Journal of Korea Multimedia Society Vol. 18, No. 8, August 2015(pp. 943-949)
http://dx.doi.org/10.9717/kmms.2015.18.8.943

광학 문자 인식을 통한 단어 정리 방법

김남규[†], 김동언[††], 김성우[†††], 권순각[††††]

Vocabulary Generation Method by Optical Character Recognition

Nam-Gyu Kim[†], Dong-Eon Kim[††], Seong-Woo Kim[†††], Soon-Kak Kwon[††††]

ABSTRACT

A reader usually spends a lot of time browsing and searching word meaning in a dictionary, internet or smart applications in order to find the unknown words. In this paper, we propose a method to compensate this drawback. The proposed method introduces a vocabulary upon recognizing a word or group of words that was captured by a smart phone camera. Through this proposed method, organizing and editing words that were captured by smart phone, searching the dictionary data using bisection method, listening pronunciation with the use of speech synthesizer, building and editing of vocabulary stored in database are given as the features. A smart phone application for organizing English words was established. The proposed method significantly reduces the organizing time for unknown English words and increases the English learning efficiency.

Key words: OCR, Character Recognition, Words Ordering, Vocabulary

1. 서 론

어학 학습을 위해 단어 정리는 필요하고, 모르는 단어의 검색을 위해서는 다양한 종이 단어장, 단어장 어플리케이션, 사전, 인터넷 검색 등이 사용되고 있다. 그러나, 일반적으로 이러한 도구들은 미리 저장되어 있는 단어를 학습해야 하거나, 학습할 단어 목록을 만들기 위해서 단어마다 키패드 입력을 통한 사전 검색을 수행하고, 종이사전을 찾아서 정리를 해야 하는 단점이 있다.

이를 해결하기 위해서는 문서로부터 단어를 인식하면, 검색을 위한 준비 및 실행 시간을 줄일 수 있을 것이다. 문서 인식은 전단계로 광학 문자 인식이 필요하다. 광학 문자 인식 (OCR : Optical Character Recognition)은 출력된 문서 및 이미지를 스캐닝을 통해 한글, 영문, 숫자 폰트에 대해 편집 가능한 텍스트로 변환하고 저장할 수 있게 해주는 기술이다. OCR의 역사는 1928년에 독일의 G. Taushek가 미리 준비된 몇 개의 표준 패턴문자와 입력문자를 비교하여 표준 패턴문자와 가장 유사한 것을 해당 문자로 선정하는 패턴 매칭기법을 이용한 문자 인식 방법을 특허로 등록하면서 시작되었다.

OCR은 매우 다양한 분야에 응용되고 있으며, 뱅킹, 의료 산업, 보안 문자와 같은 법적 산업, 저장소

※ Corresponding Author : Soon-Kak Kwon, Address: (614-714) Eomgang-ro 995, Busanjin-gu, Busan, Korea, TEL : +82-51-890-1727, FAX : +82-51-890-2629, E-mail : skkwon@deu.ac.kr
Receipt date : July 13, 2015, Revision date : July 16, 2015
Approval date : July 17, 2015
[†] Dept. of Game Engineering, Dongeui University (E-mail: ngkim@deu.ac.kr)
[††] Dept. of Computer Software Engineering, Dongeui University ((E-mail: pmzy@me.com)
[†††] Dept. of Computer Software Engineering, Dongeui University ((E-mail: libero@deu.ac.kr)
[††††] Dept. of Computer Software Engineering, Dongeui University
※ This work (Grants No. C0298779) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2015, and this work was supported by Dong-eui University Grant (2013 AA173).

및 디지털 라이브러리, 인간의 보정 또는 인간의 노력없는 문자 인식, 자동 번호판 인식, 그리고, 필기 인식 등으로 광범위하게 적용되고 있다[1,2,3,4]. OCR은 마이크로소프트와 HP에서 라이브러리를 제공하고 있다. 마이크로소프트는 오피스 도큐먼트 이미징 라이브러리를 제공하고 있으며, 마이크로소프트 오피스 도큐먼트 스캐닝이 스캔한 문서를 편집하는 것을 도와준다[5]. HP는 Tesseract 오픈소스 OCR 엔진을 개발하였으며, Tesseract는 기본적으로 C++라이브러리를 제공하고 있지만 자바 라이브러리로 Tess4J, 안드로이드 라이브러리로는 tesseract-android-tools도 같이 제공해 주고 있다[6,7,8]. Tesseract는 출력 파일의 형식이 존재하지 않으므로 출력 파일을 생성하기 위한 단계는 존재하지 않는다. 이러한 문자 인식 기술은 글자를 정확하게 인식하는 정도를 나타내는 인식률과 인식을 위한 인식 시간이 중요한 요인이다.

본 논문에서는 문서를 촬영하면 소프트웨어가 자동적으로 문서 내 단어를 인식하고, 단어와 단어의 뜻을 정리하여 단어장으로 만들어 단어 학습을 도와주는 방법을 제안한다. 제안한 방법을 통해 문자인식 기능을 이용하고 단어정리를 수행하면 단어 검색을 위한 시간단축이 가능하여 단어 학습 시간을 절약할 수 있을 것이다.

제안한 방법은 문자인식을 위해서는 Tesseract-OCR 기법을 사용하고, 사전 소프트웨어 StarDict의 사전데이터를 사용하며, 안드로이드 플랫폼으로 구현한다. 이러한 방법을 통하여 스마트폰 촬영을 통한 단어정리 및 편집, 이분법을 사용한 사전데이터 탐색, 음성합성을 이용한 발음 듣기, 데이터베이스에 저장하여 단어장 구축 및 편집의 기능을 제공한다.

본 논문은 영어 단어를 정리하는 어플리케이션으로 구현하였지만, 그 외 다양한 언어도 동일하게 적용이 가능할 것으로 기대된다.

2. TESSERACT OCR

오픈 소스 OCR 엔진인 Tesseract은 Hewlett and Packard(HP)에서 개발하였으며, 개발 이후 지속적인 성능 개선을 통해 2005년에 오픈소스로 발표되었다. 현재는 구글이 Tesseract의 일부를 지원하고 있다.

Tesseract의 작동 과정은 Fig. 1과 같다. 첫째, 이미지의 임계값을 이용하여 입력 이미지를 이진화한다. 둘째, 이진화한 이미지의 연결된 구성 요소를 분석하여 각 구성 요소의 외곽선을 추출하고, 이를 이진 데이터로 저장한다. 셋째, 텍스트 라인(text line)을 분석하여 구성요소를 체계화한 뒤, 문자의 자간에 따라 단어 단위로 나눈다. 마지막으로, 단어 단위로 나뉜 요소를 단어 단위와 페이지 단위로 인식한다. Tesseract는 출력 파일의 형식이 존재하지 않으므로 출력 파일을 생성하기 위한 단계는 존재하지 않는다. Tesseract는 조명, 각도, 폰트 모양에 따라 인식률이 저하될 수 있다[9,10,11].

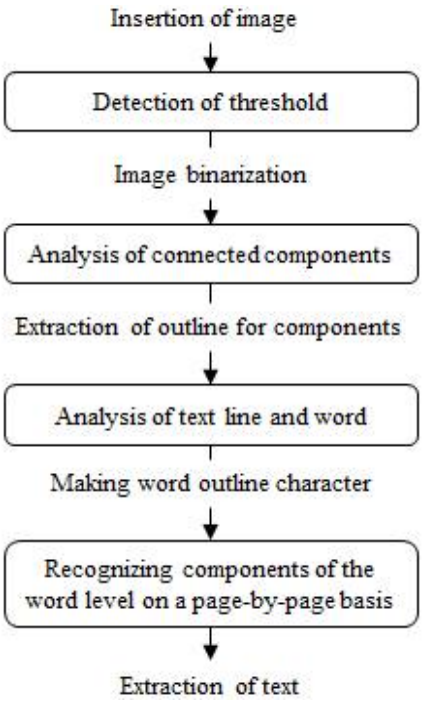


Fig. 1. Flow of Tesseract processing.

Tesseract 라이브러리는 바이너리 파일을 통해 바로 설치를 할 수 있으며, 바이너리를 통해 설치할 경우 환경변수를 자동으로 설정해주고 관련 라이브러리 및 언어팩을 자동으로 다운로드 및 설치를 수행한다. 설치가 완료되었다면 Tesseract 라이브러리는 빌드를 통해 생성하며, Visual Studio를 통해 (설치경로\tesseract-ocr\vs2008\tesseract.sin)의 파일을 열 수 있다. 그리고 솔루션 구성을 DLL_Debug, DLL_Release 중에서 필요에 따라 프로젝트 설정을 수행해 주면 된다.

3. 제안된 단어 정리 방법

본 논문에서 제안하고 있는 방법의 개요와 시스템의 구성에 대해 설명하고, 단어를 정리하고 사전데이터를 연결하기 위한 핵심 원리인 이분법과 부가적인 기능들을 설명한다. 카메라로 문서를 촬영하면 소프트웨어가 자동적으로 문서 내 단어를 인식하여 인식순서에 따라 단어를 표시시켜주고 사용자가 단어장에 추가 및 편집이 가능하게 해준다.

3.1 시스템 구조

Fig. 2는 제안하는 방법에서 사용자가 상호작용하는 시스템의 흐름을 나타낸다. 사용자는 문서의 단어를 정리하기 위해 카메라로 촬영하면 시스템에서 단어를 분석하고 단어 목록을 출력해준다. 사용자는 시스템이 출력한 단어리스트를 편집하여 단어장으로 데이터베이스에 저장할 수 있고, 단어장을 데이터베이스에서 조회할 수 있다. 사용자가 단어의 정보를 요청하면 시스템은 사전데이터에서 단어와 일치하는 정보를 조회하여 출력해준다.

3.2 주요 기능

제안한 방법의 주요 기능 구현과 작동원리에 대하여 설명한다.

3.3.1 단어정리

문자인식으로 저장된 텍스트 파일을 읽어 들여 공백과 특수문자를 없애면서 구분하고 나머지 구분한 문자를 저장하여 화면에 보이도록 리스트뷰에 출력하고, 단어 수정과 삭제 또는 사전에 등록되지 않은 단어를 모두 제거할 수 있고, 필터를 통해 원하는 문자열을 추려낼 수 있다. 그리고 리스트뷰에 출력한 단어들 중 사전에 등록된 단어의 비율을 확인할 수 있는데 이것은 사용자가 단어를 얼마만큼 수정해야 되는지, 또는 인식이 잘되었는지 확인할 수 있는 지표가 된다.

3.3.2 단어 사전 검색

오픈소스 사전인 StarDict [12]의 사전 파일을 활용, 이분법 [13]을 이용해 검색하고자 하는 단어의 위치를 빠르게 찾아내고 임의 접근하여 단어의 정보를 찾도록 수행한다. 위의 사전에 등록되지 않은 단어를 모두 제거하는 기능도 이분법을 사용해서 사전에서 데이터를 찾지 못하면 제거하도록 수행한다. Fig. 3은 이분법은 나타낸다. 이분법은 구하고자 하는 근이 존재하는 일정 구간을 반으로 이분한 후 두 개의 구간 중에서 근이 존재하지 않는 구간은 버리고 존재하는 구간은 또 반으로 이분하여 존재하지 않는 구간을 버리는 방법을 계속 n번을 시행하면 점점 구하고자 하는 근에 다가가게 되는 원리이다. 이분법을 사용하면 약 10만회 이상 읽어야 될 것을 20회 내외로 줄일 수 있다.

3.3.3 단어장 구축

단어 리스트를 편집하고 단어장으로 저장할 수 있게 한다. 안드로이드에서 제공하는 SQLite를 사용하고, 단어장을 저장할 때 테이블을 생성하여 저장한

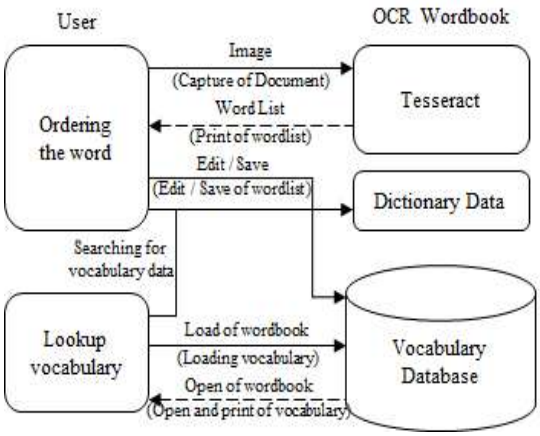


Fig. 2. Schematic diagram of the proposed system.

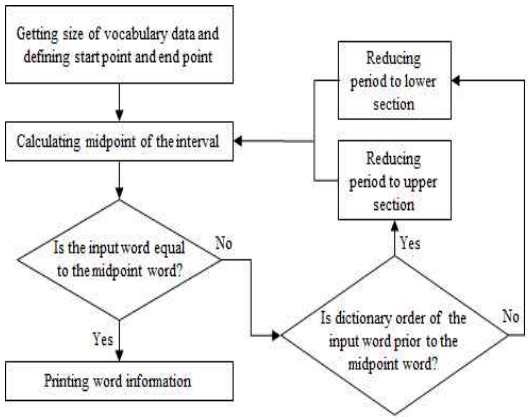


Fig. 3. Searching process of word in the dictionary.

Fig. 4. Start scene of proposed application.

다. SQLite는 MySQL나 PostgreSQL와 같은 데이터베이스 관리 시스템이지만, 서버가 아니라 응용 프로그램에 넣어 사용하는 비교적 가벼운 데이터베이스이다[14].

3.3.4 음성 합성

음성합성은 안드로이드 플랫폼을 위해 제공해주는 API를 사용한다.

4. 모의실험 및 결과

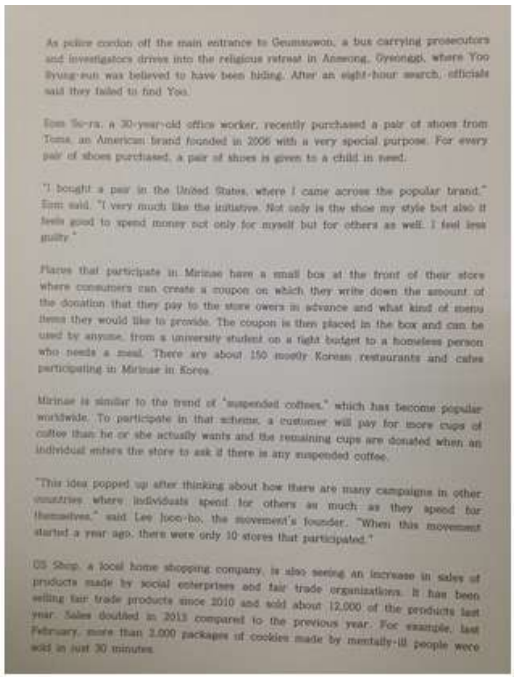
제안된 단어 정리 방법에 대해 안드로이드 스마트폰에서 구동되는 어플리케이션을 구현한다. 그림은 영어 단어 정리를 구현한 어플리케이션의 시작화면을 나타낸다. 처음 애플리케이션을 시작하면 시작 화면에서 카메라 촬영을 수행할지 데이터베이스에 저장된 단어장을 조회할지 사용자가 선택을 하게 된다. 카메라 촬영을 선택할 경우 영어 문서 촬영 후 단어 정리를 수행하게 되고, 단어장 미리 보기를 선택할 경우 단어장의 내용을 확인한 다음 단어장을 열어볼 수 있다.

단어 인식 버튼을 클릭하면, 카메라로 문서를 촬영한다. Fig. 5는 모의실험에 사용한 A4용지 영어 문서를 나타낸다.

스마트폰 액정의 넓이가 한정되어 있기 때문에 카메라로 한번에 촬영 가능한 면적은 제한적이다. 삼성 갤럭시S2 기종을 기준으로 한번에 Fig. 6의 테두리만큼 촬영이 가능하다. 인식에 성공한 단어는 Fig. 7에 표시하였다.

인식된 단어는 새로운 파일의 이름으로 단어장을 만들고, Fig. 8과 같이 화면을 구성하고 표시해준다.

단어장에서 사용자가 알고 있는 단어는 삭제할 수



As police cordon off the main entrance to Geumsuwon, a bus carrying prosecutors and investigators drives into the religious retreat in Anseong, Gyeonggi, where Yoo Byung-eun was believed to have been hiding. After an eight-hour search, officials said they failed to find Yoo.

Eom So-ra, a 30-year-old office worker, recently purchased a pair of shoes from Toms, an American brand founded in 2006 with a very special purpose. For every pair of shoes purchased, a pair of shoes is given to a child in need.

"I bought a pair in the United States, where I came across the popular brand," Eom said. "I very much like the initiative. Not only is the shoe my style but also it feels good to spend money not only for myself but for others as well. I feel less guilty."

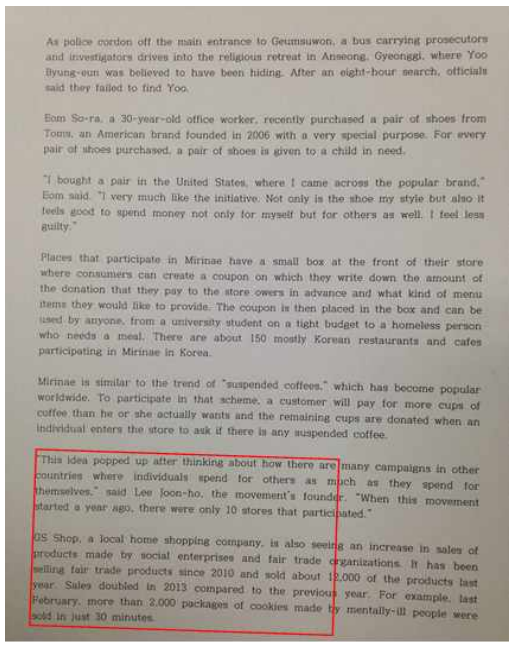
Places that participate in Mirinae have a small box at the front of their store where consumers can create a coupon on which they write down the amount of the donation that they pay to the store owers in advance and what kind of menu items they would like to provide. The coupon is then placed in the box and can be used by anyone, from a university student on a tight budget to a homeless person who needs a meal. There are about 150 mostly Korean restaurants and cafes participating in Mirinae in Korea.

Mirinae is similar to the trend of "suspended coffees," which has become popular worldwide. To participate in that scheme, a customer will pay for more cups of coffee than he or she actually wants and the remaining cups are donated when an individual enters the store to ask if there is any suspended coffee.

"This idea popped up after thinking about how there are many campaigns in other countries where individuals spend for others as much as they spend for themselves," said Lee Joon-ho, the movement's founder. "When this movement started a year ago, there were only 10 stores that participated."

GS Shop, a local home shopping company, is also seeing an increase in sales of products made by social enterprises and fair trade organizations. It has been selling fair trade products since 2010 and sold about 12,000 of the products last year. Sales doubled in 2013 compared to the previous year. For example, last February, more than 2,000 packages of cookies made by mentally-ill people were sold in just 30 minutes.

Fig. 5. Example of English document.



As police cordon off the main entrance to Geumsuwon, a bus carrying prosecutors and investigators drives into the religious retreat in Anseong, Gyeonggi, where Yoo Byung-eun was believed to have been hiding. After an eight-hour search, officials said they failed to find Yoo.

Eom So-ra, a 30-year-old office worker, recently purchased a pair of shoes from Toms, an American brand founded in 2006 with a very special purpose. For every pair of shoes purchased, a pair of shoes is given to a child in need.

"I bought a pair in the United States, where I came across the popular brand," Eom said. "I very much like the initiative. Not only is the shoe my style but also it feels good to spend money not only for myself but for others as well. I feel less guilty."

Places that participate in Mirinae have a small box at the front of their store where consumers can create a coupon on which they write down the amount of the donation that they pay to the store owers in advance and what kind of menu items they would like to provide. The coupon is then placed in the box and can be used by anyone, from a university student on a tight budget to a homeless person who needs a meal. There are about 150 mostly Korean restaurants and cafes participating in Mirinae in Korea.

Mirinae is similar to the trend of "suspended coffees," which has become popular worldwide. To participate in that scheme, a customer will pay for more cups of coffee than he or she actually wants and the remaining cups are donated when an individual enters the store to ask if there is any suspended coffee.

"This idea popped up after thinking about how there are many campaigns in other countries where individuals spend for others as much as they spend for themselves," said Lee Joon-ho, the movement's founder. "When this movement started a year ago, there were only 10 stores that participated."

GS Shop, a local home shopping company, is also seeing an increase in sales of products made by social enterprises and fair trade organizations. It has been selling fair trade products since 2010 and sold about 12,000 of the products last year. Sales doubled in 2013 compared to the previous year. For example, last February, more than 2,000 packages of cookies made by mentally-ill people were sold in just 30 minutes.

Fig. 6. Captured area for one scene.

있으며, 발음 버튼을 클릭하면 영어발음을 들을 수 있다. 또한, 현재 문서 인식의 환경이 좋지 않아 문서 인식율이 떨어졌다고 판단될 경우에는 사용자가

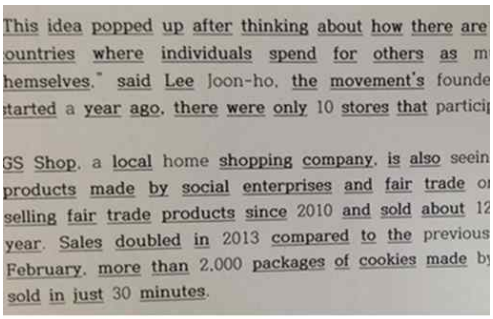


Fig. 7. Detected words after recognition.

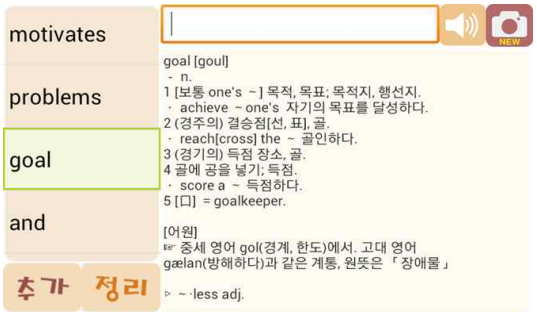
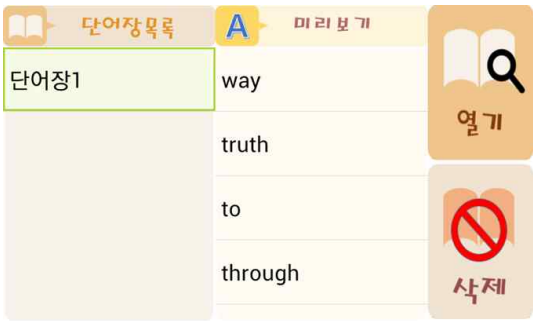


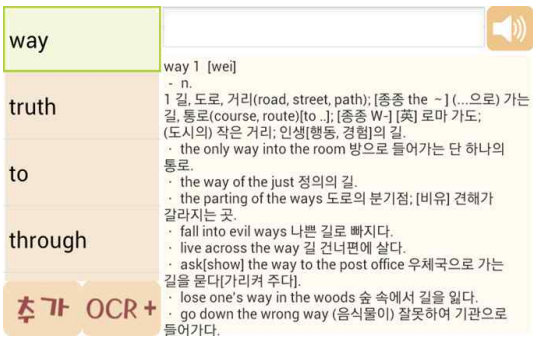
Fig. 8. Vocabulary display.

New버튼을 클릭하면 이전 촬영단계로 돌아갈 수 있다. 인식율은 조명, 촬영 거리, 문서의 기울기 등에 따라 영향을 받으므로 현재보다 인식율을 높이기 위해서는 조도를 높이든지, 촬영거리를 줄여주면 된다. 만약 인식되지 않는 단어는 추가버튼을 클릭하여 단어장에 추가할 수 있다. 단어장1 이름으로 단어장이 만들어진 경우에 Fig. 4에서 단어장 버튼을 클릭하면 Fig. 9(a)와 같이 단어장 목록을 볼 수 있다. 원하는 단어장을 선택하고 열기버튼을 클릭하면, Fig. 9(b)와 같이 단어장을 열어 정리되어 있는 단어를 볼 수 있다. 또한, 목록에 있는 단어장을 삭제할 경우에는 이름을 선택하고 삭제버튼을 클릭하면 된다.

다양한 실험환경에 대해 단어 인식율을 조사하였다. 사용자가 책을 읽을 수 있는 밝은 조명에서는 구름없는 햇빛에서와 같이 인식율이 같았으며, 촬영거리는 Fig. 6에서과 같이 A4용지의 3/4크기 이상을 촬영하면, 단어의 인식율이 떨어지지만, 그 이하를 촬영하면 인식율에는 영향이 없었다. 문자의 기울기에 대한 인식율은 Table 1에 나타낸다. 조명과 거리에는 인식율이 영향을 미치지 않는 조건에서 기울기만을 조정하였다. 문서에서 전체 단어수는 71개 였으



(a)



(b)

Fig. 9. Vocabulary utilities, (a) Preview, (b) Open.

Table 1. Recognition rate according to tilt

Tilt (degree)	Number of recognized words	Recognition ratio (%)
0	70	98.59
5	68	95.77
10	65	91.54
15	60	84.50
20	45	63.38
25	18	25.35
30	5	7.04

며, 기울기가 0인 경우에는 70개로 거의 모든 단어를 인식하였지만, 기울기가 커질수록 인식율이 떨어짐을 알 수 있다. 특히 30도 이상 기울어지면 단어의 인식율이 많이 떨어지므로 단어장을 만들기 위해서는 문서에서 단어와 카메라 기울기가 어긋나지 않도록 촬영하는 것이 필요하다. 이를 위해서는 문서 촬영이 시작될 경우에, 촬영되는 화면에 올바른 방향으로 적색의 직선을 표시해 주며, 인식되는 문서의 단어들의 방향과 직선의 방향이 같도록 카메라 방향을 조정하면 기울기가 어긋나지 않게 된다.

5. 결 론

본 논문에서는 언어 단어장을 정리하는 방법을 제안하고, 안드로이드 계열 스마트폰 어플리케이션으로 영어 단어장을 정리하는 시스템을 구현하였다. 실험결과 사용자가 촬영환경(조명, 기울기)을 적절하게 설정하여 사용하면 대부분의 영어 단어를 인식할 수 있었고, 소수의 인식하지 못한 단어는 편집하거나 추가할 수 있도록 구현하였다. 스마트폰으로 언제 어디서나 사용할 수 있는 인터페이스로 구성되어있기 때문에 사용자는 영어를 효과적으로 학습할 수 있을 것으로 기대되며, 현재 특허 출원[15]중으로서 영어 이외의 언어에도 확장 적용함으로써 다양한 언어로 상업화도 가능할 것으로 기대된다.

REFERENCE

[1] A. Singh, K. Bacchuwar, and A. Bhasin, "A Survey of OCR Applications," *International Journal of Machine Learning and Computing,* Vol. 2, No. 3, pp. 314-318, 2012.

[2] R. Mithe, S. Indalkar, and N. Divekar, "Optical Character Recognition," *International Journal of Recent Technology and Engineering*, Vol. 2, No. 1, pp. 72-75, 2013.

[3] V. Shrivastava and N. Sharma, "Artificial Neural Network Based Optical Character Recognition," *Signal & Image Processing : An International Journal*, Vol. 3, No. 5, pp. 73-80, 2012.

[4] S.-K. Kwon and S.-H. Jo, "Community-Based Travel Information System Using Augmented Reality," *Journal of Korea Multimedia Society,* Vol. 18, No. 2, pp. 97-105, 2015.

[5] Microsoft, https://www.microsoft.com (accessed Feb., 25, 2015).

[6] R. Smith. "An Overview of the Tesseract OCR Engine," *Proceeding of International Conference on Document Analysis and Recognition*, pp. 629-633, 2007.

[7] The Tesseract Open Source OCR Engine, https://github.com/tesseract-ocr (accessed Jan., 10, 2015).

[8] S.-K. Kwon, H.-J. An, and Y.-H. Choi, "Character Recognition System based on Android Smart Phone," *International Journal of Modern Engineering Research*, Vol. 2, No. 6, pp. 4091-4093, 2012.

[9] E.-B. Go, Y.-J. Ha, S.-R. Choi, K.-H. Lee, and Y.-H. Park, "An Implementation of an Android Mobile System for Extracting and Retrieving Texts from Images," *Journal of Digital Contents Society*, Vol. 12, No. 1, pp. 57-67, 2011.

[10] S.-R. Choi, E.-B. Go, Y.-J. Ha, and Y.-H. Park, "An Implementation of a System for Extracting and Retrieving Texts on Android Platform using OCR," *Proceeding of Korea Multimedia Society*, pp. 469-473, 2010.

[11] J.-W. Kim, S.-H. Kim, and C.-W. Lee, "Construction of Educational Contents Using On-Line Character Recognition," *Proceeding of Korea Multimedia Society*, pp. 309-312, 2009.

[12] Stardict, http://www.stardict.org/ (accessed Oct., 12, 2014).

[13] R. Stevenson, "The Completion of Locally Refined Simplicial Partitions Created by Bisection," *Mathematics of Computation*, Vol. 77, No. 261, pp. 227-241, 2008.

[14] SQLite, http://www.sqlite.org/ (accessed Sep., 3, 2014).

[15] S. Kwon, S. Kim, and D. Kim, *System and Method for Ordering Word Based on OCR Character Recognition*, 10-2015-0019575, Korea, 2015.

김 남 규

1995년 2월 한국과학기술원 졸업
1998년 2월 포항공과대학교 전자계산학 석사
2005년 2월 포항공과대학교 컴퓨터공학과 박사
2000년 3월~2001년 2월 ATR연구소(일본) 연구원
2005년 3월~2006년 2월 포항공과대학교 정보통신연구소 연구원
2006년 3월~2009년 8월 KT중앙연구소 선임연구원
2009년 9월~현재 동의대학교 게임공학과 조교수
관심분야: 인간과컴퓨터상호작용, 컴퓨터비전, 가상현실

김 성 우

1991년 2월 KAIST 전기및전자공학과 졸업
1993년 2월 KAIST 전기및전자공학과 석사
1999년 2월 KAIST 전기및전자공학과 박사
1999년 3월~2002년 2월 한국전자통신연구원 선임연구원
2008년 9월~2009년 8월 California State Univ. 방문 교수
2002년 3월~현재 동의대학교 컴퓨터소프트웨어공학과 교수
관심분야: 임베디드소프트웨어, 사물인터넷

김 동 언

2009년 3월~현재 동의대학교 컴퓨터소프트웨어공학과 학부과정
관심분야: 사물인터넷(Internet of Things), 영상인식

권 순 각

1990년 2월 경북대학교 전자공학과 졸업
1992년 2월 KAIST 전기및전자공학과 석사
1998년 2월 KAIST 전기및전자공학과 박사
1997년 3월~1998년 8월 한국전자통신연구원 연구원
1998년 9월~2001년 2월 기술보증기금 기술평가센터 팀장
2003년 9월~2004년 8월 Univ. of Texas at Arlington 방문 교수
2010년 9월~2011년 8월 Massey University 방문 교수
2001년 3월~현재 동의대학교 컴퓨터소프트웨어공학과 교수
관심분야: 멀티미디어신호처리, 영상통신