텍스트 추출을 위한 모바일 응용 구현

고은비*, 박영호*
*숙명여자대학교 멀티미디어과학과
e-mail: yhpark@sm.ac.kr

An Implementation of a Mobile Function for Extracting and Retrieving as Text

En-Bee Go*, Young-Ho Park*
*Dept of Multimedia Science, Sookmyung University

요 약

본 논문에서는 다양한 상황에서의 정보 접근성을 향상시키기 위해 이미지를 검색 수단으로 사용하는 검색 시스템을 제안한다. 본 콘텐츠는 안드로이드 플랫폼 기반의 스마트폰에서 이미지를 얻어 텍스트를 추출하고, 이를 검색 엔진의 키워드로 입력하여 그 결과를 출력하는 과정을 거친다. 또한, 검색 결과를 스마트폰의 내장 데이터베이스에 저장하고, 이를 관리하여 추후에 재사용할 수 있도록 한다.

1. 서론

본 논문에서는 이미지를 이용하여 모바일 검색을 실시하는 이미지 내 텍스트 검색 기능을 소개한다. 이는 모바일 기기를 통해 얻은 이미지에서 텍스트를 추출한 뒤, 추출한 텍스트를 상용 검색 엔진의 키워드로 활용하여 정보 검색을 하는 단계로 진행된다. 본 논문은 기존 상용 기능인 OCR을 직접 개발하지 않고 개발된 오픈 소스 OCR 엔진을 사용하였으며, 안드로이드 플랫폼 기반의 모바일 검색을 위한 콘텐츠를 구현하였다. 안드로이드(Android)는 무료로 제공하는 최초의 오픈 소스 모바일 플랫폼[1]으로 다음과 같은 기능을 제공한다. 첫째, 스마트폰을 사용하여 이미지 내에 존재하는 텍스트를 추출하고, 추출한 텍스트를 이용하여 즉각적인 모바일 검색을 실시하는 콘텐츠를 제안한다. 둘째, 제안하는 시스템은 안드로이드 플랫폼에서 제공하는 오토 포커스 기능을 사용하고, 이미지 내의 텍스트 영역을 지정하여 인식 범위를 축소하는 텍스트 크롭 기능을 구현하여 텍스트 인식률을 높인다. 셋째, 검색 결과를 스마트폰의 내장 데이터베이스에 저장하고, 이를 추후에 재검색하여 학습할 수 있도록 한다.

2. 관련 연구

본 장에서는 오픈 소스 OCR 엔진인 Tesseract와 웹에서 OCR엔진을 구동시키는 WeOCR에 대해 설명한다. Tesseract는 1984년부터 약 10년간 Hewlett and Packard에서 개발한 오픈 소스 엔진이다[2]. 개발 이후 지속적인 성능 개선을 통해 2005년에 Tesseract를 오픈 소스로 발표하였다. Tesseract는 레이아웃에 대한 분석이 존재하지 않아 다수의 텍스트 라인으로 이루어진 이미지는 왜곡될 수 있다. 그러므로, 입력 이미지는 반드시 TIFF 형식이어야 한다[3]. 본 논문에서 Tesseract는 OCR 서버의 한 부품으로 사용되며, 텍스트를 추출하는 주요 기능을 수행한다는 차이점이 있다. 다음으로는 WeOCR이다. WeOCR은 네트워크 상에서 문자 인식을 가능하도록 하는 OCR 시스템의 플랫폼이다. WeOCR은 사용자로부터 이미지를 입력받아 텍스트를 인식한 뒤, 그 결과를 사용자에게 전송한다[4]. 그러나 WeOCR은 자체의 OCR 엔진을 사용하지 않으므로, 오픈 소스 OCR 엔진과 연동하여 텍스트를 인식한다. 본 논문에서 WeOCR은 OCR 엔진을 웹에서 편리하게 사용할 수 있도록 하는 툴킷 역할을 하며, Tesseract와 함께 OCR 서버를 구성한다. 또한, 모바일 클라이언트와 Tesseract 간의 파일 변환과 데이터 전송을 실시한다는 차이점이 있다.

3. 이미지 내 텍스트 검색 기능

3.1 개요

본 시스템은 OCR 서버와 모바일 클라이언트로 구성되어 있다. OCR 서버는 모바일 클라이언트가 전송하는 이미지에서 텍스트를 인식, 추출한다. 모바일 클라이언트는 사용자의 요청을 받아 OCR 서버

에 전달하고 OCR 서버의 응답을 사용자에게 보여준다. 모바일 클라이언트는 OCR 서버의 텍스트 인식률을 높이기 위해 오토 포커스 기능과 텍스트 크롭 기능을 제공한다. 오토 포커스 기능은 모바일 클라이언트에 내장되어 있는 카메라 초점을 맞추는 역할을 수행하고, 텍스트 크롭 기능은 이미지의 텍스트 영역을 축소하여 텍스트 인식률을 높인다. 이 시스템은 이미지를 이용하여 모바일 검색을 실시하기 때문에 다양한 상황에서의 검색이 가능하도록 한다. 이를 통해 모바일 환경에서의 정보 접근성을 높일 수 있다.

3.2 시스템 구조

모바일 클라이언트는 텍스트 검색을 통해 OCR 서버로 이미지를 전송하여 추출한 결과 텍스트를 상용 검색 엔진의 키워드로 이용하여 모바일 검색을 한다. 키워드로 사용된 텍스트는 모바일 클라이언트의 내부 데이터베이스인 검색내역 DB에 저장되고, 검색내역 조회를 통해 열람 및 재검색이 가능하다. OCR 서버는 모바일 클라이언트로부터 받은 이미지에서 인식, 추출한 텍스트를 반환하며, WeOCR과 Tesseract로 구성된다. Tesseract는 이미지 내의 텍스트를 인식, 추출하고, WeOCR은 Tesseract와 모바일 클라이언트를 연결하는 역할을 수행한다.

3.3 통신 방법

제안하는 시스템의 작동 과정을 텍스트 검색 기능과 검색내역 조회 기능으로 나누어 나타내며, 다음과 같은 순서로 진행된다. 모바일 클라이언트가 내장 카메라나 사진첩을 통해 이미지를 얻은 뒤, 텍스트 크롭을 사용하여 이미지 내의 텍스트 영역을 잘라내고, 이를 OCR 서버로 전송한다. OCR 서버의 WeOCR은 이미지를 TIFF 형태로 변환하여 Tesseract에 전송하면, Tesseract는 이미지에서 추출한 텍스트를 텍스트 파일 형태로 WeOCR에 전송한다. WeOCR은 파일 내의 텍스트를 모바일 클라이언트에 전달한다. 모바일 클라이언트는 검색에 사용할 사용 검색엔진을 선택하고, OCR 서버로부터 얻은 텍스트를 이용하여 모바일 검색을 실시한다. 텍스트 검색 기능을 수행하면서 저장한 추출 텍스트 관련 정보를 이용하여 재검색이 가능하도록 하였다. 재검색은 추출 결과 텍스트를 상용 검색 엔진의 키워드로 전송하면 검색 결과가 모바일 클라이언트에 전송되는 과정을 거친다.

3.4 OCR 서버

OCR 서버는 Apache, PHP, 요청처리 프로그램, WeOCR, Tesseract로 구성된다. Tesseract는 오픈소스로 제공되는 OCR 엔진이며, 이미지에서 텍스트를 인식, 추출하는 중추적 역할을 수행한다. Tesseract의 입력 이미지는 TIFF형태로 제한되어 있기 때문에 모바일 클라이언트로부터 전달받은 이미지를 적합한 형태로 변환하는 과정이 필요하다. WeOCR은 요청 처리 프로그램으로부터 입력받은 이미지를 Tesseract에서 사용가능한 형태로 변환해 주고, Tesseract의 텍스트 추출 결과를 요청 처리 프로그램에 전달하는 중간 연결부의 역할을 한다. Tesseract의 텍스트 추출 결과는 임시 텍스트 파일로 저장되고, OCR 서버 프로그램은 이 파일을 파싱하여 모바일 클라이언트에 텍스트 형태로 응답한다. 이 과정에서 텍스트 추출 결과 파일을 임시로 저장함으로써 저장 공간의 낭비를 줄이는 효과를 얻을 수 있다.

3.5 모바일 클라이언트

본 절에서는 모바일 클라이언트의 주요 기능을 텍스트 검색 기능과 검색내역 조회 기능으로 나누어 설명한다.

3.5.1 텍스트 검색 기능

텍스트 검색 기능은 이미지에서 OCR 서버에서 추출한 텍스트를 키워드로 이용하여 모바일 검색을 수행한다. 텍스트 검색 기능은 다음과 같은 단계로 진행된다. 첫째, 모바일 클라이언트가 내장 카메라를 이용하거나 내부 메모리에서 이미지를 선택한다. 둘째, 이미지를 텍스트 크롭 한다. 텍스트 크롭은 사용자가 가변 프레임을 반복적으로 사용하여 최종적으로 결정한 텍스트 영역을 새로운 파일로 저장한다. 셋째, 저장된 이미지를 OCR 서버로 전송하여 텍스트를 추출하고, 이를 모바일 클라이언트가 전달 받아 키워드로 설정한다. 넷째, 키워드가 사용자 의도와 다른 경우 수정하고, 구글, Dictionary, 네이버, 위키피디아 중 사용자가 선택한 포털의 검색 결과를 출력한다. 추출한 텍스트의 관련 정보는 검색이 수행되는 시점에 모바일 클라이언트의 내장 데이터베이스인 검색내역 DB에 저장되어 검색 내역 조회 기능에 사용된다.

3.5.2 검색내역 조회 기능

검색내역 조회 기능은 모바일 클라이언트의 내장 데이터베이스인 검색내역 DB에 접근하여 저장되어 있는 정보를 조회한다. 저장되어 있는 정보는 텍스

트 검색 기능을 통해 추출한 텍스트에 대한 세부 정보(검색 날짜, 텍스트 크롭한 이미지, 검색 키워드, 검색결과 URL)이다.

4. 구현

4.1 모바일 클라이언트 구현

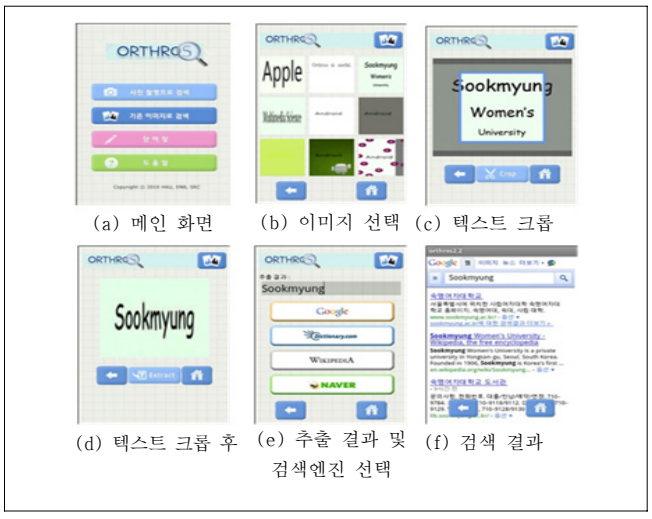
모바일 클라이언트는 윈도우 PC에서 안드로이드 2.2 플랫폼을 기반으로 Eclipse Ganymede를 사용하여 구현하였으며, 테스트 단말기로 HTC 사의 넥서스원(GGL-NX1) 모델을 사용하였다.

4.2 인터페이스 구현

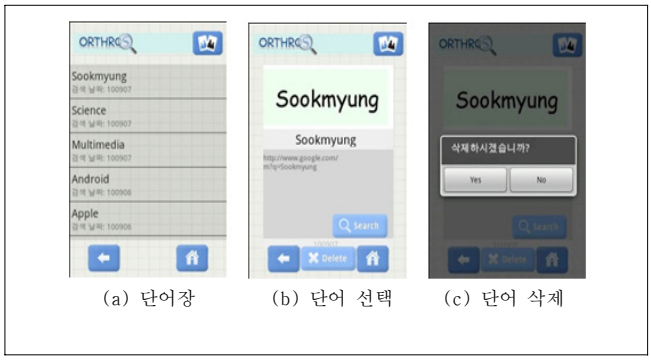
모바일 클라이언트의 인터페이스 구현은 이미지를 이용한 텍스트 검색 기능과 재검색을 위한 검색내역 조회 기능으로 구분하여 설명한다.

그림 1는 모바일 클라이언트의 내장 메모리를 이용한 텍스트 검색 과정순차적으로 나타낸다. 그림 1(a)는 메뉴를 나타내고, '기존 이미지로 검색' 버튼을 클릭하면 그림 1(b)가 나타난다. 하나의 이미지를 선택하고 이에 텍스트 크롭 기능을 적용하면 각각 그림 1(c)와 그림 1(d)를 볼 수 있다. 그림 1(e)는 OCR 서버는 텍스트 추출 결과와 검색하고자 하는 상용 검색 엔진을 선택하는 화면이며, 그림 1(f)는 원하는 검색 엔진으로 검색한 결과이다.

그림 2(a)는 검색내역을 나타내는 화면이고, 그림 2(b)는 검색 세부 정보를 나타낸다. 세부 정보의 삭제 버튼을 선택하면, 삭제 의사를 확인하기 위한 화면이 그림 2(c)와 같이 나타난다.



(그림 1) 텍스트 검색 기능 화면



(그림 2) 검색내역 조회 기능 화면

5. 결론 및 향후 연구

본 논문에서는 사용자의 정보 접근성 증대를 위하여 이미지를 이용한 모바일 검색을 위한 콘텐츠를 제안하였다. OCR 서버는 이미지에서 텍스트를 추출하기 위해 오픈 소스 OCR 엔진인 Tesseract를 사용하고, 모바일 클라이언트는 안드로이드 플랫폼 기반의 스마트폰을 사용하는 모바일 어플리케이션이다. 제안하는 시스템의 기능은 이미지를 이용한 텍스트 검색과 검색내역 조회로 나뉜다. 이미지를 이용한 텍스트 검색 기능은 OCR 서버와 모바일 클라이언트가 연동하여 이미지에서 텍스트를 추출한 뒤, 그 텍스트를 상용 검색 엔진의 키워드로 입력하여 모바일 검색을 실시한다.

참 고 문 헌

[1] G. Chang, C. Tan, G. Li, and C. Zhu, "Developing Mobile Applications on the Android Platform," *Mobile Multimedia Processing: Fundamentals, Methods, and Applications*, Springer, 2010

[2] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. of *Intl' Conf. on Document Analysis and Recognition(ICDAR)*, 2007

[3] "OCR - Optical Character Recognition," https://help.ubuntu.com/community/OCR

[4] "WeOCR Project," http://weocr.ocrgrid.org/