이미지데이터 활용을 위한 지능형 인식 라이브러리 연구 개발

곽희규*, 김성헌*, 이정우*, 유지훈*, 이현주*
*㈜인지소프트 정보기술연구소
e-mail : hkkwag@inzisoft.com

R&D of Intelligent Document Recognition Library for utilizing image data

Hee Kue Kwag*, Sung Hun Kim*, Jung Woo Lee*, Ji Hun Yoo*, Hyun Joo Lee*
*InziSoft Co., Ltd. Information & Technology Lab.

요 약

본 연구는 공공기관이 소장한 이미지데이터 활용성을 높이기 위한 전문검색서비스 구현 시 필수적인 문서인식시스템의 고도화에 있으며, 주요한 연구방향은 공공기관이 소장하고 있는 데이터의 분석을 통해 이미지분석 기술 및 라이브러리를 개발하고 특화된 지식베이스를 구성하는 것이다. 또한, 향후 확장성을 고려하여 지식베이스를 지속적으로 관리할 수 있는 툴을 개발하는 것이다. 본 연구는 현재 지능형 인식 라이브러리를 결합한 프로토타입(prototype) 시스템 개발이 완료된 바, 방대한 국가기록원내 소장자료를 대상으로 다양한 성능평가를 위한 테스트베드 구축이 진행되고 있다.

1. 서론

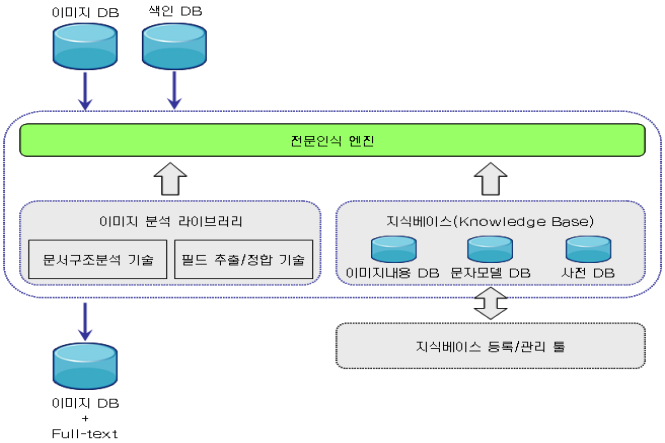
역사, 문화, 사회적 자료를 소장하고 있는 도서관 및 공공기관에서는 국가지식의 대국민서비스 측면에서 기존 종이문서로 보관되어 온 자료의 이미지화 작업을 활발히 진행하였고, 인터넷을 통한 자료의 접근성을 크게 증가시켰다. 그러나 정보검색은 이미지데이터에 대한 수작업 색인정보에 한정되어 있어 정보검색의 정확성 및 효율성이 떨어질 뿐만 아니라 전문(full text)에 대한 검색은 많은 시간비용이 소요되고 있는 실정이다. 따라서 소장하고 있는 이미지데이터에 대한 정보검색의 속도와 정확성을 향상시키고 활용을 극대화하기 위해서는 문서인식을 통한 전문검색(Full Text Retrieval)서비스 제공이 요구되고 있다.

전문검색서비스 실현은 이미지데이터에 대한 전문인식 기술의 결합이 필요하고[1-7], 인식기술의 성능이 전체적인 서비스의 품질을 좌우한다. 과거 전자도서관 구축에 상용 문자인식 기술을 접목한 사례가 있지만, 성능의 불안정성에 의해 실효성을 거두지는 못했다. 국내 상용 문자인식 기술은 문서의 구조적인 특징, 문서이미지의 품질 변화에 따라 인식성능이 크게 변하는 불안정성으로 인해 문자인식 기술을 전문검색에 활용하는데 큰 장애요인으로 작용하고 있다.

본 연구는 국가기록원에서 소장하고 있는 이미지데이터의 활용을 극대화하기 위해, 소장자료의 분석으로부터 문서인식에 필요한 인식라이브러리 및 지식베이스(knowledge base)를 구축함으로써 전문검색서비스 구축을 위한 기술 기반을 마련하는데 있다.

2. 연구내용

본 연구의 목표는 국가기록원이 소장한 이미지데이터의 활용성을 높이기 위한 전문검색서비스 구현 시 필수적인 문서인식시스템의 고도화에 있으며, 주요한 연구방향은 국가기록원이 소장하고 있는 데이터의 분석을 통해 이미지분석 기술 및 라이브러리를 개발하고 특화된 지식베이스를 구성하는 것이다. 또한, 향후 확장성을 고려하여 지식베이스를 지속적으로 관리할 수 있는 툴을 개발하는 것이다.



(그림 1) 특화된 문서인식시스템 모델

국가기록원 이미지데이터 활용을 위한 특화된 문서인식시스템 모델 연구 개발은 먼저 소장 자료 분석을 통한 문서 유형 분류와 유형별 문서 구조를 파악하는 작업을 선행해야 한다. 다양한 문서구조와 블록 정보를 표현할 수 있는 데이터 구조를 설계하고, 문서별로 정보를 등록 및 관리할 수 있는 도구를 설계하고

개발한다. 또한, 기존 문서 구조 분석 방법 중 소장 자료의 문서구조 분석에 가장 적합한 방법론을 선택하여 라이브러리를 구성하고, 추출된 문서구조를 등록된 DB 와 비교하여 보다 정확한 정보를 추출하는 방법 등도 고려한다. 각 블록의 문자열 추출 및 인식 결과를 향상시키기 위해서는 블록의 폰트 정보와 검색에 사용되는 용어사전을 활용하는 것이 바람직하다. 블록별로 선택적으로 사용할 수 있도록 한글, 한/영, 한/한 문자인식기를 구성하고, 용어 사전은 특화된 분야에 따라 나누어서 구성한 후 인식 과정에서 선택된 사전을 참조하여 인식 결과를 최적화할 수 있도록 구성한다. 본 연구의 주요 내용은 다음과 같다.

◇소장 자료 분석 및 지식베이스 연구
◇이미지내용 DB 등록 모듈 설계 및 개발
◇문서 영상 전처리 라이브러리 설계 및 개발
◇문서구조분석 라이브러리 설계 및 개발
◇이미지내용 DB 정합 라이브러리 설계 및 개발
◇문자 모델 설계 및 개발
◇문자 모델 DB 관리도구 개발
◇사전(dictionary) 설계 및 개발
◇사전 DB 관리도구 개발

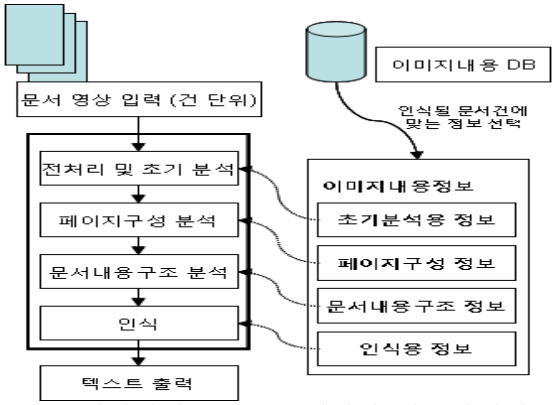
3. 연구 방법

1) 이미지내용 DB 연구 개발

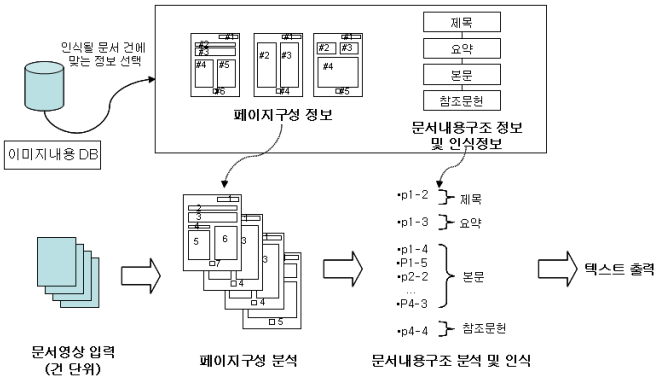
이미지내용 DB 는 문서구조분석 결과를 향상시키기 위한 지식베이스로서, 초기 문서 분석용 정보, 페이지 구성 정보, 문서내용 구조 정보, 인식용 정보로 구성하였다.

2) 이미지내용 DB 정합 알고리즘 개발

이미지내용 DB 정합 알고리즘은 문서구조분석 결과의 검증 역할을 수행하는 것으로, 전체적인 문서인식처리 과정은 전처리 및 초기 분석 단계, 페이지 구성 분석 단계, 문서내용구조 분석 단계, 인식 단계로 수행하였다. 문서영상의 문서구조 분석 결과를 등록된 이미지내용 DB 와 매칭하고, 정확한 분석 결과를 도출하기 위해 블록구조 정보(추출된 블록들의 위치 정보와 배치 관계를 매칭), 키워드 정보(인식 결과를 이용하여 등록된 키워드를 매칭), 직선 및 블록 구분자 정보 등과 비교하여 구조분석 결과를 보정하였다.



(그림 5) 이미지내용 DB 를 결합한 블록다이어그램



(그림 6) 이미지내용 DB 결합 정합 예시

4. 결론

본 연구는 현재 지능형 인식 라이브러리를 결합한 프로토타입(prototype) 시스템 개발이 완료된 바, 방대한 국가기록원내 소장자료를 대상으로 다양한 성능평가를 위한 테스트베드 구축이 진행되고 있다. 이번 연구개발의 결과를 통해 소장자료의 자동 인식이 안정적으로 이루어지면 전문검색서비스 구축을 위한 기반이 마련되고 정확하고 많은 색인들이 자동으로 추출되면서 검색 시 효율성과 정확성을 동시에 만족시킬 수 있을 것으로 기대하고 있다.

외국의 유수한 도서관 소장자료 디지털화를 위한 연구개발에 비해 국내의 경우 국가적인 측면에서 문서인식기술의 연구 개발에 대한 투자는 전무했었다. 이번 국가기록원의 연구개발은 소장자료 디지털화 선진사례에 부합하는 것으로, 향후 국내에서 문자인식 및 문서인식 기술의 고도화를 촉진할 수 있는 계기가 될 수 있을 것으로 전망하고 있다.

참고문헌

[1] Digital Heritage Publishing Ltd., "The electronic version of Siku Quanshu", http://www.skqs.com
[2] T. Keaton et. al., "Keyword spotting for cursive document retrieval", Proc. on DIA'97
[3] E.A. Galloway, "The Heinz electronic library interactive online system", The Public Access Computer Systems Review, 1998.
[4] M. Droettboom, "Using the Gamera framework for the recognition of cultural heritage materials", JCDL'02.
[5] S. Hara, "OCR for CJK classical texts", PNC'00.
[6] M. Kojima, "Automatic recognition of Tibetan Buddhist text by computer", PNC'00.
[7] T. Shih, "Transformation of palace archives of Ming and Ching Dynasties onto CD-ROM and Internet", PNC'00.

<<사사(acknowledgement)>>
본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사 드린다.