한국컴퓨터정보학회 동계학술대회 논문집 제25권 제1호 (2017. 1)

이미지파일에 포함된 개인정보추출에 관한 연구

이민석[O], 김숙현[*] 윤지애[*], 원유재[*]
[O*]충남대학교 컴퓨터공학과
e-mail: {minsuk805, kimsh2983}@nate.com,
dbswlo548@naver.com, yjwon@cnu.ac.kr[*]

A Study on Detecting Personal Information from Image Files

Minsuk Lee[O], Sukhyeon Kim[*], Jiae Yoon[*], Yoojae Won[*]
[O*]Dept. of Computer Science & Engineering, Chungnam National University

● 요 약 ●

최근 정보통신기술의 비약적 발전에 따라 문서 제작 과정 또한 디지털 방식의 형태가 주를 이루게 되었다. 하지만 이와 더불어 문서를 통한 개인 정보 유출의 문제 또한 대두되게 되었다. 본 논문에서는 이미지 형식의 문서의 유출 방지를 위해 광학문자인식(OCR)을 활용한 문자인식 기능과 개인정보 검출 기능을 통합적으로 수행 한하여 기존 OCR엔진과의 차별점을 두었다. 또한 원하는 경로의 파일 탐색을 가능하도록 하고, 선택한 경로에 저장되어 있는 이미지파일 내의 검출 문자들을 정규표현식을 사용해 특정한 개인정보 패턴과 매칭하여 문서 내 포함된 개인정보를 반환하여 출력한다. 이러한 개인정보 검출 결과 중요 개인정보가 포함된 파일을 사용자에게 별도로 통보하도록 한다. 따라서 본 논문에서는 기존의 개인정보 검출 과정의 번거로움을 극복하여 사용자의 편의 향상과 더불어 문서를 통한 개인정보의 유출을 사전에 방지 할 수 있도록 하였다.

키워드: 개인정보(Personal Information), 광학문자인식(Optical Character Recognition), 문자인식(Character Recognition), 정규표현식(Regular Expression)

I. Introduction

국내 인터넷과 스마트폰 보급률은 꾸준히 증가하였고, 이에 따라 인터넷 또는 웹서비스를 이용하는 사용자도 크게 증가하였다. 서비스를 사용하는 사용자가 증가함에 따라 서비스를 사용하고자 하는 사용자가 옳은 사용자인지 인증하는 과정은 매우 중요해졌다. 따라서 각종 인증 시스템이 등장하였고, 국내에서는 금융거래 시 공인인증서를 의무적으로 사용하도록 규정을 세워 사용자 인증에 신뢰성을 높였다. 그러나 2015년 3월 공인인증서 의무사용 규정이 폐지되었고, 이에 따라 다양한 보안 시스템에 대한

요구가 증가하고 있다. 이러한 흐름에 맞춰 최근에 핀테크 시장에서는 새로운 인증시스템이 등장하고 있다.

시장에 등장하고 있는 인증시스템 중 가장 주목받는 분야는 생체정보를 이용한 생체인증 분야이다. 기존에 사용하던 아이디와 패스워드를 이용한 인증의 경우, 서비스 제공자는 사용자가 사용할 아이디와 패스워드에 대해서 철저한 보안이 필요하고, 사용자 또한 보안 강화를 위해 복잡하고 어려운 패스워드를 이용하면서 주기적으로 변경을 요구하는 등 현실적인 어려움이 존재했다. 이에 비해 생체인식의 경우 인증에 사용자의 생체정보를 이용하는데, 생체정보는 사용자의 고유한 정보이기 때문에 앞서 아이디와 패스워드를 사용하였을 시 발생한 관리에 대한 현실적인 문제들이 발생하지 않으면서도 간단하게 인증할 수 있다는 장점이 있다.

본 논문에서는 최근의 추세를 반영하고, 별도의 생체인식 장치를 필요로 하지 않고 사용자가 디바이스의 입력장치를 사용할 때 발생하는 다양한 정보를 분석하여 인증에 사용하고자하는 행위 기반 인증을 제시하고자 한다. 이에 먼저 최근에 등장한 인증방식들에 대한 간단한 내용을 2장에서 소개할 예정이다. 3장에서는 본 논문에서 제안하고자 하는 인증방식을 기술한다. 4장은 결론으로 우리가 제시한 인증방식에 대한 실험을 통해 실질적으로 사용자 인증에 사용될 수 있을지 분석한다.

II. Related Works

1. TesseractOCR

광학문자인식(Optical Character Recognition, OCR)은 출력된 문서 및 이미지 스캐닝을 통해 한글, 영문, 숫자 폰트에 대해 편집 가능한 텍스트로 변환하고 저장할 수 있게 해주는 기술이다[1]. 광학문자인식 기술은 매우 다양한 분야에 응용되고 있으며, 뱅킹, 의료산업, 자동 번호판 인식, 필기 인식 등 광범위하게 적용되고 있다.

TesseractOCR은 1984년부터 약 10년간 HP에서 개발한 오픈소스 OCR엔진으로 개발 이후 지속적인 성능 개선을 통해 2005년 오픈소스로 발표되었다. 현재는 구글이 Tesseract의 일부를 지원하고 있다[2]. TesseractOCR은 문자인식단계와 학습단계로 크게 2개의 단계로 구분된다. 첫 번째, 문자인식단계에서 TesseractOCR에 명령어를 통해 jpg파일을 넣으면 인식 결과로 텍스트 파일이 생성된다. 두 번째, 인식률이 낮은 jpg파일에 대해 인식률을 높이기 위해 문자학습단계를 진행한다[3]. 문자 학습을 하기 전, jpg파일 이름 형식을 학습하고자하는 언어명, 폰트종류, 학습을 진행할 때 분류할 수 있는 숫자순으로 변경한다[4]. 다음 Fig. 2는 TesseractOCR의 동작 흐름도를 나타낸다.

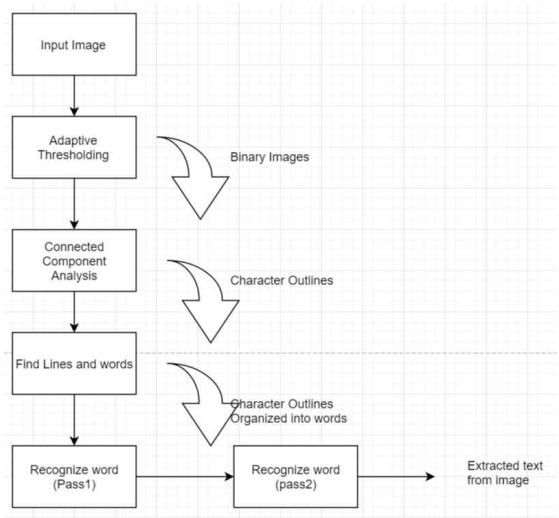

Fig. 1. TesseractOCR Operation Flowchart

2. 정규표현식

정규표현식의 사전적인 의미는 특정한 규칙을 가진 문자열의 집합을 표현하는데 사용하는 형식 언어이다. 주로 Programming Language나 Text Editor 등 에서 문자열의 검색과 치환을 위한 용도로 쓰이고 있다. 입력한 문자열에서 특정한 조건을 표현할 경우 일반적인 조건문으로는 다소 복잡할 수도 있지만, 정규표현식을 이용하면 간단하게 표현 할 수 있다.

정규표현식은 사용자로부터 값을 입력 받는 부분에서 유효성 체크 외에도 Email Check. File확장자 Check, 주민등록번호 Check, 문자열 공백제거, 문자열 첫 글자 대문자로 치환 등의 다양한 형태의 유효성검사에 사용된다.

주민등록번호	[01][0-9]{5}[[:space:],~-]+[1-4][0-9]{6}|[2-9][0-9]{5}[[:space:],~-]+[1-2][0-9]{6}
이메일	(₩w+₩.)*₩w+@(₩w+₩.)+[A-Za-z]+
여권번호	[a-zA-Z]{2}[-~.[:space:]][0-9]{7}
운전면허번호	[0-9]{2}[-~.[:space:]][0-9]{6}[-~.[:space:]][0-9]{2}
핸드폰번호	01[016789][-~.[:space:]][0-9]{3,4}[-~.[:space:]][0-9]{4}
외국인등록번호	[01][0-9]{5}[[:space:]~-]+[1-8][0-9]{6}|[2-9][0-9]{5}[[:space:]~-]+[1256][0-9]{6}
신용카드번호	[34569][0-9]{3}[-~.[:space:]][0-9]{4}[-~.[:space:]][0-9]{4}[-~.[:space:]][0-9]{4}
건강보험번호	[1257][-~.[:space:]][0-9]{10}
IP주소	((([0-9])|([1-9]₩d{1})|(1₩d{2})|(2[0-4]₩d)|(25[0-5]))₩.){3}(([0-9])|([1-9]₩d{1})|(1₩d{2})|(2[0-4]₩d)|(25[0-5]))
계좌번호	[0-9]{2}[-~.[:space:]][0-9]{2}[-~.[:space:]][0-9]{6}|[0-9]{3}[-~.[:space:]]([0-9]{5,6}[-~.[:space:]][0-9]{3}|[0-9]{6}[-~.[:space:]][0-9]{5}|[0-9]{2,3}[-~.[:space:]][0-9]{6}|[09]{2}[-~.[:space:]][0-9]{7}|[0-9]{2}[-~.[:space:]][0-9]{4,6}[-~.[:space:]][0-9]|[09]{5}[-~.[:space:]][0-9]{3}[-~.[:space:]][0-9]{2}|[0-9]{2}[-~.[:space:]][0-9]{5}[-~.[:space:]][0-9]{3}|[0-9]{4}[-~.[:space:]][0-9]{4}[-~.[:space:]][0-9]

Fig. 2. Regular Expression of Personal Information

본 논문에서 제안하는 방법을 사용하기 위해 개인정보의 일정하게 반복되는 패턴을 분석하여 이를 Fig. 2와 같은 형태로 프로그램 소스코드에 접목하여 추출된 문자들과 매칭하는 방식으로 활용한다.

III. The Proposed Scheme

1. 개인정보 추출 프로그램

본 논문에서는 이미지파일에 포함된 개인정보 추출 방법을 제안한다. 이를 위해 개발한 프로그램의 주요 기능은 이미지파일 내에서의 문자인식 및 문자학습기능, 검출된 문자로부터의 개인정보 검출 기능으로 대분 할 수 있다. 다음 Fig. 3은 개인정보 검출 프로그램의 구조를 나타낸다.

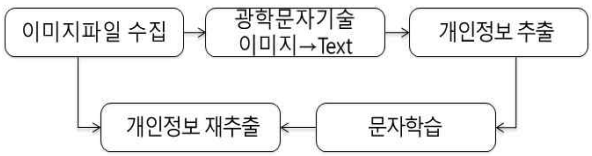

Fig. 3. Architecture of Personal Information Extraction Program

개인 PC내에서 이미지 파일을 수집해 광학문자인식(OCR)기능을 사용하여 수집한 이미지 파일 내 문자들을 텍스트 형식으로 전환하여 저장한다[5]. 이후 텍스트 형태로 저장된 내용과 정규표현식을 활용한 개인정보패턴을 이용하여 개인정보를 추출하여 사용자에게 반환하게 된다. 이 과정에서 인식률이 떨어지는 문자들은 별도의 문자학습 과정을 거친 뒤 개인정보를 재추출하여 프로그램의 정확성과 효율성을 높인다.

2. 개인정보 검출 방법

개인정보 검출 프로그램에는 두 가지의 기능으로 나눌 수 있다.

첫 번째 이미지파일 내의 문자 인식과 인식된 문자의 학습 기능에서, 사용자는 프로그램의 탐색기 기능을 활용하여 직접 개인 PC 내에서 개인 정보 검출을 여부를 확인하고 싶은 폴더들을 선택해 해당 경로 내의 파일을 확인한다. 정렬된 파일은 사용가자가 특정 확장자를 지정하여 이미지파일들로 정렬이 가능하다. 이 때 정렬된 이미지파일에 포함된 문자를 광학 문자 인식(OCR)기능을 통해 이미지 내 문자를 텍스트로 변환하고 각 파일들의 추출 결과를 텍스트 파일 형태로 저장하게 된다. 이러한 방식으로 저장된 텍스트 파일 형태의 문자들은 지속적 학습을 통해 점차 인식률을 향상 시킬 수 있고 이에 따라 보다 정확하게 파일 내의 문자를 추출 할 수 있다.

두 번째 개인정보 검출 기능에서는 일정하게 반복되는 개인정보들의 패턴을 파악해 정규표현식을 작성하고 이를 활용한 소스코드를 프로그램에 접목 시켜 특정 경로에서 인식되고 학습되어진 문자들과 각각의 개인 정보 패턴 코드를 매칭 시킨다. 프로그램은 이러한 매칭 결과에 따라 포함된 개인정보를 반환하여 사용자에게 파일 내의 포함된 개인정보를 알린다.

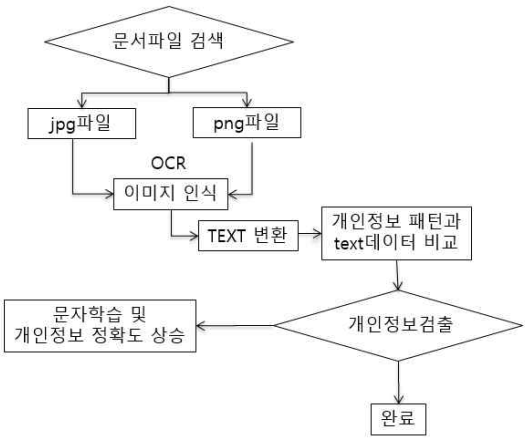

Fig. 4. Flow of Detecting Personal Information in Image Files

3. 실험 결과

최종 구현 결과는 특정 폴더 내 이미지 파일에서의 개인 정보 검출과 검출의 정확성을 높이기 위한 문자 학습 기능으로 구분하여 설명한다.

특정 폴더 내 이미지파일의 개인 정보 검출 화면은 Fig. 5와 같다. 사용자가 본 소프트웨어를 실행하면 Fig. 5의 좌측 그림과 같이 탐색기 화면이 나타난다. 이 탐색기 화면에서 개인 정보 검출 기능을 사용하고자 하는 특정 폴더를 선택하면 Fig. 5의 우측 상단과 같이 해당 폴더에 존재하는 이미지 파일이 리스트 형태로 출력된다. 이후, 상단의 메뉴에서 파일>개인정보검출을 선택하면 Fig. 5의 우측 하단과 같이 해당 폴더 내 이미지 파일에 포함되어 있는 개인정보 내역이 출력된다.

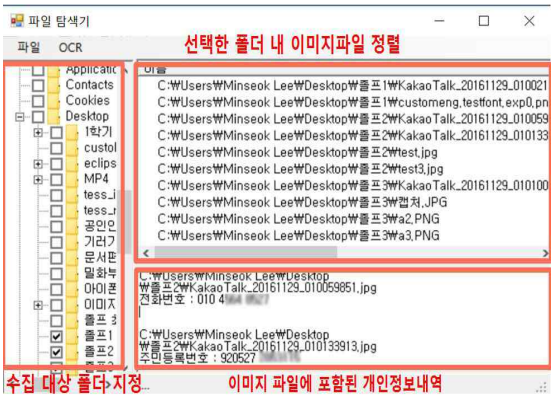

Fig. 5. Personal Information Detection Result Screen in Image Files

이러한 개인 정보 검출 과정 중 이미지파일에서 텍스트로의 전환과정 중 인식의 정확성이 다소 떨어지는 문제를 보완하기 위해 아래의 Fig. 6과 같은 문자인식과 학습 기능을 추가로 구현하였다. 초기의 문자인식 결과 Fig. 6의 좌측 상단의 그림과 같이 숫자 외의 한글 인식에서는 그 정확성이 다소 낮은 것을 확인할 수 있다. 이러한 문제의 해결의 위해 프로그램 상단 메뉴에서 OCR기능을 사용하여 Fig. 6의 아래 그림과 같은 문자학습 기능을 거치게 되면, 우측 상단의 그림과 같이 문자 인식의 정확성이 향상된 것을 확인할 수 있다.

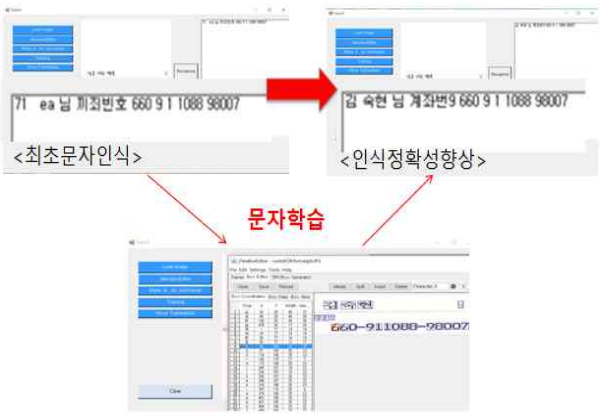

Fig. 6. Improved Accuracy of Personal Information Through Character Learning

IV. Conclusions

본 논문에서는 이미지파일에서 개인정보를 추출하는 방법을 제안하였다. TesseractOCR과 인식된 문자를 수정하는 툴을 별도로 사용하여 이미지에서의 문자 인식과 문자학습을 통해 인식률을 높이는 과정을 하나의 소프트웨어에로 통합하여 보다 효율적인 개인정보 추출을 할 수 있게 하였다. 또한 원하는 경로에 있는 이미지 파일을 탐색하여 이미지파일에 포함된 텍스트를 개인정보 패턴과 매칭 시켜 정보 포함 여부를 참과 거짓으로 반환하여 출력함으로써, 사용자의 개인정보를 포함한 문서의 부주의적 유출을 방지한다.

따라서 본 논문에서는 사용자 및 개인정보 관리자의 편의를 향상시

키고 개인정보를 포함한 문서의 부주의적 유출을 방지하여 정보유출에 따른 피해를 사전에 방지할 수 있다.

Acknowledgment

This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in

SW)(R7115-16-1007) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

[1] Victor Oh lsson, “Optical Character and Symbol Recognition using Tesseract”, pp.1-77, July 2016

[2] J Kim, S Kim, J Yoon, YI 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, Oct 2015.

[3] Ray Smith “An Overview of the Tesseract OCR Engine”, Proc. of ICDAR, Vol. 2, pp.629-633, 2007.

[4] EB Goe, YJ Ha, SL Choi, GH Lee, YH Park, “An Implementation of an Android Mobile System for Extracting and Retrieving Texts from Images” Journal of The Korea Digital Contents Society, Vol12, No.1, pp57-67, March 2011.

[5] Md. Abul Hasnat, Muttakinur Rahman Chowdhury and Mumit Khan, “Integrating Bangla script recognition support in Tesseract OCR”, Proc. of the Conference on Language and Technology 2009 (CLT09), Lahore, Pakistan, pp.1-5, 2009.