The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 15, No. 5, pp.183-192, Oct. 31, 2015. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2015.15.5.183
JIIBC 2015-5-23

이미지와 문서 분석을 통한 개인 정보 자동 검색 시스템

Auto Detection System of Personal Information based on Images and Document Analysis

조정현*, 안철웅**

Jeong-Hyun Cho*, Cheol-Woong Ahn**

요 약 본 논문에서는 통신 판매사에서 사용하는 문서와 이미지 파일에서 개인 정보의 유출을 방지할 수 있는 개인 정보 자동 검색(PIAD, Personal Information Auto Detection) 시스템을 제안한다. 제안하는 시스템은 개인 정보를 포함하는 신분증과 계약서 이미지를 자동으로 검색하고 그 결과를 사용자에게 전달하고, 문서상의 개인 정보 또한 검출할 수 있다. 본 시스템은 빠르고 정확한 검색을 위하여 선별 과정과 분석 과정으로 나뉘고, 분석 과정은 SURF, 침식과 팽창, FindContours 알고리즘들을 사용한다. 제안하는 PIAD 시스템은 272장의 입력 이미지들 중 267장을 선별 및 검출함으로써 98% 이상의 정확도를 보였다.

Abstract This paper proposes Personal Information Auto Detection(PIAD) System to prevent leakage of Personal informations in document and image files that can be used by mobile service provider. The proposed system is to automatically detect the images and documents that contain personal informations and shows the result to the user. The PIAD is divided into the selection step for fast and accurate retrieval images and analysis which is composed of SURF, erosion and dilation, FindContours algorithm. The result of proposed PIAD system showed more than 98% accuracy by selection and analysis steps, 267 images detection of 272 images.

Key Words : Personal Information Auto Detection(PIAD), SURF, Mobile Service Provider

Ⅰ. 서 론

스마트폰의 성장 속도는 2007년 1억2천만 개의 판매로 시작해 2013년도는 9억 6천 7백만 개(967million)이상의 가입자로 급성장하게 되었다.[1] 특히, 대한민국은 IT 기술이 발달되어 인터넷 이용률이 높고 2013년 7월 현재 스마트폰 이용자의 가입수는 3,595만 명을 넘어섰다.[2] 한국은 2009년 11월 아이폰이 출시된 후 2013년까지 4천만 명의 스마트 가입자를 기록했고, LTE 가입자는 2,399만 명을 기록하는 등 한국이 스마트 강국임을 증명하고 있다.

최근 늘어나는 스마트폰 가입자와 함께 심각한 문제점은 개인 정보에 대한 유출 사고가 늘어나고 있다는 것이다. 개인 정보 유출은 개인에게는 명의 도용, 보이스 피싱과 같은 금전적 손해와 정신적 피해를 주고, 기업은 기업 고객의 정보 유출에 따른 기업의 이미지 손상, 집단 손해 배상으로 기업에 타격을 준다. 현재 개인 정보 유출 실태는 통신 3사와 금융기관 11개, 여행사, 불법 도박 사이트, 인터넷 쇼핑몰 등에서 개인정보 1천230만 건이다.[3]

개인 정보 유출 피해 사례 건수는 2013년 177,736건으

*정회원, 영남이공대학교 컴퓨터정보과
**정회원, 계명문화대학교 스마트웹콘텐츠과(교신저자)
접수일자 : 2015년 8월 1일, 수정완료 2015년 9월 7일
게재확정일자 : 2015년 10월 9일

Received: 1 August, 2015 / Revised: 7 September, 2015 / Accepted: 9 October, 2015
**Corresponding Author: homepig@kmcu.ac.kr
Dept. of Smart web contents, Keimyung College University, Korea

로 꾸준히 증가되다가 2014년 11월에는 34,000건이 줄어든 143,670건이었다.[4] 개인 정보 유출 피해는 그림 1과 같이 2013년도의 개인 정보 침해 신고 상담 건수의 75%인 129,000건이 주민등록번호 등 타인 정보의 훼손·침해·도용임을 보여준다. 즉, 개인 정보의 잘못된 관리와 사용으로 인해 개인 및 기업에게 상당한 손해를 발생시키게 하는 원인이 된다.

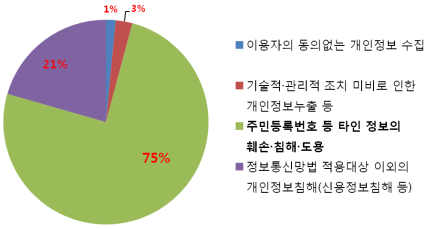


그림 1. 2013년도 개인 정보 침해 신고 상담 건수 (출처: 인터넷 침해 대응)
Fig. 1. Counselling Count of Privacy Infringement Report(KrCERT/CC)

개인 정보 유출은 통신사, 금융기관, 인터넷 쇼핑몰 등 다양한 기업에서 관리 소홀로 1천 만 건 이상이 발생하였고, 이 중 420여 만 건에 달하는 통신사 개인정보의 경우 본사 해킹이 아닌 판매점에서 유출된 것으로 보인다. 이들이 컴퓨터 파일 형태로 보관 중이던 개인정보는 LG유플러스 250만 건, KT 7만6천여건, SK브로드밴드 159만여건 등 총 420만여 건이다. 유출된 개인정보는 이름, 주민등록번호와 전화번호, 주소, 계좌번호 등이 포함된 것으로 알려졌다.[3] 통신사가 관리하는 개인 정보보다 판매사가 보유하는 개인 정보는 그 관리가 소홀한 문제점이 있다는 것을 확인할 수 있다. 통신 판매사가 관리하는 개인 정보는 신분증 복사 이미지, 계약서 이미지, 기기 변경 신청 이미지와 문서 파일 형태로 판매사 개인 PC에 저장 관리되어 그 허점이 더욱 큰 것이 현실이다.

본 논문에서는 통신 판매사 등에서 고객 유치 및 관리를 위해서 발생되는 고객 신분증의 복사 이미지, 고객의 정보가 포함된 계약서 이미지, 기기 변경 신청서 이미지, 그리고 고객의 개인 정보가 기록된 문서 파일들에서 자동으로 신분증 이미지, 계약서 이미지, 신청서 이미지, 문서들을 자동으로 검색, 추출하고 그 결과를 사용자와 관리 서버에 전달하여 삭제할 수 있도록 하는 시스템을 기술한다.

본 논문의 2장에서는 개인 정보 보호를 위한 기존 방법에 대하여 기술하고, 3장에서는 본 제안 시스템인 개인 정보 자동화 시스템이 개인 정보를 포함하는 이미지를 어떻게 분석하는지를 설명한다. 4장에서는 제안 시스템에 대한 실험 결과를 제시하고 5장에서 결론을 맺는다.

II. 개인정보 보호 방법

1. 개인 정보 보호를 위한 PC 통합 솔루션

개인정보보호법에 명시된 개인정보란 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 말한다.[5] 법에 명시된 개인 정보를 추출, 수집 및 활용을 할 경우 법적인 제재를 받게 된다.

정보 보호를 위한 방법은 통합 PC 및 인터넷 웹에서 솔루션 도입으로 시작된다.[6, 7] 통합 PC 솔루션은 PC 접근 제어, PC 방화벽 기능, 특정 URL 차단 등과 같은 PC 보안 기능, PC 내 자원 관리, 불법 S/W관리, 보안 패치 기능을 담은 자산 관리 기능, USB, SMTP, 화면 보호기, 프린트 보안 등을 차단할 수 있는 정보 유출 방지 기능 등을 포함한다. 대표적인 솔루션은 에스원 PS, 라온시큐어, V3 MSS 등이 있으나 가격 정책이 높기 때문에 일반인이 사용할 수 없는 기업용 솔루션이다.

그리고 개인 문서 정보를 보호하고 유출을 방지하기 위하여 DRM(Digital Right Management), DLP(Data Loss Prevention), ECM(Enterprise Contents Management) 등과 같은 보안 솔루션을 통합 구축하고 있다.[8, 9] 기업 내에서 관리하는 문서는 DRM을 통해서 반드시 인코딩과 디코딩을 수행하도록 함으로써 정보 유출을 차단하는 것이다. 대표적인 솔루션은 fasoo, 클라우드 DRM T Bizpoint 등이 있다. 그러나 이 방법은 솔루션에 들어가는 비용이 높다는 단점과 일반인이 사용하기에는 여러 가지 제약을 가지고 있다.

PC에 저장된 개인 정보를 검색하여 삭제하는 등의 개인 정보 검색 솔루션들이 있다. 대표적인 솔루션은 Privacy i, PC 필터, i-Safer 등의 제품이 있으며 검색을 통해 개인 정보를 검출 및 관리하고, 개인 정보의 현황 파악을 할 수 있는 리포터를 제공한다.[10, 11] 그리고 서버

와 연결되어 있는 클라이언트 PC의 개인 정보 검색 및 관리를 수행할 수 있다. 개인 정보 검색은 전화 번호, 주민등록번호, 면허 번호, 신용카드 등이 되고 파일을 읽어서 검색하고자 하는 정보들의 패턴이 검색되면 해당 파일은 임시 또는 영구 삭제를 한다. 개인 정보 검색 솔루션은 MS office, PDF와 같은 정형화된 문서 파일만 가능하다. 이미지로 저장되어 있는 개인 정보는 검색할 수 없는 단점을 가진다.

2. 신분증 인식

사용자들의 개인 정보는 PC 상에 존재하는 데이터와 계약, 업무 처리를 위한 신분증 복사 또는 스캔을 통한 이미지 저장 등으로 복제된다. 이미지에서 개인 정보 인식 대상은 주민등록증, 자동차 면허증 등으로서 사용자 정보인 주민등록번호, 주소, 사진 등을 추출하여 데이터베이스에 저장하거나 인명 검색 등에 활용한다.[12, 13] 이러한 시스템은 특정 스캐너, 바코더, 카메라를 통해 신분증을 스캔한 후 그 결과를 이용하는 방식으로 금융권, 대기업, 보안 업체에서 많이 사용한다. 신분증의 특성을 파악하여 얼굴 인식, 문자 인식, 패턴 인식과 같은 다양한 인식 알고리즘을 이용하여 정보를 추출한다. 대표적인 제품은 Plustek, 아이콤정보시스템, InziSoft 등이 있다. 그러나 통신 판매사는 이런 스캐너와 처리를 위한 서버가 존재하지 않고 일반적인 스캐너에 신분증, 계약서를 스캔한 후 PC에 저장하고 통신사에는 해당 이미지와 문서를 전송한다. 이 과정에서 발생하는 스캔된 개인 정보인 계약서, 신분증 복사본 등은 통신판매사의 PC에 고스란히 그대로 저장된 채 남아있게 되는 문제점을 가지게 된다.

제안한 PIAD 시스템은 PC에 저장된 개인 정보를 검색하고 그 결과에 대하여 사용자와 서버(선택적)에 정보를 전달한 후 삭제를 하게 된다. 사용자 편의성을 위한 직관적 UI를 제공하고 MS Office 문서, PDF, 그리고 이미지로 저장되어 있는 개인 정보를 검색할 수 있다.

Ⅲ. 이미지 내의 개인 정보 검색

통신 판매사에서 관리하는 개인 정보 파일은 4가지 종류가 있는데 신분증의 복사 이미지, 계약을 위한 고객의 정보가 포함된 계약서 이미지, 무선 휴대 전화기기의 변경을 위한 신청서 이미지, 그리고 고객의 개인 정보인 전화 번호, 주민등록증을 포함한 문서 파일들이다. 그림 2는 판매사에서 관리하는 대표적인 이미지들이며 신분증의 위치가 고정되지 않고 회전되어 있거나 전체 이미지 내에 잡음이 존재하기도 한다. 계약 이미지는 많은 글자가 포함되어 있고, 이미지 내에는 고객과 상담하면서 직접 쓴 문자가 존재한다. 이러한 이미지들은 이미 개발된 기존 제품 또는 방법으로는 개인 정보가 포함되어 있는지를 판별하지 못하고 있다.

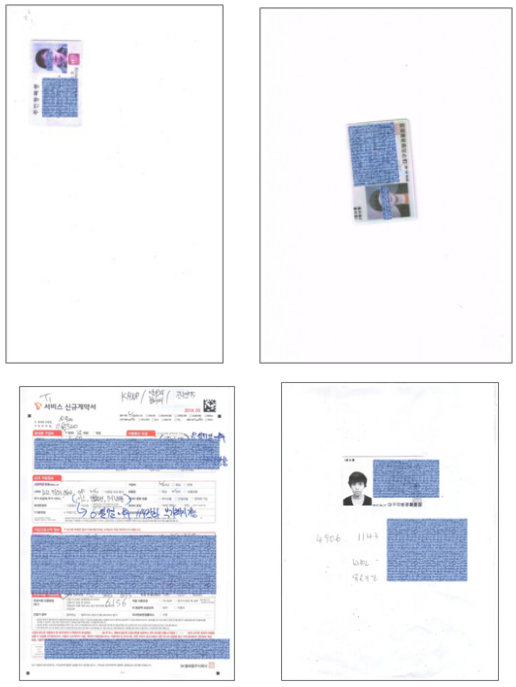


그림 2. 통신 판매사에서 사용하는 개인 정보가 포함된 이미지들
Fig. 2. Images that contain personal information used by mobile service provider

1. PIAD 시스템 개요

PIAD 시스템은 보안 인증을 통하여 승인된 PC만 사용 가능하다. 통신사에서 판매점에 할당한 P-CODE, PC의 MAC 주소, 상호명, 연락처 정보를 검증하여 인증이 되어야 한다. 인증이 된 PC는 전체 디렉토리를 검색하여 이미지와 문서를 분리한다. 판매사는 어느 폴더에 개인 정보 이미지를 저장해 두는지 기억할 수 없기 때문에 사용자가 지정한 모든 폴더와 그 하위 폴더를 검색하게 된다.

다음 단계는 검색이 완료된 후 이미지 파일들을 대상

으로 선별을 한다. 선별 단계는 분석 속도를 향상시키고 시스템의 리소스를 줄이기 위한 과정이고 선별에서 추출된 이미지들은 실제 개인 정보가 포함되었는지를 확인하기 위한 분석 단계를 거치게 된다. 분석 단계는 개인 정보가 포함된 계약서 이미지와 신분증이 포함된 이미지를 구분하여 분류한다. 분석이 완료된 후 최종 결과를 사용자에게 보여주고 필요 시 서버에 저장을 한다. 그림 3은 시스템의 각 단계를 보여준다.

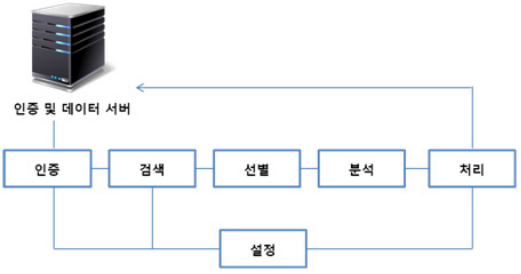


그림 3. 개인 정보 자동 검색 시스템 개요도
Fig. 3. Personal Information Auto Detection System (PIAD)

2. 선별(Selection)

통신 판매사에서 관리하는 개인 정보를 포함하는 이미지들은 신분증의 위치, 회전, 크기가 모두 불명확하고 잡음과 손글씨(hand writing)가 이미지 내에 존재하기 때문에 기존 방식으로 처리할 경우 많은 시스템 리소스와 시간이 필요하다. 본 논문에서는 판매사에서 관리하는 이미지들의 특성을 정리하여 분석에 필요한 시스템 리소스와 시간을 줄이기 위한 방안을 제시한다.

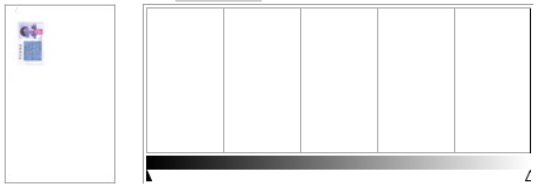
[특징 1] 신분증이 포함된 이미지는 전체적으로 흰색 여백이 많이 존재한다. 계약서 등은 글자를 제외한 흰색 여백이 많다.

입력 이미지는 특징 1을 판단하기 위하여 색상 이미지를 그레이 이미지로 변환한 후 히스토그램을 추출한다. 히스토그램 처리는 이미지 처리에서는 상당히 짧은 시간과 적은 리소스가 필요하기 때문에 PC 상에 존재하는 많은 이미지들을 선별할 수 있는 좋은 방법이다.[14] 신분증과 계약서의 배경 색상이 흰색이거나 그에 가까운 값이기 때문에 모든 히스토그램의 빈을 사용하지 않고 빈의 값이 240 이상의 경우만 고려한다.

그림 4(a)는 일반적인 이미지의 히스토그램 추출 결과이고, 그림 4(b)는 신분증 이미지의 히스토그램이다. 입력된 이미지가 분석 단계가 필요한지를 판단하기 위해 식 1과 같이 히스토그램의 빈의 개수가 전체 영역 내에서 차지하는 비율이 임계치 th_{hist_low}, th_{hist_high} 사이를 만족해야 한다.

$$th_{hist\_low} < percent\left(\sum_{n=\alpha}^{255} Hist_n\right) < th_{hist\_high} \quad (0 \le \alpha \le 255) \qquad (1)$$

(a) 일반 이미지



(b) 개인 정보가 포함된 이미지

그림 4. 일반 이미지와 통신 판매사에서 사용하는 이미지의 히스토그램 차이
Fig. 4. Histogram difference between a general image and mobile service provider's image

3. 분석(Analysis)

가. 개인정보를 포함하는 신청서, 계약서 이미지 분석

선별된 이미지들은 개인 정보 포함 여부를 판단하기 위해 이미지마다 분석을 수행한다. 판매사에서 관리하는 계약서 등의 스캔한 이미지는 상단 또는 하단 영역에 특징 2가 존재한다.

[특징 2] 계약 정보를 포함된 계약서 이미지, 휴대 전화기기 변경 신청서 등은 상단 또는 하단 - 스캔된 이미지가 상하 바뀌었을 경우 - 영역에 템플릿 매칭(template matching)을 위한 단어가 포함되어 있고 글자들로 구성되어 있기 때문에 굴곡 특성을 가지는 변곡점들이 존재한다.

스캔된 이미지는 상단 또는 하단의 일정 영역 내에 신청서 또는 계약서 등의 단어를 포함한다. 전체 영역에서 단어 검색은 많은 시간이 소요되기 때문에 특징 2를 이용하여 전체 영역에서 상단과 하단 10%의 영역을 ROI(Region of Interesting)로 설정한 후 SURF (Speeded-Up Robust Features) 특징점을 추출하여 그 개수가 th_{surf} 이상이고 템플릿 이미지와 SURF 특징점을 비교하여 th_{surf_match} 이상일 때 개인 정보를 포함하는 계약서 또는 신청서 이미지로 판단한다.[15, 16]

SURF 알고리즘은 SHIF 알고리즘의 문제점인 처리 속도를 빠르게 하기 위한 것으로써 크기와 회전 불변의 특징을 추출할 수 있다.[17] Harris가 제안한 헤이시안 행렬식(Hessian matrix)과 기존의 헤리스 코너 검출기(Hessian based blob detector)가 크기에 따른 불변을 해결할 수 없기 때문에 가우시안 2차 미분검출기(Laplacian of Gaussian, LoG)를 이용하였다. B. Bay는 처리 속도를 빠르게 하기 위해서 적분 영상(Integral image)과 근사화된 헤이시안 검출기의 행렬식(식 2)을 사용하였다.

$$H(x,y,\sigma)=\begin{bmatrix} LI_{xx}(x,y,\sigma)\ LI_{xy}(x,y,\sigma) \\ LI_{xy}(x,y,\sigma)\ LI_{yy}(x,y,\sigma)\end{bmatrix} \quad (2)$$

여기에서 $LI(x,y,\sigma)$는 x,y 위치의 입력 영상과 σ 분산을 갖는 가우시안의 x 방향 2차 미분값 $\frac{\partial^2}{\partial x^2}g(\sigma)$과의 컨벌루션 값을 의미하고, LI_{xy}와 LI_{yy}는 xy 방향과 y 방향의 2차 미분값과의 컨벌루션 값을 의미한다.

적분 영상은 원점으로부터 각 픽셀의 위치까지의 사각형 영역의 모든 화소값을 더한 것을 의미한다. 따라서 적분영상은 어떤 크기의 사각형 영역이든 4번의 연산을 통해 특정 사각형 내의 모든 화소의 합을 구할 수 있다(식 3). 적분 영상에서 빠른 속도로 특징점 검출을 위해 그림 5와 같이 근사화된 헤이시안 사각필터를 이용한다. 그리고 크기에 대한 불변을 위해서 입력 영상의 크기를 변형하지 않고 그림 6과 같이 사각 필터의 크기를 조절하여 처리한다.

$$II(x,y)=\sum_{i=0}^{x}\sum_{j=0}^{y}I(x,y) \quad (3)$$

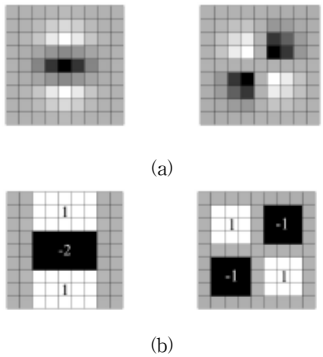


그림 5. 가우시안 2차 미분(a)과 근사화한 사각필터(b)
Fig. 5. Laplacian of Gaussian(a) and Hessian filter

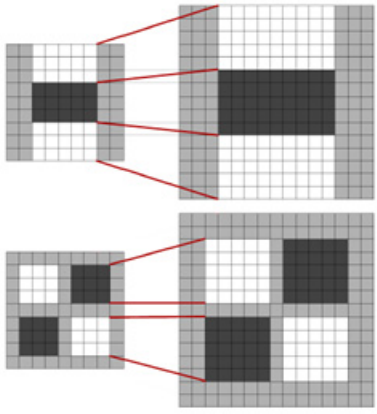


그림 6. 사각필터 스케일 변화를 이용하여 이미지의 크기 불편 특징점 추출
Fig. 6. Extraction of scaling invariant features using Hessian filter scale changes

방향 불변의 성분 추출을 위해서 관심점(interest point)의 6s 만큼의 주위 픽셀들의 방향 성분을 Haar Wavelet 변환을 통해 추출한 응답 결과값들과 60도 크기의 부채꼴 윈도우를 슬라이싱하면서 나온 합의 결과 중 가장 긴 벡터가 해당 포인터의 방향 성분이 된다. 추출된 방향대로 관심점 주변 20s의 윈도우를 회전한 후 Haar 웨이블릿 변환을 이용하여 x, y 방향에 대한 변환값이 최종적인 Descriptor가 된다.

본 논문에서는 전체 이미지의 SURF 특징점을 추출하지 않고 특징 2와 같이 상단 또는 하단의 10% 영역 내에서 특징점을 추출한다. 전체 이미지의 특징점 추출보다 처리 시간이 빠른 장점이 있다. 그림 7(a)는 입력 이미지의 상단 10% (1600x236) 영역만을 ROI로 설정한 것이고, 해당 이미지를 SURF 알고리즘으로 처리했을 경우 결과가 그림 7(b)에 있다.

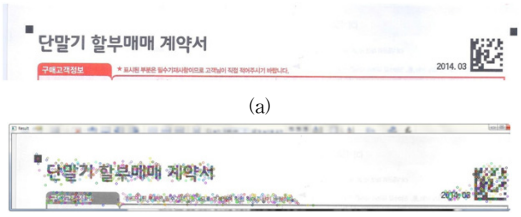


그림 7. 입력 이미지와 SURF 특징점 추출 결과
Fig. 7. Input image and SURF-based feature extraction

계약서 또는 신청서와 같은 이미지는 그림 7(b)와 같이 글씨의 특징을 가지고 있기 때문에 추출된 특징점 개수가 th_{surf}이상일 경우가 되지만 단순하게 특징점 개수만을 사용할 경우 글씨만 있는 이미지, 예를 들어 프리젠테이션 자료 등, 와 구분이 되지 않아서 동시 검출이 발생한다. 따라서 정확한 이미지 검출을 하기 위해서 템플릿 매칭을 수행한다. 이미지 상에서 존재할 수 있는 특정 단어에 대한 템플릿 이미지를 만들어 둔 후 특징점 매칭을 수행하여 매칭율이 th_{surf_match} 이상일 경우 찾고자 하는 개인 신분 정보가 포함된 이미지임을 확신할 수 있다. 그림 8은 템플릿 이미지와 그림 7(b)의 입력 이미지간의 SURF 특징점 매칭 결과이다. 매칭이 되었다고 판단되는 특징점은 템플릿 이미지의 특징점과 FLANN matcher 알고리즘을 통해 만들어진 최소 거리의 3배수 이내에 존재해야 한다. 그 개수가 전체 특징점 개수의 20% 이상일 경우 해당 이미지는 개인 정보가 포함된 이미지로 판단하게 된다.

나. 신분증을 포함하는 이미지 분석

[특징 3] 신분증을 포함하는 이미지는 전체 이미지 영역에서 특정 지점에 사각형의 형태로 포함되어 있고, 해당 영역의 크기는 전체 스캔한 이미지의 특정 크기 이내로 존재한다.

그림 8. 탬플릿 이미지와 입력 이미지 간의 SURF 특징점 매칭
Fig. 8. Matching of SURF features on between Template image and Input image

신분증을 포함하는 이미지는 특징 3과 같이 전체 스캔한 이미지에서 특정 영역의 일정 크기로 존재함으로, 신분증이 존재하는 위치를 찾고 그 지점에서 픽셀간의 연결을 그룹핑하여 사각형 영역을 인식한다. 인식한 사각형의 크기가 전체 영역의 특정 비율을 만족하게 되면 해당 이미지는 신분증을 포함하는 것으로 판단한다. 이미지 분석의 속도를 향상하고 정확도를 높이기 위해서 영역의 위치 검색 전에 이미지의 노이즈를 제거한다. 노이즈 제거는 침식(erosion)과 팽창(dilation)을 이용하여 불필요한 사각형 검색을 수행하지 않게 한다. 입력 영상의 노이즈 제거 후 사각형 영역을 찾기 위해 Suzuki가 제안한 FindContours 알고리즘을 사용하였다.[18]

FindContours 알고리즘을 사용하기 위해 먼저 입력 영상을 threshold 또는 canny edge를 사용하여, 에지 픽셀은 255로 그렇지 않은 경우는 0으로 처리한 이진 영상으로 변환한다. 이진 영상의 에지의 픽셀을 따라가면서 외부 영역에 해당하는 contours와 그 내부에 존재하는 holes로 검출한다. 그림 9는 샘플 이미지에서 contours (대쉬 라인)와 holes(도트 라인)를 반복적으로 표현하고 있다. 이렇게 검출된 영역들은 트리 형태로 표현이 가능하다. 그림 9에서 c0은 h00과 h01 hole을 자식 노드로 둘 수 있고, h00은 c000의 자식 노드를 가지게 된다. 이 때 이진 영상 변환 시 임계치 값을 높게 잡을 경우 contour와 hole의 수는 상대적으로 줄어들고, 반대로 임계치 값이 낮을 경우 많은 영역이 검출된다.

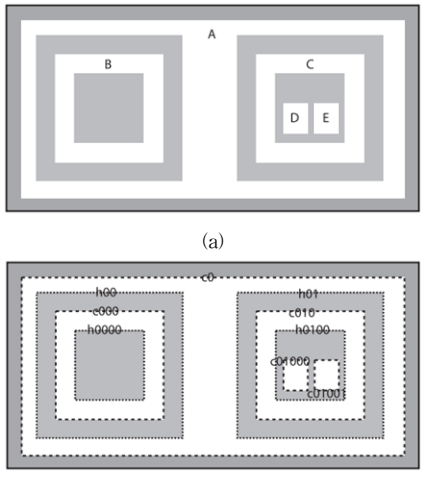


그림 9. 이미지의 에지 변화에 따른 Contour와 Hole 검출
Fig. 9. Detection of Contour and Hole following the Change of Image Edges

그러나 임계치 값 설정만으로 처리하기에는 작은 사

각 영역부터 큰 영역까지 다양하게 분포하게 된다. 신분증을 포함하는 이미지의 특성 3에 따라 contour 영역의 크기가 th_{region}이상인 경우의 영역만 체크한다. 그림 10은 입력 영상과 임계치, 그리고 th_{region}에 따른 검출된 영역을 보여준다. 임계치가 250인 경우 전체 영역이 나타나는 것을 확인할 수 있다.

입력 영상이 신분증을 포함한 영상인지 최종 판단은 둘러싼 사각형 영역의 가로와 세로 길이가 전체 입력 영상에서 신분증으로 판단이 요구되는 최소 크기의 가로와 세로간의 비율보다 큰 영역이 2개 미만일 경우이다.

Ⅳ. 실험 결과

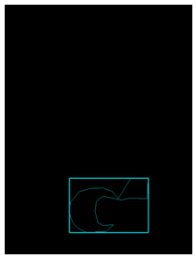
PIAD 시스템은 실행과 함께 설정된 기본 디렉토리를 검색하고, 사용자가 임의의 시간을 설정하여 검색할 수 있도록 한다. 화면 오른쪽에는 검색된 결과를 이미지 썸네일로 표현하여 어떤 이미지가 검색되었는지 보여주며, 엑셀, PDF 등의 문서 내에 포함되어 있는 전화번호, 주민번호, 여권번호의 개인 정보를 문자열 패턴 인식으로 검색할 수 있는 기능을 제공한다. PIAD 시스템의 환경은 표 1과 같으며, 시스템에 사용된 여러 임계치 값은 테스트를 통해서 얻은 상수값이다.

표 1. 시스템 환경 및 임계치 상수값
Table 1. System Environment and Threshold Values

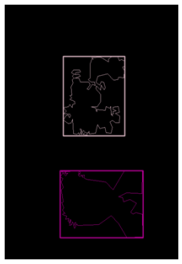
항 목	내 용	
PC 환경	Windows 7 64 bit, i5-3230M 2.60GHz, 4.0G Memory	
임계치	thhist_low	60 %
	thhist_high	99.5 %
	thsurf	110개 이상
	thsurf_match	thsurf의 20% 이상
	thregion	250

(a) 입력 이미지 (원 이미지)

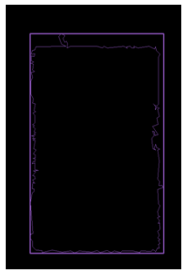
(b) 입력 이미지 (신분증 확대)



(c) 임계치 100



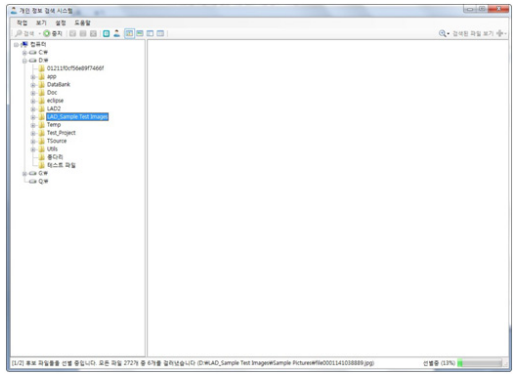
(d) 임계치 200



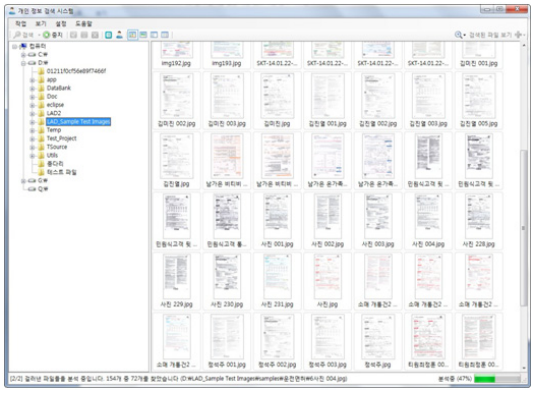
(e) 임계치 250

그림 10. 입력 영상과 임계치에 따른 검출 영역
Fig. 10. Input image and Detection Area with threshold

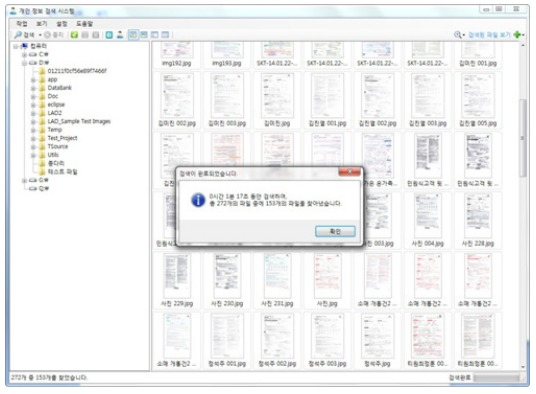
그림 11은 PIAD 시스템의 사용자 인터페이스로서 검색하고자 하는 디렉토리를 선택하고 검색을 시작하면 (b)와 같이 선별 과정을 거치고 (c)는 선별된 이미지들의 분석에 대한 결과를 사용자에게 보여준다. 사용자는 검색된 결과를 직접 확인할 수 있고 일괄 또는 부분 삭제가 가능하다.



(a) PIAD 초기 사용자 인터페이스 및 선별 진행 중인 상태



(b) 분석 진행 중인 상태



(c) 분석 완료 및 검색 결과를 보여줌

그림 11. PIAD 시스템 사용자 인터페이스
Fig. 11. PIAD System User Interface

본 논문에서 제안한 PIAD 시스템의 성능을 평가하기 위하여 입력 영상은 총 272장을 사용하였다. 테스트 영상은 표 2와 같이 판매점에서 스캔한 개인 정보 이미지, 개인 정보를 포함하는 계약서 이미지, 일반 이미지, 그리고 텍스트가 많이 포함되어 있는 이미지 등 다양한 형태를 가진다. 본 논문에서 제안한 3가지 특성에 따라 N 세트는 선별 단계에서 검출이 되고, T 세트는 이미지 내에 포함되는 색상 정보에 따라 선별 단계에서 검출이 될 수도 있고 그렇지 않을 수 있다. 그러나 분석 단계에서는 모두 미검출이 되어야 한다. L 세트와 C 세트는 선별 단계에서는 검출되지 않아야 하며 모두 분석 단계에 포함되어야 한다.

표 2. 테스트 이미지 분류
Table 2. Category of Test Images

이미지 타입	장수	개인 정보	처리단계	
			선별	분석
운전면허증, 주민등록증 (L Set)	82	O	미검출	검출
개인 정보를 포함하는 계약서, 신청서 (C Set)	70	O	미검출	검출
텍스트가 많이 포함된 이미지 (T Set)	32	X	검출 또는 미검출	미검출
일반 영상 - 자연, 인물, 동물, 물체 (N Set)	88	X	검출 또는 미검출	미검출

PIAD 시스템이 처리하는 시간은 메모리의 상태 및 현재 사용자가 사용 중인 운영체제의 환경에 의존적이기 때문에 30분 간격으로 10회의 처리 시간을 측정하여 그 평균을 확인하였다. 그림 12는 각 회차별 처리 시간을 나타내는 그래프로써, 10회 평균의 결과는 78.5초가 나왔다. 본 시스템은 일반 이미지와 분석할 이미지를 빠르고 정확하게 선별하여 이미지 분석 시간을 최소화 할 수 있는 장점을 가진다.

그러나 처리 시간이 빠르다고 해서 좋은 성능은 될 수 없다. 즉, 개인 정보를 포함하는 이미지를 정확하게 검색했는지를 확인해야 하며, 검색하기 위해 제안한 분석 과정으로 SURF, FindContour, 노이즈 제거의 결과는 표 3과 같다.

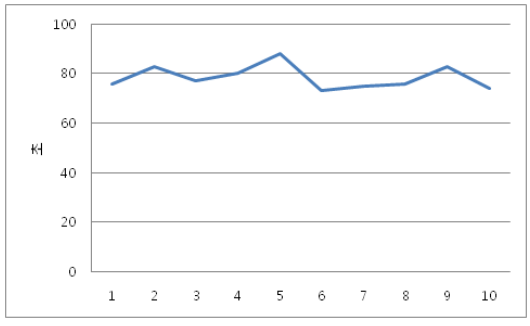


그림 12. PIAD 처리 시간
Fig. 12. PIAD Running Time

표 3. 이미지별 처리 결과
Table 3. Result according to Images Types

이미지 타입	장수	처리단계			
		선별		분석	
		검출	미검출	검출	미검출
L Set	82	**4**	76	**75**	1
C Set	70	**0**	0	**70**	0
T Set	32	**26**	6	**0**	6
N Set	88	**88**	0	**0**	0

표 3에서 T 세트는 이미지 특성상 흰색 배경이 많기 때문에 선별에서 미검출된 이미지수가 나타나지만 분석에서 미검출이 되고 있다. 이것은 SURF 알고리즘과 템플릿 매칭을 통해 동일한 텍스트 정보가 없기 때문에 미검출 이미지로 남게 된다. 반면에 C 세트는 선별 단계에서 검출되지 않고 분석 단계에서 100% 검출이 된 것을 확인할 수 있다. 즉, 본 논문에서 제시한 분석 방법과 이미지 특성이 일치한다는 것을 알 수 있다. L 세트는 스캔된 이미지 정보가 본 논문에서 제시한 특성이 아닌 경우 선별 단계에서 미리 검출되어지는 문제점이 있는 것을 확인하였다. 그림 13과 같이 이미지 전체가 신분증인 경우 일반 이미지로 판단하게 되고, 스캔된 이미지의 원본이 많이 접혀 있어서 FindContours를 이용한 ROI를 얻지 못해서 분석 단계에서 미검출 되는 경우가 발생하여 정확도가 90% 정도가 되었다. 본 논문에서 제시한 PIAD 시스템을 통해 실험한 결과 전체 입력 영상 272장 가운데 선별 단계에서 118장을 선별하였고, 분석 단계에서 147장을 찾았고 5장을 미검출하였다. 즉, 전체 272장 중 267장에 대한 정확한 정보를 추출하였다.

(a) 신분증 전체가 스캔된 이미지

(b) 스캔된 원본이 특이한 이미지

그림 13. 미검출된 이미지의 특징
Fig. 13. Characteristic of Detection Failed Images

V. 결 론

본 논문에서는 통신사와 판매사들의 개인 정보 유출에 대한 피해 사례가 늘어남에 따라 개인 정보에 대한 효율적인 관리를 위한 PIAD 시스템을 제안하였다. 제안한 PIAD 시스템은 통신 판매사에서 관리하는 고객 신분증 이미지, 고객의 정보가 포함된 계약서 또는 신청서, 그리고 문서들을 자동으로 검색, 추출하여 사용자가 삭제할 수 있게 한다. 실험 결과로는 검색된 272장의 이미지들 중 정상적으로 처리된 이미지가 267장으로써 98% 이상의 정확도를 보였다. 본 논문에서 제안한 방법은 통신 판매사들이 보다 빠르고 안전하게 고객 정보를 관리함으로써 개인 정보 유출을 차단하고 보호할 수 있는 시스템으로 기대된다.[19]

References

[1] The Statics Portal, “Number of smart phones sold to end users worldwide from 2007 to 2013 (in million units)”, http://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007/.

[2] Korea Internet Security Agency, “Study on Usage Patterns of Age and Gender for Smartphone Users”, Internet & Security Focus, pp. 35-51, 2013.

[3] “Personal Information Loss of Mobile Service Provider”, http://www.zdnet.co.kr/news/news_view.asp?artice_id=20140311154043.

[4] Korea Internet Security Agency, “Counselling Count of Personal Information Protection”, http://isis.kisa.or.kr/.

[5] Ministry of Security and Public Administration, “The Personal Information Protection Act”.

[6] H. Y. Hwang and N. Y. Kim, “Personal Information Protection System for Web Service”, The Journal of The Institute of Internet, Broadcasting and Communication, VOL. 11, No. 6, pp. 261-266, December 2011.

[7] W. J. Kang, “An Efficient Privacy Preserving Method based on Semantic Security Policy

Enforcement", The Journal of The Institute of Internet, Broadcasting and Communication, VOL. 13, No. 6, pp. 173-186, December 2013.

[8] J. K. Baek and J. P. Park, "A Study on Personal Information Control and Security in Printed Matter", Journal of Korea Academia-Industrial cooperation Society, Vol. 14, No. 5, pp. 2415-2421, 2013,

[9] B. Kim, J. I. Lim, Y. H. Jo, "Privacy Situation and Countermeasures of Financial Apps based on the Android operating System", The Journal of The Institute of Internet, Broadcasting and Communication, VOL. 14, No. 6, pp. 267-272, December 2014.

[10] M. S. Seo, and D. W. Park, "The Solution for Personal Information Protection Act of PC", Journal of the Korea Institute of Information Security and Cryptology, Vol. 22, No. 8, pp. 21-25, 2012.

[11] J. H. Cho, C. W. Ahn, and J. H. Jun, "A Feature Point Tracking Method By Using Template Matching and Buffer", The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), VOL. 14, NO. 4, pp.173-179, 2014.

[12] H. J. Lee, "Character Recognition for Machine Reader Zone of Electronic Identity Card", Master´s Thesis, Ajou University, 2014.

[13] S. D. Park, Y. U. Woo and G. B. Kim, "Recognition of Resident Registration Cards Using ART-1 and PCA Algorithm", Journal of Korea Institute of Information and Communication Engineering, Vol. 11, No. 9, pp. 1786-1792, 2007.

[14] R. Sankar and G. Ivkovic, "An Adaptive Image Quality Assessment Algorithm", International Journal of Advanced Smart Convergence(IJASC), Vol. 1, No. 1, pp. 6-13, 2012.

[15] L. Neumann and J. Matas, "Real-Time Scene Text Localization and Recognition," the 25th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3538-3545, 2012.

[16] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," European Conference on Computer Vision, Vol. 3951, pp. 404-417, 2006.

[17] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int'l J. Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.

[18] S. Suzuki and K. Abe, "Topological structural analysis of digital binary images by border following," Computer Vision, Graphics and Image Processing, Vol. 30, pp. 32-46, 1985.

[19] J. S. Kim, " A Study on V.M.D(Visual Merchandising Design) Environment of Mobile Telecommunication Company Store", Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 14, No. 4, pp. 1589-1594, 2013.

저자 소개

조 정 현(정회원)

- 1988년 : 경북대학교 전자공학과(학사)
- 1990년 : 경북대학교 컴퓨터공학과(공학석사)
- 2005년 : 경북대학교 컴퓨터공학과(공학박사)
- 2012 ~ 현재 : 영남이공대학교 컴퓨터정보과 교수

<주관심분야 : 영상처리, 네트워크>

안 철 웅(정회원)

- 1993년 : 경북대학교 컴퓨터공학과(학사)
- 1995년 : 경북대학교 컴퓨터공학과(공학석사)
- 2009년 : 경북대학교 컴퓨터공학과(공학박사)
- 2001년 ~ 현재 : 계명문화대학교 디지털콘텐츠학부 교수

<주관심분야 : 멀티미디어, 이미지 처리, 빅데이터 처리>