성된 학습데이터를 추가함으로써 손쉽게 지원할 수 있다.

서식인식 알고리즘에는 픽셀 값의 제곱차를 이용하는 제곱차 매칭방법과 템플릿과 입력 영상의 곱을 제곱하여 더하는 상관관계 매칭방법이 있다[3]. 제곱차 매칭방법은 템플릿 T를 탐색 영역 I에서 이동시켜 가며 픽셀 간 차이의 값을 제곱하여 합계를 계산한다. 이때 완벽하게 일치하면 0을 반환하지만, 일치하지 않을수록 값이 커진다. 상관관계 방법은 완벽하게 일치하면 큰 값이 나오고, 일치하지 않을수록 작은 값이 나오거나 0이 나온다. 본 논문에서는 상관관계 매칭 방법을 개선하여 적용한다.

2.2 OCR 인식기술

OCR이란 이미지 스캔으로 얻을 수 있는 문서의 글자 이미지를 컴퓨터가 편집 가능한 글자형식으로 변환하는 기술이다[19]. OCR은 머신비전, 인공지능의 연구 분야로 시작되었다. 렌즈, 거울 등의 광학 기술을 이용한 광학문자인식과 스캐너 또는 알고리즘에 의한 문자 인식은 다른 영역으로 생각되었으나 현 시대에서는 광학문자인식이라는 말이 디지털 문자 인식을 포함한다.[4].

2.3 개인정보 검출 방법

개인정보를 검출하는 방법으로는 가장 많이 이용되는 것은 정규 표현식을 사용하는 방법, 패턴매칭을 사용하는 방법이 있다[5]. 정규 표현식 방법은 순차적으로 정규식과 비교하여 검출하는 방식이고, 패턴매칭 방법은 텍스트 문자열 전체를 하나의 패턴으로 규정하고 텍스트 전체를 비교하는 방법이다 본 논문에서는 패턴매칭 방법을 이용하여 개인정보를 검출한다. 패턴매칭 방법은 개인정보 자릿수, 띄어쓰기, 특수기호 등의 패턴을 이용하여 매칭이 되는지를 먼저 검출하고 체크디지트 단계에서 검출하는 방법이다.

2.4 이미지에서의 각도산출 방안

이미지에서의 기울기의 각도를 산출하는 방안은 Fig. 1과 같다.

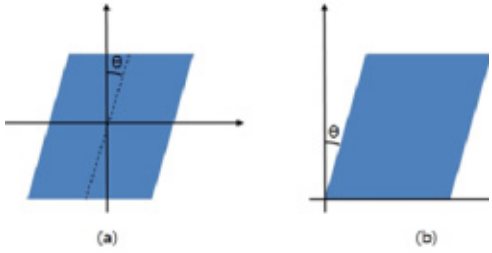

Fig. 1. Image rotation angle calculation method
(a) Corner based angle calculation
(b) Center based angle calculation

(a)는 센터 기반으로 각도를 산출하는 방법을 나타낸 것이고, (b)는 코너 기반으로 각도를 산출하는 방법을 나타낸 것이다. (a) 및 (b)를 참조하면, 이미지 문서의 중심 또는 코너를 기준으로 하는 x-y 좌표에서 y축과 이미지 문서의 변이 이루는 각도를 산출할 수 있다.

3. 이미지에서 개인정보 검출 및 보호 시스템의 설계

3.1 개인정보 검출 및 보호 시스템

본 논문에서는 전자문서에서 정형/비정형 이미지 내의 기밀정보(텍스트 형태로 존재하는 대외비 문자열)와 개인정보를 검출하기 위한 서식인식 기술, OCR 인식율을 높이기 위한 이미지 보정 기술, OCR을 이용한 데이터 추출기술, 이미지에서의 개인정보 검출 기술을 제안한다. 제안하는 전체 시스템 구성도는 Fig. 2과 같다.

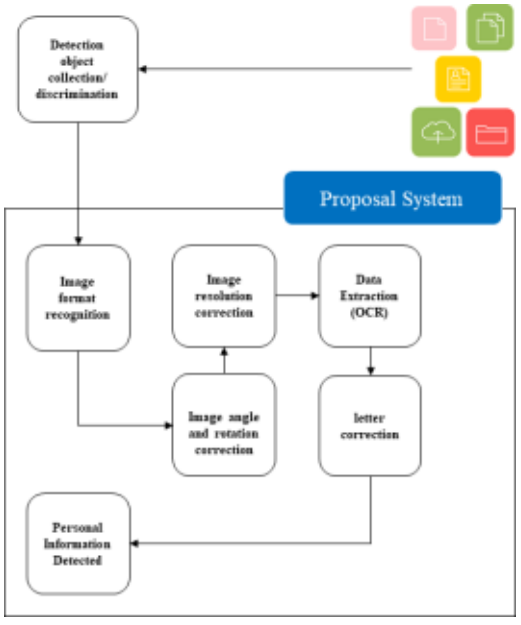

Fig. 2. Proposed System Overall Diagram

문서 내 이미지, 정형/ 비정형 이미지가 수집 및 식별이 되면 먼저 이미지에 대해 서식인식을 한다. 서식인식이 안 된 이미지는 OCR을 위한 보정작업을 한다. 이미지를 색상보정을 한 후 이미지의 왜곡도를 조사하여 회전 및 각도 보정을 진행한다. 이미지 색이 선명해짐으로써 왜곡 도에 대한 수치가 정확히 계산될 수 있다. OCR 추출 시 글자의 크기에 영향을 받기 때문에 이미지에 대해 A4 기준으로 해상도 보정 후 OCR 인식을 진행한다.

OCR 인식 시 개인정보가 아님에도 개인정보로 잘못 탐지하는 부분을 수정하기 위해 별도의 글자보정 작업을 진행 후 개인정보가 있는지 탐지한다.

3.2 이미지 서식인식 방안

제안시스템에서 문서 내 이미지 또는 정형/ 비정형 문서가 검색이나 추출이 되면 이미지 문서에 대한 서식인식을 진행한다.

이미 학습된 데이터를 가지고 먼저 이미지 서식인식을 해서 개인정보가 포함된 이미지이면 개인정보 영역을 검출한다. 이미지 보정이나 OCR 과정을 거치지 않기 때문에 빠른 속도로 다음 이미지를 검색할 수 있다. 이미지 서식인식에서 영상 밝기 그대로 패턴매칭을 하여 서식에 대한 유사도 값을 도출할 수 있으나 패턴 매칭 탐지율을 올리기 위해 영상의 밝기에 의해 민감하므로 매칭 전 정규화 과정이 필요하다. 이미지 서식인식 과정은 Fig. 3과 같다.

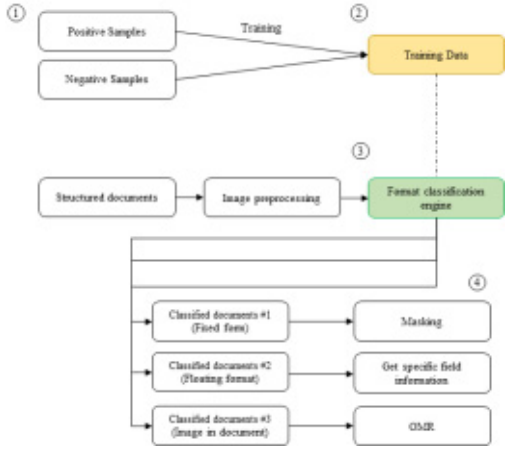

Fig. 3. Image Classification Processing

① 이미지 서식에 매칭되는 긍정적인 샘플 군과 잘못 매칭이 된 이미지 또는 매칭되는 서식과 상관없는 부정적인 샘플 군을 입력한다.

② 입력된 샘플데이터 근거로 이미지 학습을 통해 학습 데이터파일을 만든다. 샘플 수에 비례하여 인식률이 높아진다.

③ 서식분류 엔진에 서식데이터 파일을 탑재 후 다중

이미지를 가질 수 있는 PDF나 TIFF 문서면 300DPI jpg 형태로 변환하고, 단일 이미지면 200DPI로 변환한다.

④ 서식분류 환경 파일 옵션에 따라 해상도를 변경한다.

⑤ 학습데이터 이미지와 이미지 서식인식 대상 이미지를 그레이 형태로 이미지를 전처리한다. 문자, 선, 객체 등 문서 내 특징들을 분석하여 유사도가 높은 서식으로 자동 분류한다.

⑥ 분류된 서식에 따라 특정 데이터 취득, 마스크 데이터 가공 등 후처리를 진행한다.

개인정보의 위치에 따라 고정서식, 유동 서식, 문서 내 이미지(OMR 처리)로 정의한다.

고정서식은 기본증명서, 인감증명서 등 개인정보 위치가 고정된 서식이고, 유동 서식은 가족관계증명서, 주민등록등본 등 개인정보 위치가 세대 구성에 따라 유동적인 서식이다. 문서 내 이미지 서식은 문서 내 신분증, 통장사본 등 문서 안에 스캔 이미지가 첨부된 서식을 의미한다.

고정서식의 경우 지정 위치에 마스크를 하고, 유동 서식이면 세대 구성원의 이미지 색상을 비교하여 유동적으로 마스킹을 한다. 문서 내 이미지면 문서 페이지를 이미지로 변환 후 신분증, 통장사본 등을 패턴 매칭하여 해당 이미지를 추출 후 마스크 하도록 구성한다.

이미지 서식인식 방법은 왼쪽에서 오른쪽으로 위에서 아래로 픽셀 단위로 학습데이터와 매칭한다. 유사도를 정할 때는 상관계수 방법을 적용하여 검칠된 수치를 정규화 계수로 나누어 정한다. 수식은 Eq. (1)과 같다.

$$R(x,y)=\frac{\sum_{x',y'}^{(T'(x',y')\cdot I'(x+x',y+y'))}}{\sqrt{\sum_{x',y'}^{T'(x',y')^2}\cdot\sum_{x'y'}^{I'(x+x',y+y')^2}}} \quad (1)$$

3.3 이미지 외곡도 보정 방안

이미지 문서의 센터 기반 각도를 산출하는 기법과 코너 기반으로 각도를 연산하는 알고리즘을 병행해서 사용한다. 코너 기반으로 각도를 산출할 경우 회전으로 이미지가 손상이 심하므로 센터 기반으로 각도를 산출한다. 1.5° 이상 또는 -1.5° 미만이면 센터 기반으로 산출된 각도로 각도 보정을 한다. 센터 기반으로 산출된 값이 1.5°~-1.5° 사이 값일 경우에는 코너 기반으로 각도를 산출하여 각도 보정을 한다. 센터 기반은 중심점을 찾아 산출하기 때문에 각도가 조금 왜곡되었을 경우 센터를 찾기가 쉽지 않다. 코너 기반은 전체 이미지의 외곽선에 대해 각도를 산출하기 때문에 작은 왜곡도 에는 각도 산출 정확도가 높다. 산출된 각도가 0.08°~-0.08° 사이 값에 대해서는 보정을 하지 않는다. 오히려 작은 각도의 각도 보정을 통해 글자의 폰트 형태가 훼손되면서 OCR 인식률에 영향을 줄 수 있다.

3.4 이미지에서의 개인정보 검출 방안

개인정보 검출 엔진에서는 텍스트로 보정된 데이터를 받아서 개인정보 검출을 진행한다. OCR 엔진을 통해 오탐률이 있는 개인정보에 대해 정확도를 가지는 검출 방안을 제시하며 구성은 Fig. 4와 같다.

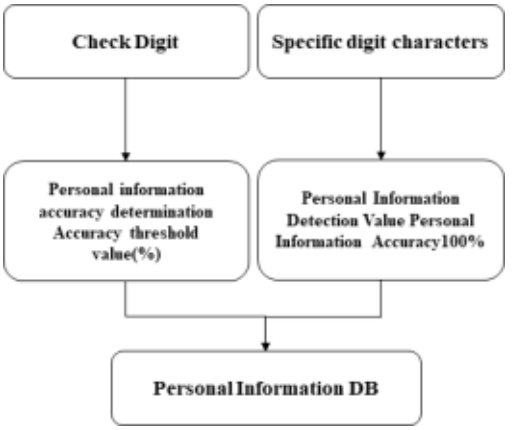

Fig. 4. Diagram of Personal Information Detection Engine in Image Document

이미지에서의 개인정보 검출 엔진은 한국인터넷진흥원에서 고시한 11종의 개인정보를 검출할 수 있도록 하였다. 개인정보 중에는 체크디지트를 포함하는 개인정보, 특정 자릿수에 의해 판별되는 개인정보가 있다. 또한, 특정 자릿수에 판별되는 개인정보 중에는 한글을 포함하는 개인정보, 한글을 포함하지 않는 개인정보로 나뉜다.

개인정보 검출 정책에 한글을 포함하지 않는 개인정보를 검출할 때는 OCR 엔진의 언어 팩을 영문, 숫자로 설정한 후 추출한다. 이는 영문이나 숫자가 조합하여 한글로 인식되는 경우를 방지하기 위함이다. 한글을 포함하는 개인정보면 먼저 OCR 엔진의 언어 팩을 영문, 숫자로 개인정보 영역을 인식하고 한글이 필요한 영역만을 한글 언어 팩으로 설정한 후 글자를 검출한다. 단, 특정 한글 단어만 검출이 필요한 경우에는 한글 언어 팩으로만 설

정한 후 OCR 엔진에 삽입한다.

개인정보가 체크디지트나 특정 자릿수에 검출된 개인정보는 정확도 100%의 값으로 데이터베이스에 저장하고, 개인정보 검출 엔진에서 매칭이 실패된 개인정보는 개인정보판별 정확도 알고리즘을 통해 개인정보를 다시 검사하고 개인정보 정확도 임계치 이상의 개인정보만을 데이터베이스에 정확도 수치와 함께 저장한다. 미검출 된 개인정보 정확도를 판별하는 알고리즘은 Eq. (2)과 같다.

$$Z=(\sum_{i=1}^{k}x^{k}\cdot\ \frac{c}{100})+(\sum_{j=1}^{l}y^{l}\cdot\ (\frac{c}{100}\cdot\ g)) \qquad (2)$$

개인정보 검출 엔진에서는 “100”에서 개인 정보패턴 글자 수를 나누어 글자 1점당 점수를 구한다. 이는 숫자로 검출된 한 글자의 점수가 되며, 여기에 g를 곱하게 되면 영문자 점수가 된다. 영문자에 점수를 주는 이유는 숫자 “1”이 “l”(영문자 L의 소문자)로 추출될 수 있고, 숫자 “0”이 “o”(영문자 L의 소문자)나 “O”(영문자 L의 대문자)로 표현될 수 있기 때문이다. 백분율로 환산된 글자마다 점수를 합하여 개인정보 검출 임계치보다 크면 개인정보로 인식한다.

4. 비교분석 및 성능평가

4.1 시험환경 및 방법

전자문서에서 이미지 서식인식과 광학문자인식을 이용한 개인정보를 보호하기 위한 시험환경은 Table 1과 같이 구성하였다.

Table 1. Proposed System Test Environment

Division	Explanation
Format Recognition Similarity	over 90%
Privacy accuracy	over 90%
DPI	72DPI, 96DPI, 150DPI, 200DPI, 300DPI
Number of Classification Recognition Samples	1755 EA(3219 Page)
OCR sample card	12966 EA(Server 3EA)
Document Format	docx, doc, xlsx, xls, ppt, pptx, pdf, hwp, html, mht, jpg, jpeg, bmp, png, gif, tif, tiff

시험을 수행하는 동안 다른 프로세스에 의한 제안 프로그램 속도의 영향도를 줄이기 위해 3대를 PC를 준비하여 평균치를 냈다.

개인정보에 대한 검출 임계치는 서식인식 90%, OCR 글자 정확도 80%, 개인정보 정확도 90%로 설정하였다. 시료는 문서 포맷을 무작위로 웹 사이트 크롤링 문서, 공인기관 발급문서(주민등록등본, 가족증명서, 특허증 등), 신분증 사진, 일반문서, 금융문서로 무작위로 구성하였다.

서식인식 시료는 1755개 파일, 3219페이지, 신분증, 공문서, 금융문서 샘플 군으로 나누어 분석하였고, OCR 엔진 성능평가용 12966개의 시료는 각 서버 당 4322개를 설정하였다.

4.2 성능평가

4.2.1 서식인식엔진 성능평가

서식인식 엔진 시료 1755(3219페이지) 개를 A, B, C 샘플 군으로 나누어 성능평가를 실행하였다.

이미지 서식인식 처리시간은 작게는 0.06초부터 크게는 0.38초까지 분석시간이 걸렸으며, 서식인식시간이 평균치에서 크게 벗어나거나 인식시간이 특이하게 오래 걸리는 경우는 없었으며, OCR, 상용 OCR 엔진, PAID 시스템과 이미지 분석시간을 비교하였고, Table 2과 같다.

Table 2. Average Analysis Speed by Image Format Recognition Engine

Division	Time(Second)
Tesseract OCR[6]	2~15
Commercial OCR	1~8
PAID System[7]	0.8
Proposal System Recognition	Total Time : 790.8171 Tatal Page : 3219 Avg. 0.245672

기존 시스템 중에서 가장 우수했던 PAID 시스템 서식인식 방안보다 분석속도가 약 0.5초 향상되었다.

4.2.2 OCR엔진 성능평가

OCR 엔진 시료 12966개를 3대의 PC에 배포하고, 성능평가를 시행하였다. 서버별 OCR 엔진 장당 처리시간 평균은 Table 3과 같다.

제안시스템 3개 서버 평균 2.2초를 기록하여 상용 OCR보다 2.3초, Tesseract OCR보다 7.3초가 빠른 것으로 측정되었다.

Table 3. OCR Engine Character Data Analysis Time Comparison

Division	Average analysis time(Second)
Proposal System 1	2.2216
Proposal System 2	1.8653
Proposal System 3	2.0871
Proposal System Avg.	2.2046
Commercial OCR	1~8(4.5)
Tesseract OCR[6]	2~15(9.5)

4.3 비교분석

서식인식 엔진은 시료 1755(3219페이지) 개를 A, B, C 샘플 군으로 나누어 비교분석을 진행하였고, OCR 엔진은 시료 12966개를 대상으로 비교분석을 진행하였다.

4.3.1 서식인식엔진 비교분석

서식인식 엔진 샘플 타입별로 제안시스템 엔진, 한글 처방전 문자인식 시스템 엔진[3], KCR-AlexNet 엔진, PAID 시스템 엔진에 대해 인식률과 미탐률을 측정하였으며, 이미지 서식인식 인식률 및 미탐률 표는 Table 4과 같다.

Table 4. Recognition rate and undetected rate of image format recognition by sample

Division	Recognition rate	Not detected
A Type	99.2%	0.8%
B Type	99.7	0.3%
C Type	98.1	1.9%
Proposal System	99%	N/A
Hangul Prescription Character Recognition System[12]	83.1%	N/A
KCR-AlexNet[10]	86%	N/A
PAID System[7]	96%	N/A

A타입 시료 분석에서는 인식률 99.2%, 미탐률 0.8%를 기록 하였고, 미탐 된 이미지는 너무 밝게 스캔 되거나 어둡게 스캔 되어 윤곽을 찾지 못하는 경우 검출을 하지 못하였다.

B타입 시료 분석에서는 인식률 99.7%, 미탐률 0.3%를 기록 하였고, 미탐 된 이미지는 영문 주민등록등본, 패턴 인식한 부분에 신분증이 겹쳐있는 경우 검출을 하지 못하였다. 이는 추가 이미지 서식인식 학습을 통해 보안이 가능했다.

C타입 시료 분석에서는 인식률 99.8%, 오탐률 0.2%를 기록하였다. 금융서류에는 패턴 인식된 부분과 계약서 뒷면 서류와 패턴 유사도가 비슷하여 오탐이 발생되었다.

제안시스템 서식인식 엔진은 한글 처방전 문자인식 시스템 서식인식 엔진보다 15.9%, KCR-AlexNet 서식인식 엔진보다 13%, PAID 시스템 서식인식 엔진보다 3%의 인식률 향상을 보였다.

4.3.2 OCR엔진 비교분석

공개소스 기반 Tesserct OCR 엔진, 상용 OCR 엔진에 대해 DPI별로 영문/ 숫자, 국문 글자 인식률을 측정하고, 제안시스템 OCR 엔진, Tesserct OCR 엔진, 상용 OCR 엔진, 한글 처방전 문자인식 시스템 OCR 엔진, KCR-AlexNet OCR 엔진과 글자 인식률을 비교하였다.

OCR 엔진별 글자 인식률 비교 표는 Table 5와 같다.

Table 5. Comparison of Character Recognition Rate by OCR Engine

Division	Recognition rate
Tesseract OCR[6]	69%
Commercial OCR	92.5%
Hangul Prescription Character Recognition System[12]	83.21%
KCR-AlexNet OCR[10]	86%
Proposal System OCR	92.3%

제안시스템 OCR의 글자 인식률 대비 Tesseract OCR 23.3%, 한글 처방전 문자인식 시스템 OCR 9.09%, KCR-AlexNet OCR 6.3%로 글자 인식률이 향상되었다.

5. 결론

본 논문에서는 전자문서에서 이미지 서식인식, OCR 인식률을 향상하기 위한 이미지 보정, OCR 엔진 분석, 이미지에서의 개인정보 분석 방안, 이미지 서식인식과 OCR 글자 데이터 활용방안을 제시하였다.

이미지 서식인식 엔진을 이용하여 OCR 엔진의 분석을 거치지 않고 개인정보를 찾아 마스크 하였고, 이미지 보정을 통해 OCR 엔진 글자 인식률을 향상했다. 또한, 이미지에서의 개인정보 분석을 통해 개인정보 정탐률을 높였다.

실험 결과, 기존의 서식인식 방안보다 분석속도가 0.5초 향상되었으며 이미지 서식인식 인식률은 평균 99%를

기록하였다. OCR 엔진은 상용 OCR보다 2.3초, Tesseract OCR 7.3초가 향상된 것으로 나타났으며 OCR 엔진 글자 인식률은 한글 처방전 문자인식 시스템 엔진보다 9.09%, KCR-AlexNet OCR 엔진보다 6.3%, Tesseract OCR 엔진보다 23.3% 향상되었다.

향후, 제안하는 기술과 인공지능 기술을 활용하여 문서내의 데이터를 식별하여 정합성 검증 사물 인식 처리에 대한 연구가 요구된다.

References

[1] I. G. Cheon, T. Y. Young, "Basic image processing", KiHanJae, 1999.

[2] D. H. Jang, "Implementation of Digital Image Processing", PC ADVANCE, 1999.

[3] https://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software (accessed Oct. 31. 2019)

[4] https://docs.opencv.org/4.1.2 (accessed Oct. 31. 2019)

[5] https://en.wikipedia.org/wiki/Regular_expression (accessed Oct. 31. 2019)

[6] Ray Smith, "An Overview of the Tesseract OCR Engine", Google Inc., 2007.

[7] J. H. Cho, C. W. Ahn, "Auto Detection System of Personal Information based on Images and Document Analysis", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol 15 No 5, pp.183-192, 2015.
DOI:https://doi.org/10.7236/JIIBC.2015.15.5.183

[8] J. W. Kim, S. T. Kim, J. Y. Yoon, Y. I. Joo, "A Personal Prescription Management System Employing Optical Character Recognition Technique", *Journal of the Korea Institute of Information and Communication Engineering*, Vol 19, No. 10, pp.2423-2428, 2015.
DOI:https://doi.org/10.6109/jkiice.2015.19.10.2423

[9] S. C. Park, "Design and Implementation of Personal Information Identification and Masking System Based on Image Recognition", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol 17 No 5, pp.1-8, 2017.
DOI:https://doi.org/10.7236/JIIBC.2017.17.5.1

[10] Y. G. Kim, "*Improvement of Korean Characters Recognition Performance Using CNN and Feature Extraction*", Ph.D dissertation, Pusan National University, 2017.

[11] G. W. Joe, "*A Personal Information Detection Method of Image File*", Master's thesis, Jeonbuk National University, 2018.

[12] S. H. Lee, J. H. Joen, H. S. Hong, D. H. Kang, M. H. Park, "Korean Prescription Character Recognition System Using OCR Technology", *Korean Institute of Information Scientists and Engineers 2017 Conference*, Korea, pp.362-364, 2017.

백 종 경(Jong-Kyung Baek) [정회원]

- 2011년 2월 : 숭실대학교 정보과학대학원 정보보안학과 (공학석사)
- 2020년 2월 : 숭실대학교 대학원 컴퓨터학과 (공학박사)

〈관심분야〉
정보보안, 정보통신, 암호학, AI, Cloud

지 윤 석(Yoon-Seok Jee) [정회원]

- 1998년 8월 : 한양대학교 산업대학원 전자계산학과 (공학석사)
- 2020년 2월 : 숭실대학교 대학원 IT정책경영학과 (박사수료)

〈관심분야〉
정보통신 기반시설 및 제어시스템 보안, 정보보안관리 실태평가 제도, 사이버 보안정책

박 재 표(Jae-Pyo Park) [종신회원]

- 1998년 2월 : 숭실대학교 대학원 컴퓨터학과 (공학석사)
- 2002년 8월 : 숭실대학교 대학원 컴퓨터학과 (공학박사)
- 2010년 3월 ~ 현재 : 숭실대학교 정보과학대학원 교수

〈관심분야〉
정보보안, 보안평가 및 인증, 디지털포렌식, FinTech