카메라형 광학식문자판독기술(OCR)을 활용한 오프라인 중고서점의 장서 디지털 데이터화 관리 방안 제안

구자민°, 함승모*, 김우제*, 심현동**, 류기동***
°*서울과학기술대학교 소프트웨어분석설계학과
**대보정보통신
***ECS 텔레콤
e-mail: coojamin@seoultech.ac.kr°,{sm1130,wjkim}@seoultech.ac.kr*, shimhd@dbcs.co.kr**, dryu@ecstel.co.kr***

An Efficient Management Strategy of A Offline Second-Hand Bookstore With Camera Type OCR Technology

Koo Ja Min°, Ham Seung Mo*, Kim Woo Je*, Shim Hyun Dong**, Ryu Ki Dong***
°*Dept. of Software Analysis and Design, Seoul National University Of Science & Technology
**DAEBO Communication & Systems
***ECS Telecom

● 요 약 ●

본 논문에서는 카메라형 OCR (Optical Character Reader) 기술을 이용해 오프라인 중고서점의 효율적 장서관리 시스템을 구축하기 위한 디지털 데이터화 관리시스템 방안을 제안한다. OCR은 광학적으로 인식할 수 있는 문자를 컴퓨터가 읽을 수 있도록 하는 기술이다. 원리적으로 문자 한 개를 수십 개의 모눈으로 분할해 특정한 모눈의 흑백 또는 자획형상 특징에 의해 문자를 판독한다. 이 논문에서는 OCR 기술을 활용함으로써 디지털 데이터화의 효과는 물론 적용 환경의 개선효과를 기대해 볼 수 있는 오프라인 중고서점 시장을 목표로 했다. 오프라인 중고서점에서 보유하고 있는 장서의 디지털 데이터화는 기업형 중고서점과의 경쟁에 있어서도 생존을 위해 필요한 요소이다. 카메라형 OCR 기술을 활용한 장서 디지털 데이터화는 오프라인 중고서점 판매자가 도서재고 검색 및 판매 관리 효율을 높이도록 도와줄 뿐 아니라, 도서판매 유형, 소비자 분석과 수요 예측을 가능하게 한다. 또한 소비자에게 오프라인 중고서점에서 보유하고 있는 희귀 장서와 중고서적들을 검색해 구입할 수 있는 편의를 제공할 것이다. 오프라인 중고서점 판매를 촉진하고 활성화시킨다면 출판의 선순환적 구조를 만드는 데 기여할 것으로 예상된다.

키워드: 광학식문자판독기술(Optical Character Recognition), 오프라인 중고서점(Offline Second-Hand Bookstore), 장서관리(Collection Management Policy), 카메라형 OCR(Camera Type OCR)

I. 서 론

이 논문은 오프라인 중고서점의 효율적인 장서관리를 위해 카메라형 OCR을 활용하는 방안을 제시한다. 오프라인 중고서점에 카메라형 OCR 도입이 필요한 이유는 다음과 같다.

오프라인 중고서점의 판매자 스마트폰, 태블릿 컴퓨터와 같은 디지털 디바이스에 카메라형 OCR을 간편히 설치해 활용한다면 장서를 디지털 데이터화 함으로써 관리 환경을 개선하고 효율을 높일 수 있을 것이다. 이는 오프라인 중고서점에 컴퓨터, 포스시스템 외에 스캐너와 같은 전산 비품을 구축하는 비용을 절감하고, 소비자들에게 가격적 요인을 통한 오프라인 중고서점 이용 동인(Key driver)을 제공해 줄 것이다. 이로써 고객들의 중고서적 구매 니즈가 증가한 출판시장환경에서 오프라인 중고서점의 매출 증대와 활성화를 기대할 수 있다.

이 논문은 다음과 같이 진행된다. 오프라인 중고서점의 카메라형 OCR 도입 필요성을 제시하고, 2장에서 OCR 기술에 대한 동향 및 관련사례를 분석한다. 3장에서 오프라인 중고서점에 카메라형 OCR 기술 적용시 발생 가능한 편익과 문제점을 파악하고 이를 보완해낼 기술 방법을 제시한다. 4장에서 결론과 향후 과제를 다룬다.

2. 관련 연구

2.1 OCR 기술 동향 및 사례

2.1.1 국내 동향

국내의 OCR 기술에 대한 연구는 1970년대부터 상업적 용도로 널리 사용되었으며 최근에는 여권 처리, 보안 문서 처리(수표, 재무 문서, 청구서), 우편물 추적, 출판, 소비재 포장(뱃치 코드, 로트 코드, 만료일), 임상적 용도 등과 같이 자동화된 작업을 위해 사용되고 있다. 여기에는 OCR 판독기와 소프트웨어를 사용할 수 있으며 바코드 판독 및 제품 검사와 같은 추가 기능이 있는 스마트 카메라와 비전 시스템도 사용할 수 있다.[4]

현재 사용 중인 문자 인식 방법은 문자패턴의 표현 방법과 분류 방법에 따라 원형 정합(Template matching) 방법, 통계적 방법, 분석적 방법 등으로 나누어진다.

원형 정합 방법은 문자의 패턴을 배열 형태로 분류하여 원형 패턴과 비교하여 가장 유사한 형태를 찾아내는 방법이다. 이 방법은 초기에 많이 사용하였으나 주로 하나의 고정된 형식의 문자에 대해서만 사용가능하다는 문제점으로 인해 현재는 사용 빈도가 낮다.

통계적 문자 인식 방법은 인식대상에서 특징 벡터를 추출하여 문자인식을 하는 것이다. 구조 분석적 문자 인식 방법은 문자의 구성 원리에 입각하여 자획 등과 같은 문자를 구성하는 기본 요소와 그들의 연관성을 추출하여 문자를 인식하는 것이다. 이 방법은 이론적인 정립이 잘 되어 있고 방법이 단순한 장점을 가지고 있으나 특징 문자에 대한 규칙이 활자체에 따라 매우 다양해지므로 인식 시간이 오래 걸린다는 단점이 있다.[6]

그림 1. 라벨상의 OCR
Fig 1. OCR on the label

2.1.2 해외 동향

OCR 기술은 도서관에서도 아주 중요한 기술로서, 현재 소장한 방대한 인쇄 자료를 디지털화 할 때 필요한 핵심 요소이다. 구글은 이미 도서검색프로젝트를 진행하면서 광학문자인식 기술을 활용해서 스캔한 이미지에서 책 본문 검색을 할 수 있도록 지원하고 있다. 이는 결국 도서관에서도 역사적 가치가 높아서 디지털화 해야 하는 자료나 공공 재산에 속하는 자료의 경우에는, 광학문자기술을 적용해서 이용자들의 검색 편의성을 도와야 한다는 것을 간접적으로 보여주고 있다.

네덜란드국립도서관은 신문 자료를 디지털화 하는 사업을 수행하면서 이 광학문자기술 연구에 더욱 박차를 가하고 있다. 예를 들어 자료가 17-18세기의 신문이라서 훼손이 있거나 글자 폰트가 현저히 다르다면 문자 인식의 정확성은 더 떨어질 수 있기 때문에 네덜란드 국립도서관은 민간 기업 및 대학과 협력하여 그 인식 정확도를 높이기 위한 기술 발전에 더욱 힘을 쏟고 있다.[1]

3. 장서관리 적용 방식

3.1 OCR 기술 적용시 시장 편익

기존까지 다량의 장서를 보유한 환경에서 판매관리를 위해 상용된 기술은 바코드[5]이며 최근에는 QR코드[6]가 바코드만큼 널리 사용되고 있다. 바코드를 QR코드와 비교할 경우 비용측면에서 바코드는 QR코드보다 1/10가량 저렴한 편이지만 정보 집적도가 낮고 훼손에 취약하다. QR코드와 스마트태그[7] 방식은 인식률이 좋고 정보를 많이 담을 수 있지만, 정보의 디지털 데이터화보다는 암호화를 통한 보전 측면에 적합하다.

한편 이 두 가지 방식의 단점을 보완하면서 서적에 인식 코드를 부착하지 않고도 디지털 데이터화를 통해 사용자가 바로 활용할 수 있는 가독성 있는 활자 데이터를 제공하는 데 적합한 기술로 OCR이 주목된다.

사무환경의 기술 개선을 통해 업무 효율성을 높이는 것은 중요한 일이다. 그러므로 서류와 장서의 정보를 실시간으로 인식해 이미지, 활자 데이터로 관리하는 OCR 기술 중에서도 스캐너와 카메라형의 장단점을 이해하고 적합한 방식을 차용한다면 중고서적의 방대한 장서 물량을 관리하는 데 필요한 시간과 비용을 절감하는 편익을 제공할 것이다.

3.2 디지털 데이터화 적용 방식

OCR 기술을 활용하여 오프라인 중고서점에서 보유한 장서 DB를 구축하기 위한 시스템 구성도는 그림2와 같다.

---

5) 영숫자나 특수글자를 기계가 읽을 수 있는 형태로 표현하기 위해 굵기가 다른 수직 막대들의 조합으로 나타내, 광학적으로 판독이 가능하도록 한 코드. 상품포장에 인쇄되어 가격표시를 하거나 책표지에서 도서 관리를 위한 정보를 나타내는 등 물품을 구분하기 위한 다양한 용도로 사용되는 인식 코드. 예를 들어, 바코드를 카운터에 있는 광학 스캐너로 문지르면 포스(POS) 컴퓨터가 상품 번호를 가격리스트 데이터베이스와 대조하여 정확한 양을 금전등록기에 기록하는 방식으로 작업이 이루어진다. 코드화하는 방법으로는 세계 상품 코드(Universal Product Code; UPC) 체계가 오늘날 가장 널리 사용되고 있다. (한글글꼴용어사전: 세종대왕기념사업회)

6) QR코드 [Quick Response Code] 바코드보다 훨씬 많은 정보를 담을 수 있는 격자무늬의 2차원 코드이다. 스마트폰으로 QR코드를 스캔하면 각종 정보를 제공받을 수 있다. (두산백과)

7) RFID [radio frequency identification]. 극소형 칩에 상품정보를 저장하고 안테나를 달아 무선으로 데이터를 송신하는 장치. (시사상식사전, 박문각)

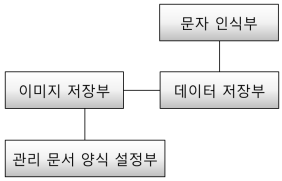

그림2. 장서 디지털 데이터화 시스템 구성도
Fig 2. Digitalizing System Configuration

관리 문서양식 설정부는 중고서적 등록 및 도서관리 를 하기 위한 양식 서류들에 대한 인식 영역을 설정해 둔 것이다. 이미지 저장부는 디지털 이미지의 저장 영역이다. 중고서적의 사진, 제목, 출판사 정보 등의 데이터 저장부이다. 문자, 이미지 인식이 이뤄진다.

3.2.1 OCR 기본 문자 인식

다음 절에서 카메라를 이용한 OCR 과정을 설명하기 전에 기본적으로 문자를 인식하는 과정을 그림3을 통해 알아보기로 한다.

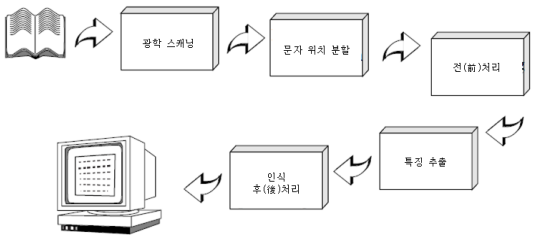

그림3. OCR시스템의 문자 인식 과정[8)][5]
Fig 3. Character Recognition Process

우선 인식하고 싶은 부분을 광학 스캐너를 이용하여 아날로그 형식의 문서를 디지털화 한다. 텍스트가 포함된 부분이 분할 과정에 의해 추출되며 노이즈 제거 등의 전처리 과정을 거친다. 각 문자의 고유 특징들은 전 단계에서 추출된 특징들과 비교되어 얻어지고 맥락적인 정보는 원본의 문자와 숫자들을 기반으로 재구성된다.

3.2.2 카메라형 OCR

이 논문에서 제시하는 방법은 스캐너 장치를 이용한 방법이 아닌, 카메라를 이용하여 얻어진 이미지에서 문자를 인식하는 방법이다. 우선적으로 개인이 보유한 스마트폰, 태블릿 컴퓨터와 같은 디지털 디바이스에 촬영 가능한 카메라 렌즈가 장착되어 있어야 한다. 다음으로 장서관리를 위한 OCR 소프트웨어 또는 애플리케이션을 설치해야 한다. 프로그램의 정상구동 중 촬영된 정보의 디지털 데이터화가 이뤄진다. 디지털 디바이스의 카메라 렌즈를 통해 수집된 이미지 정보가 처리되어 그림4와 같은 전개로 장서정보를 저장하게 된다.

8) 문자인식(Character Recognition)이란 시각정보를 통하여 문자를 인식하고 의미를 이해(Understanding)하는 인간의 능력을 컴퓨터로 실현하려는 패턴인식(Pattern Recognition)의 한 분야이다.[4]

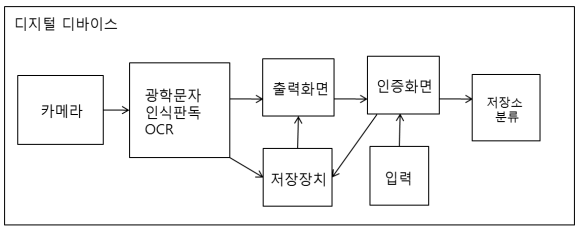

그림4. 디지털 디바이스 탑재된 OCR 활용 장서 저장법
Fig 4. Collection Storage Process with Digital Device

카메라를 통해 장서를 촬영하면 앞 절에서 언급한 OCR 시스템의 기본 문자 인식 과정에 의해 문자를 인식한다. 인식한 화면이 디지털 디바이스에 출력되면 올바르게 인식이 되었는지, 저장할 것인지를 결정하여 저장소에 분류한다.

3.3 카메라형 OCR 적용 이점 및 한계점 보완

3.3.1 시간효율성 및 사용자 편의성

스캐너와 카메라형 OCR 기술 중에서도 오프라인 중고서적의 장서관리를 하는데 무엇이 적합한지 판단해 활용하려면 각각의 장단점을 이해할 필요가 있다.

스캐너 장치를 이용하는 방법의 경우, 데이터화가 필요한 장서 부위를 스캐너에 안착해 인식시키므로 카메라로 촬영하는 방식보다 시간이 소요될 수밖에 없다. 또한, 인식 결과가 만족스럽지 못할 때 인식과정을 반복하여 정보를 취득하므로 소요시간이 증가한다. 카메라형 OCR의 경우, 인식결과를 확인하고 오류를 파악한 다음, 재촬영을 통해 작업을 수정하므로 스캐너 장치 활용에 비해 저장 방법이 수월하다.

스캐너 장치의 경우 사용자들이 이용할 수 있는 물리적인 크기가 점차 줄어들어 최근에는 스캔 기능을 탑재한 마우스까지 등장한 상태다. 스캐너 장치가 공간을 차지하는 제약사항은 줄어든 상황이다. 하지만 스캐너 장치의 소형화는 사용자가 원하는 정보의 디지털 데이터화를 위해 장치를 다룰 때, 사용자의 움직임을 간소화 시켜주지 못한다. 왜냐하면 소형의 스캐너 장치는 카메라 렌즈로 촬영하는 방식과 비교했을 때 사용자가 여러 번에 걸쳐 기기를 움직여 정보를 인식해야하기 때문이다.

오프라인 중고서점에서 보유한 다량의 장서가 책꽂이에 꽂혀있거나 한데 모여 있는 상태에서 스캐너를 통해 도서정보를 취합하기는 어렵다. 따라서 카메라형 OCR로 촬영해서 10권이상의 장서 제목을 인식하는 방식이 적용된다면 오프라인 중고서점의 관리자는 훨씬 편리하게 장서 정보를 취득하고 저장할 수 있을 것이다.

카메라형 OCR은 사용자가 보유한 디지털 디바이스(스마트폰, 태블릿 컴퓨터 등)에 소프트웨어를 설치하거나 애플리케이션 형태로 다운받아서 사용할 수 있는 것이 큰 장점이다. 이는 스캐너와 같은 입력장치 설치비용을 절약하고 기존의 컴퓨터 장치에 데이터 정보를 동기화시키는 방식으로 이용할 수 있으므로 경제적인 효율성을 높일 것으로 기대된다.

3.3.2 한계점 및 보완점

카메라형 OCR에도 한계는 있다. OCR 기술이 지속적으로 발전

하면서 끊임없이 지목되고 있는 것은 판독특성에 의한 문자 및 이미지 인식률, 즉 정확도 측면이다.

하지만 이 문제 역시, 촬영거리와 데이터 처리속도의 한계를 보완해낸 상태다. 본 논문의 기술용도와 유사한 상용제품 중 카메라형 OCR 엔진을 탑재한 도로의 차량용 번호판 판독 카메라가 그것이다. 이는 이미 전세계 모든 차량의 번호판을 날씨 조건에 지장 없이 고속 주행하는 차량의 이미지까지 높은 정확도로 처리할 수 있는 기술로써 보급된 상태이다. [2]

그러므로 오프라인 중고서점에서 장서관리를 위해 정적으로 비치되어 있는 상태의 도서를 촬영할 때 반드시 필요한 판단기준은 카메라형 OCR이 다양한 디자인 서체를 인식해야하는 정확성이 될 것이다.

이러한 문제 역시 다른 기술영역의 발전과 더불어 해결 대안이 나오고 있다. 웹 서비스 신청 단계에서 신청자가 실제 인간 사용자임을 확인하기 위해 사용되는 텍스트 기반 캡차(text-based CAPTCHA) 부분에서 변형된 문자를 OCR로 인식하는 이미지-텍스트 융합 캡차가 기존 이미지 기반 캡차보다 사용자에게 편리하고 신뢰성이 증진될 수 있음을 입증되었기 때문이다.[3] 따라서 이미지 인식률과 다양성은 현재의 기술로 충분히 보완할 수 있는 상태다.

4. 결 론

이 논문은 현존하는 OCR 기술 수준이 카메라로 촬영된 이미지에서 활자를 식별해 디지털 데이터화 할 수 있는 단계임을 이해하고, 이를 오프라인 중고서점의 장서 디지털 데이터화에 활용함으로써 매장에 필요한 실질적인 도서 재고관리 및 운영효율을 높일 수 있도록 지원하기 위한 목적에서 진행하였다.

촬영용 렌즈가 있는 디지털 디바이스에 카메라형 OCR 기술 소프트웨어를 설치해 활용함으로써 오프라인 중고서점의 관리자가 장서의 표지이미지를 촬영하였을 때 식별가능한 도서 정보를 디지털 데이터화 함으로써 활용가능하게 하는 것이다.

향후 OCR 기술을 탑재할 수 있는 디지털 디바이스의 유형이 다양해지면서 기술의 추후 연구가 가능할 것으로 보인다. 이러한 기술 융합은 도서정보를 저장 관리해주는 서비스를 발전시켜 소비자에게 제공함으로써 희소가치가 있는 장서의 발굴과 지식활용에 기여할 것으로 예상한다.

또한, 오프라인 중고서점이 OCR을 활용해 도서재고 검색과 판매 효율을 증진시키고, 수요 분석 지표로써 활용한다면 사업 활성화에 도움을 줄 수 있다.

오프라인 중고서점이 활성화되면 오늘날 출판시장이 기업형중고서점에 편향되어 있는 것을 견제하고 출판 생태계의 건전한 선순환 구조 기능을 점진적으로 회복하는데도 긍정적인 효과를 줄 것으로 기대한다.

나아가 카메라형 OCR 기술의 가치를 알고, 오프라인 중고서점 사업에 적용하는 방안을 모색해본 이 논문의 시도를 참고로 하여 보다 심도 있는 관련 연구들이 이뤄진다면, 새로운 사업기회와 기술력 제고에도 기여할 것으로 기대된다.

참고문헌

[1] Library Term, *Library Lab Webzine* Vol. 68, The National Library of KOREA, December 2010.

[2] Microscan Systems, Inc., *Optical Character Recognition System*, http://www.microscan.com /ko/Technology/

[3] Moon Kwang Ho and Kim Yoo Sung. *"Reliable Image-Text Fusion CAPTCHA to Improve User-Friendliness and Efficiency,"*CAPTCHA 27 DOI: 10.3745/KIPSTC.2010.17C.1.027.

[4] Myung-Jae Lim, Sung-Kyung Hyun, Ji-Eun Park, and Ki-Young Lee, *"Image Processing for Mobile Information Retrieval Service"*, Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 11, No.1, Feb 2011.

[5] Optical Character Recognition, Line Eikvil, December 1993.

[6] Won-Gyo Jung, Sang-Sung Park, Young-Geun Shin, Dong-Kyu Ahn, and Dong-Sik Jang, *"Optical Character Recognition System Using The Document Form Identification"*, Journal of the KOREA Society of Computer Information. Vol. 16. No. 1. pp.155-161. June 2008.