



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년03월29일
(11) 등록번호 10-1721063
(24) 등록일자 2017년03월23일

(51) 국제특허분류(Int. Cl.)
G06K 9/32 (2006.01) G06K 9/62 (2006.01)
(52) CPC특허분류
G06K 9/3258 (2013.01)
G06F 21/6245 (2013.01)
(21) 출원번호 10-2016-0163392
(22) 출원일자 2016년12월02일
심사청구일자 2016년12월02일
(56) 선행기술조사문헌
KR1020090024563 A
KR1020160068441 A
KR1020160091307 A

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
원유재
경기도 성남시 분당구 미금로 215, 813동 1901호
(금곡동, 청솔마을대원아파트)
김숙현
대전광역시 유성구 대학로81번길 32-36 306호
(뒷면에 계속)
(74) 대리인
특허법인 공간

전체 청구항 수 : 총 5 항

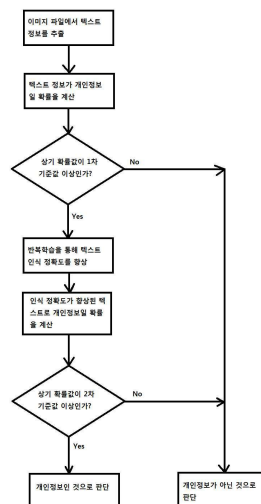
심사관 : 이종경

(54) 발명의 명칭 이미지 파일에 포함된 개인정보 검색 방법 및 그 방법을 구현하는 프로그램을 기록한 기록매체

(57) 요약

본 발명에 의한 개인정보 검색 방법은, 이미지 파일에서 텍스트 정보를 추출하는 제1 단계; 텍스트 정보가 개인정보일 확률을 계산하는 제2 단계; 제2 단계의 확률값이 1차 기준값 미만이면 개인정보가 아닌 것으로 판단하는 제3 단계; 제2 단계의 확률값이 1차 기준값 이상이면 반복학습을 통해 텍스트 인식 정확도를 향상시키는 제4 단계; 제4 단계에서 인식된 텍스트로 개인정보일 확률을 계산하는 제5 단계; 제5 단계의 확률값이 2차 기준값 이상이면 개인정보인 것으로 판단하는 제6 단계;를 포함하는 것을 특징으로 한다.

대표도 - 도4



(52) CPC특허분류

G06K 9/6226 (2013.01)

(72) 발명자

윤지애

충청북도 청주시 흥덕구 오송읍 오송생명3로 111,
605동 1201호(오송호반베르디움아파트)

이민석

대전광역시 서구 청사로 282, 18동 1508호(둔산동,
수정타운)

이 발명을 지원한 국가연구개발사업

과제고유번호 R7115-16-1007

부처명 미래창조과학부

연구관리전문기관 정보통신기술진흥센터

연구사업명 정보통신기술인력양성

연구과제명 2016년 SW중심대학_충남대

기 여 율 1/1

주관기관 충남대학교 산학협력단

연구기간 2016.03.01 ~ 2017.02.28

명세서

청구범위

청구항 1

이미지 파일에서 텍스트 정보를 추출하는 제1 단계;

텍스트 정보가 개인정보일 확률을 계산하는 제2 단계;

제2 단계의 확률값이 1차 기준값 미만이면 개인정보가 아닌 것으로 판단하는 제3 단계;

제2 단계의 확률값이 1차 기준값 이상이면 반복학습을 통해 텍스트 인식 정확도를 향상시키는 제4 단계;

제4 단계에서 인식된 텍스트로 개인정보일 확률을 계산하는 제5 단계;

제5 단계의 확률값이 2차 기준값 이상이면 개인정보인 것으로 판단하는 제6 단계;

를 포함하는 것을 특징으로 하는 이미지 파일에 포함된 개인정보 검색 방법.

청구항 2

청구항 1에 있어서,

상기 제4 단계의 반복학습의 횟수는 4회인 것을 특징으로 하는 개인정보 검색 방법.

청구항 3

청구항 1에 있어서,

상기 제4 단계의 반복학습은, 인식된 텍스트로 개인정보일 확률을 판단할 때, 개인정보일 확률 향상도가 소정의 기준의 이하가 될 때까지 반복하는 것을 특징으로 하는 개인정보 검색 방법.

청구항 4

청구항 1에 있어서,

상기 제4 단계에서 추출한 텍스트는 숫자로만 구성된 것을 특징으로 하는 개인정보 검색 방법.

청구항 5

컴퓨터에 설치되는 컴퓨터 프로그램을 저장하는 기록매체로서, 상기 컴퓨터에 설치되면 상기 컴퓨터가 청구항 1 내지 청구항 4 중 어느 하나의 방법을 실시하도록 하는 컴퓨터 프로그램을 저장하는 기록매체.

발명의 설명

기술 분야

[0001] 본 발명은 이미지 파일에 포함된 개인정보 검색 방법 및 그 방법을 구현하는 프로그램을 기록한 기록매체에 관한 것으로, 더욱 상세하게는, 이미지 파일에서 개인정보를 찾아낼 때 계산량이 적으면서도 개인정보 검출 정확도가 높은 개인정보 검색 방법 및 그 방법을 구현하는 프로그램을 기록한 기록매체에 관한 것이다.

배경 기술

[0002] 개인정보보호법의 시행으로 개인정보를 취급하는 기관 및 기업에서 개인정보를 안전하게 관리하는 시스템을 구

축하는 것이 필수적이다. 텍스트 형태로 저장되어 있는 경우는 개인정보 관리가 비교적 용이하지만, 이미지 파일로 저장된 개인정보의 경우 개인정보인지를 판단하는 것조차도 용이하지 않다. 예를 들면 이동통신서비스 가입시 주민등록증, 계약서 등의 스캔 이미지에는 개인정보가 기재되어 있을 가능성이 크지만 그 스캔 이미지 내의 이미지화된 텍스트 개인정보를 찾아내는 것이 용이하지 않다.

[0003] 특허문헌 1은 개인정보 자동 검색 방법에 관한 것으로, 이미지 파일에서 개인 정보를 검출하는 기술이 기재되어 있다.

[0004] 그러나 특허문헌 1에 기재된 방법은 필요한 계산량에 비해 개인정보를 찾아내는 정확도가 높지 않다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) KR 10-2016-0016201 A(공개일: 2016. 2. 15.)

발명의 내용

해결하려는 과제

[0007] 본 발명에서 해결하고자 하는 과제는, 이미지 파일에서 개인정보를 찾아낼 때 계산량이 적으면서도 개인정보 검출 정확도가 높은 개인정보 검색 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 본 발명에 의한 개인정보 검색 방법은, 이미지 파일에서 텍스트 정보를 추출하는 제1 단계; 텍스트 정보가 개인정보일 확률을 계산하는 제2 단계; 제2 단계의 확률값이 1차 기준값 미만이면 개인정보가 아닌 것으로 판단하는 제3 단계; 제2 단계의 확률값이 1차 기준값 이상이면 반복학습을 통해 텍스트 인식 정확도를 향상시키는 제4 단계; 제4 단계에서 인식된 텍스트로 개인정보일 확률을 계산하는 제5 단계; 제5 단계의 확률값이 2차 기준값 이상이면 개인정보인 것으로 판단하는 제6 단계; 를 포함하는 것을 특징으로 한다.

[0010] 상기 제4 단계의 반복학습의 횟수는 4회일 수 있다.

[0011] 상기 제4 단계의 반복학습은, 인식된 텍스트로 개인정보일 확률을 판단할 때, 개인정보일 확률 향상도가 소정의 기준의 이하가 될 때까지 반복하는 것일 수 있다.

[0012] 상기 제4 단계에서 추출한 텍스트는 숫자로만 구성된 것일 수 있다.

[0013] 컴퓨터에 설치되면 컴퓨터가 상기 방법을 실시하도록 하는 컴퓨터 프로그램은 컴퓨터 프로그램을 저장하는 기록 매체에 저장될 수 있다.

발명의 효과

[0015] 본 발명의 개인정보 검색 방법은, 이미지 파일에서 개인정보를 찾아낼 때 계산량이 적어서 검출시간이 짧으면서도 개인정보 검출 정확도가 높다.

도면의 간단한 설명

[0017] 도 1은 이미지 파일로부터 개인정보를 추출한 결과 화면의 예

도 2는 문자 인식의 학습을 명령하는 화면의 예

도 3은 주민등록번호일 확률을 판단하는 예

도 4는 본 발명에 의한 개인정보 검색 방법의 흐름도의 예

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명은 다양한 변형을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아

니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

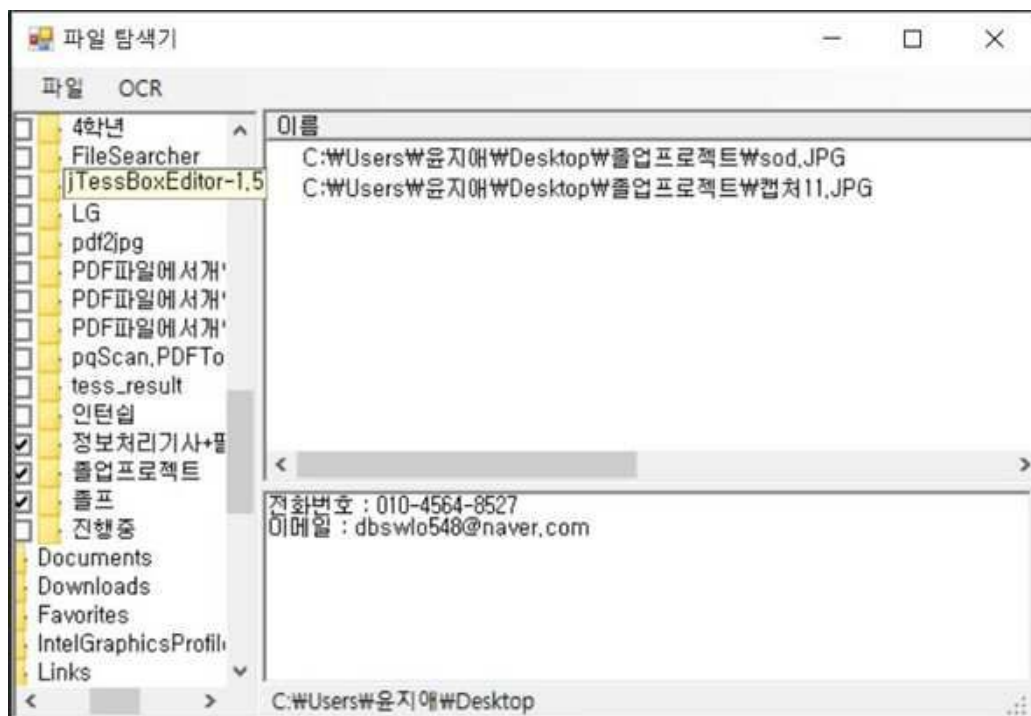
- [0019] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.
- [0021] 아래에서는 본 발명을 첨부된 도면을 통해 더욱 상세히 설명한다.
- [0023] OCR(Optical Character Recognition)은 광학적인 방법으로 인쇄 혹은 수기로 된 문자를 인식하여 디지털로 인코딩하는 기술이다. 이 기술을 이용하여 이미지 파일에서 문자를 판독해서 정리하고 자동으로 컴퓨터에 텍스트 파일로 저장할 수 있다.
- [0024] 현재 다양한 OCR이 상용화되어 있다. 그 예로, TesseractOCR은 다양한 문자인식 서비스나 이메일 스팸 필터링에도 적용하고 있다. 또한 TesseractOCR은 낮은 인식률을 문자학습을 통해 향상시킬 수도 있다.
- [0026] 이미지 파일에서 텍스트를 추출해낸 경우, 그 텍스트와 개인정보의 유사도를 비교하여 그 텍스트가 개인정보일 확률을 계산해낼 수 있다.
- [0027] 유사도를 비교할 때 리벤슈타인 디스턴스 알고리즘(Levenshtein Distance Algorithm)을 이용할 수도 있다. 리벤슈타인 디스턴스 알고리즘은 두 개의 단어에서 유사성을 검사하는 알고리즘으로써 두 개의 문자열을 비교할 때 사용한다.
- [0029] 도 1은 이미지 파일로부터 개인정보를 추출한 결과 화면의 예이다.
- [0030] 도 1에서는, 이미지 파일로부터 개인정보인 전화번호와 이메일을 추출해내서 표시하고 있다.
- [0031] 도 2는 문자 인식의 학습을 명령하는 화면의 예이다.
- [0032] 도 2의 화면에서는, 문자 인식의 학습을 명령하는 “Training”이라는 버튼이 존재한다.
- [0034] 확률을 계산하는 방법의 예로는 문자인식으로 추출된 데이터와 사전에 프로그램에 등록한 개인정보 패턴과 비교하여 계산할 수 있다.
- [0035] 도 3은 주민등록번호일 확률을 판단하는 예이다.
- [0036] 도 3과 같이 주민등록번호 패턴과 추출된 데이터를 비교한다고 가정할 때, 각 한자리씩 매칭하여 주민등록번호 패턴과 유사한지 비교를 한다.
- [0037] 000000-0000000 은 총 14개의 자리수로 되어있는 패턴으로, 추출데이터와 비교하였을 때 13개의 자리수가 맞다면 $(13/14) \times 100 = 92.9\%$ 로 결과가 나타나게 된다.
- [0038] 따라서이미지파일에서 개인정보가 포함될 가능성은 92.9%로 판단할 수 있다.
- [0039] 본 발명에서는 주민등록번호, 이메일, 핸드폰번호 등의 개인정보를 검출하는 것을 목적으로 한다.
- [0040] 그런데 이미지 파일로부터 텍스트를 추출해낼 때, 최초로 추출해낸 텍스트는 부정확할 수 있고 학습을 반복할수록 정확도는 증가한다.
- [0041] TesseractOCR(광학문자인식기술 오픈소스 엔진의 한 종류)로 문자인식을 한 후 리벤슈타인 디스턴스 알고리즘(Levenshtein Distance Algorithm)을 이용하여 개인정보일 확률을 구하는 실험을 했더니, 학습을 반복할수록 개인정보일 확률이 증가하였고, 일정횟수의 학습이 반복되면 그 확률은 더 이상 증가하지 않았다. 실험결과에 의하면 4회 학습한 이후에는 더 이상 확률이 증가하지 않는 경우가 많았다.
- [0042] 1회 추가 후에도 개인정보일 확률이 증가한 정도가 소정의 기준치(예를 들면, 1%) 이하이면, 한계구간에 도달하여 더 이상 개인정보일 확률이 증가하지 않는 것으로 판단할 수 있다.
- [0044] 본 발명에서는 계산량을 줄이기 위해 최초의 텍스트 추출에서 개인정보일 확률이 1차 기준값(예를 들면 20%) 미만이면, 개인정보가 아닌 것으로 판단하여 더 이상의 문자 인식 학습을 않는다.
- [0045] 그러나 1차 기준값 이상이면, 반복 학습을 통해 텍스트 인식 정확도를 향상시킨 후 개인정보일 확률을 다시 판단하고, 그 확률이 2차 기준값(예를 들면 80%) 이상이면, 개인정보인 것으로 판단한다.
- [0046] 상기 반복 학습은, 개인정보일 확률 향상도가 소정의 기준의 이하가 될 때까지 반복할 수도 있고, 실험의 결과,

확률 향상의 한계치에 도달하는 비율이 높은 4회만 반복할 수도 있다.

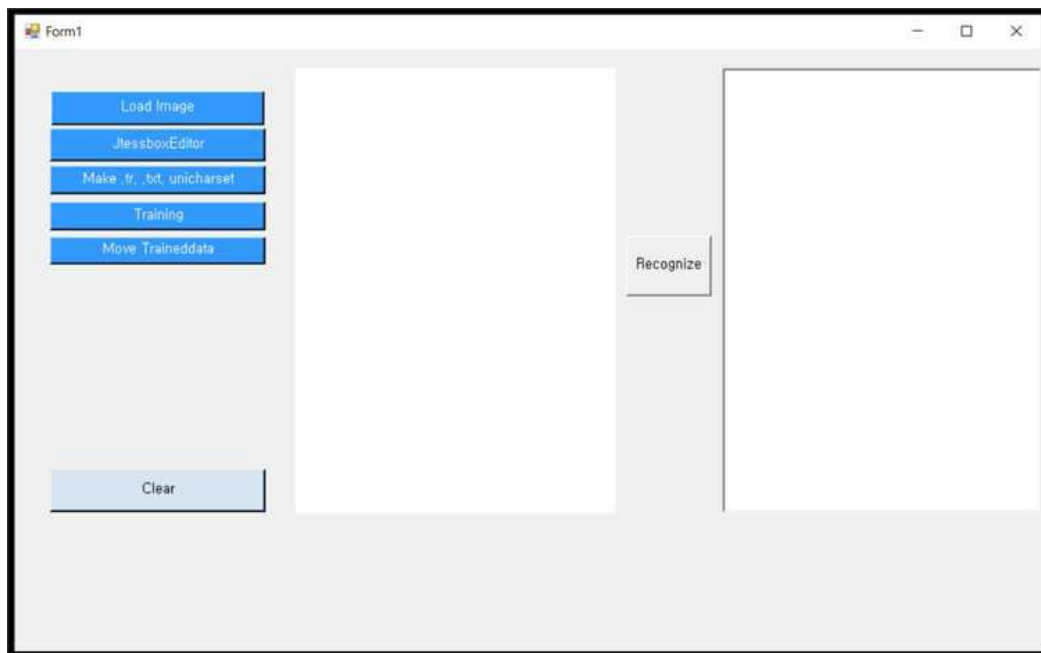
- [0047] 또한 숫자만을 읽어들이도록 하면 모든 문자를 읽어들이도록 할 때보다 OCR에 의한 인식이 용이하므로, 숫자만을 읽어들이어 개인정보 여부를 판단하도록 구현하면 계산량을 줄일 수 있다.
- [0049] 도 4는 본 발명에 의한 개인정보 검색 방법의 흐름도의 예이다.
- [0051] 따라서 본 발명에 의한 개인정보 검색 방법은 다음의 단계를 포함한다.
- [0052] (1) 제1 단계: 이미지 파일에서 텍스트 정보를 추출하는 단계
- [0053] (2) 제2 단계: 텍스트 정보가 개인정보일 확률을 계산하는 단계
- [0054] (3) 제3 단계: 제2 단계의 확률값이 1차 기준값 미만이면 개인정보가 아닌 것으로 판단하는 단계
- [0055] (4) 제4 단계: 제2 단계의 확률값이 1차 기준값 이상이면 반복학습을 통해 텍스트 인식 정확도를 향상시키는 단계
- [0056] (5) 제5 단계: 제4 단계에서 인식된 텍스트로 개인정보일 확률을 계산하는 단계
- [0057] (6) 제6 단계: 제5 단계의 확률값이 2차 기준값 이상이면 개인정보인 것으로 판단하는 단계
- [0059] 만일 제2 단계에서 구한 확률이 2차 기준값 이상이면 바로 개인정보인 것으로 판단할 수도 있다.
- [0060] 따라서 제3 단계와 제4 단계는 다음과 같이 변경될 수도 있다.
- [0062] 제3 단계: 제2 단계의 확률값이 2차 기준값 이상이면 개인정보인 것으로 판단하고 1차 기준값 미만이면 개인정보가 아닌 것으로 판단하는 단계
- [0063] 제4 단계: 제2 단계의 확률값이 1차 기준값 이상이고 2차 기준값 미만이면 반복학습을 통해 텍스트 인식 정확도를 향상시키는 단계
- [0065] 본 발명에 의한 개인정보 검색 방법이 컴퓨터상에서 구현되려면, 컴퓨터에 설치될 때 컴퓨터가 본 발명에 의한 개인정보 검색 방법을 실시하도록 하는 컴퓨터 프로그램이 필요하다.

도면

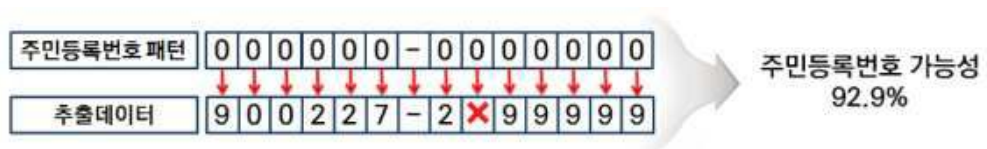
도면1



도면2



도면3



도면4

