DOI: 10.3745/KIPSTB.2010.17B.2.125

이미지데이터 활용을 위한 문서인식시스템 연구 및 개발

곽 희 규†

요 약

본 연구는 공공기관이 소장한 이미지데이터의 검색 및 열람 등의 활용성을 높이기 위한 전문검색서비스 구현 시 필수적인 문서인식시스템의 고도화를 목표로 한다. 주요한 연구방향은 공공기관이 소장하고 있는 데이터를 사전에 분석하여 문서이미지 전처리 및 문서구조분석 기술을 개발하고, 문서인식 과정에서 활용하기 위한 이미지내용DB, 문자모델DB, 용어DB로 구성되는 특화된 지식베이스를 구축하는 것이다. 또한, 지식베이스 관리도구를 개발하여 향후 다양한 형태의 문서이미지로의 확장을 가능하게 한다. 최근 본 연구는 국가기록원에서 소장하고 있는 이미지데이터에 적합한 문서구조분석 라이브러리와 특화된 지식베이스를 결합한 문서인식 프로토타입 시스템 개발을 완료했다. 향후 본 연구의 결과는 방대한 소장자료의 검색 및 활용을 극대화할 전문검색시스템 연계를 위한 성능평가 및 테스트베드 구축에 활용될 것이다.

키워드 : 전문검색, 문서인식, 문서구조분석, 지식베이스, 이미지내용, 문자모델, 용어사전

Research and Development of Document Recognition System for Utilizing Image Data

Kwag, HeeKue†

ABSTRACT

The purpose of this research is to enhance document recognition system which is essential for developing full-text retrieval system of the document image data stored in the digital library of a public institution. To achieve this purpose, the main tasks of this research are: 1) analyzing the document image data and then developing its image preprocessing technology and document structure analysis one, 2) building its specialized knowledge base consisting of document layout and property, character model and word dictionary, respectively. In addition, developing the management tool of this knowledge base, the document recognition system is able to handle the various types of the document image data. Currently, we developed the prototype system of document recognition which is combined with the specialized knowledge base and the library of document structure analysis, respectively, adapted for the document image data housed in National Archives of Korea. With the results of this research, we plan to build up the test-bed and estimate the performance of document recognition system to maximize the utilization of full-text retrieval system.

Keywords : Full Text Retrieval, Document Recognition, Document Structure Analysis, Knowledge Base, Image Layout & Property, Character Model, Word Dictionary

1. 서 론

역사, 문화, 사회적 자료를 소장하고 있는 도서관 및 공공기관에서는 국가지식의 대국민서비스 측면에서 기존 종이문서로 보관되어 온 자료의 이미지화 작업을 활발히 진행하였고, 인터넷을 통한 자료의 검색 및 열람 등의 접근성을 크게 증가시켰다. 그러나 고문서 이미지데이터에 대한 정보검색은 수작업 색인정보에 한정되어 있어 정보검색의 정확성 및 효율성이 떨어질 뿐만 아니라 문서 전체 내용에 대한 검색은 많은 시간비용이 소요되고 있는 실정이다. 따라서 소장하고 있는 고문서 이미지데이터에 대한 정보검색의 속도와 정확성을 향상시키고 문서 전체 내용에 대한 검색과 활용성을 개선하기 위해서는 문서인식을 통한 전문검색(full-text retrieval)서비스 제공이 요구되고 있다.

전문검색서비스의 실현은 이미지데이터에 대한 전문인식기술의 결합이 필요하고, 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

※ 본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사 드린다.
† 정 회 원 : (주)인지소프트 책임연구원
논문접수 : 2009년 11월 4일
수 정 일 : 1차 2010년 2월 19일
심사완료 : 2010년 3월 4일

문서의 구조적인 특징, 문서이미지의 품질 변화에 따라 인식성능이 크게 변하는 불안정성으로 인해 문자인식 기술을 전문검색에 활용하는데 큰 장애요인으로 작용하고 있다[1-4]. 반면 국외 대학교, 연구기관, 공공기관 등에서 진행하는 전자도서관 구축 및 소장 자료 디지털화 사례에서 보면, 대상 기관과 기업, 대학교의 연구기관 등이 참여하여 대상 자료를 분석하고 그 자료에 특화된 이미지처리 및 문서분석 기술, 문자인식 기술을 연구 개발하여 적용시키고 있다[5-13]. 특히 해당기관의 이미지데이터로부터 특화된 문서구조정보를 DB화하여 문서인식에 활용하려는 연구사례가 있는데, 사전에 이미지데이터로부터 추출한 논리적인 문서구조정보를 이용하여 제목과 저자 등의 텍스트 정보를 추출하는 연구[16], 사전에 저장된 문서의 물리적인 구조정보를 확률기반의 매칭 알고리즘을 적용하여 문서를 분석하는데 활용하는 연구[17]가 해당된다. 세계적으로 문서인식기술은 이미 연구단계를 넘어서 실용화 단계에 접어들었지만, 모든 문서 자료를 완벽하게 인식할 수 있는 일반화된 기술은 개발되지 않은 것이 현실이다. 따라서 국가적인 지식이나 기관의 소장 자료의 검색 및 정보공유를 위한 디지털화 사업에서는 기존 방법의 한계점을 극복하기 위한 방법론과 다양한 필드테스트에 의한 휴리스틱 노하우(heuristic knowhow)를 활용하고 있다.

본 연구는 국내 대학교, 도서관, 공공기관 등에 소장되어 있는 고문서 이미지데이터의 전문검색 서비스 구현을 위한 문서인식시스템 구성에 대한 연구 및 개발 방법에 대해 제안하고 있으며, 실제 국가기록원 소장 자료 디지털화를 위한 연구 개발 사례를 들어 설명한다. 본 논문의 2장에서는 국내외 고문서 디지털화 동향 및 방법에 대한 조사 내용을 요약하고, 3장 및 4장에서는 본 논문에서 제안하는 문서인식 시스템에 대한 내용과 방법에 대해 기술한다. 마지막으로 4장 및 5장에서는 연구 및 개발의 결과를 통한 성능 평가와 결론을 맺는다.

2. 국내외 고문서 디지털화 동향 및 방법 조사

2.1 국내 사례 조사

국내에 존재하는 방대한 고문서들은 역사, 문화, 정치, 경제, 공공분야 등 다양한 분야로 분류되어 해당 기관에서 소장하고 있다. 이와 같은 고문서들은 그 시대의 생활상이나 제도 및 경제, 정치, 문화적인 상황 등을 이해하는데 중요한 단서가 된다는 점에서 그 역사적 가치나 보존가치가 높다. 현재 국내에 소장되어 있는 고문서에서 역사적인 사료의 경우 대부분 한자로 이루어져 있고 필사본 또는 목판본 이미지 형태이며, 공공기관의 정치, 경제, 사회 전반에 걸친 고문서의 경우 한글, 한자, 영숫자 등이 복합적으로 이루어져 있고 활자본 또는 타이프라이터 등으로 인쇄된 형태를 갖는다. 궁극적으로 자료가 어떤 형태이든지 국내 기관에 소장되어 있는 고문서 자료의 완전한 활용서비스를 위해서는 자료의 이미지화 작업과 해당 이미지에 대한 텍스트데이터 변환 기술, 자동 색인기술 및 검색기술이 함께 결합되어야 한다. 그러나 고문서 자료의 다양한 인쇄 특징이 소장기간이 오래되어 원본의 노후화에 의한 문제와 결합되어 이미지데이터의 텍스트데이터 변환 기술에 의한 디지털화에 가장 큰 걸림돌로 인식되고 있다[14, 15]. 따라서 국내 소장자료 디지털화 사업은 고문서에 대한 텍스트데이터 획득 과정에 집중되어 있고, 크게 수작업 입력(manual typing)에 의한 방법과 문자인식 기술을 활용한(OCR-based) 방법으로 분류된다.

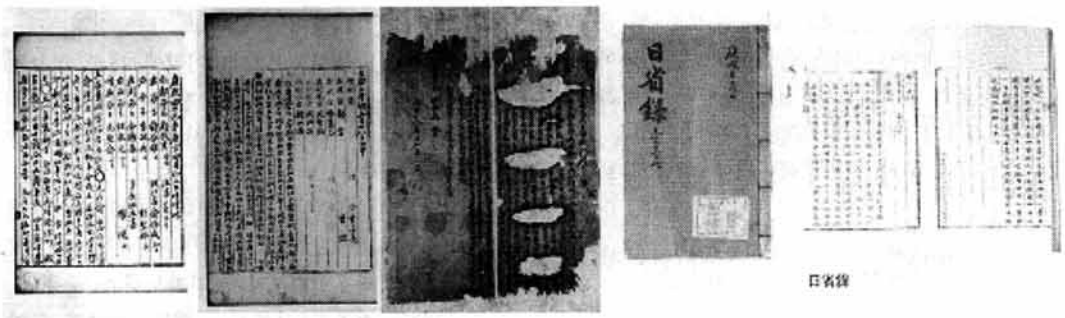
2.1.1 수작업 입력에 의한 방법

국내 역사, 문화적인 사료의 경우 대부분 수작업 입력에 의한 디지털화 방법에 의해 텍스트데이터를 획득하는데, 이것은 오랜 소장기간뿐만 아니라 필사본, 목판본과 같은 문자의 비정형성이 매우 크기 때문이다(그림 1).

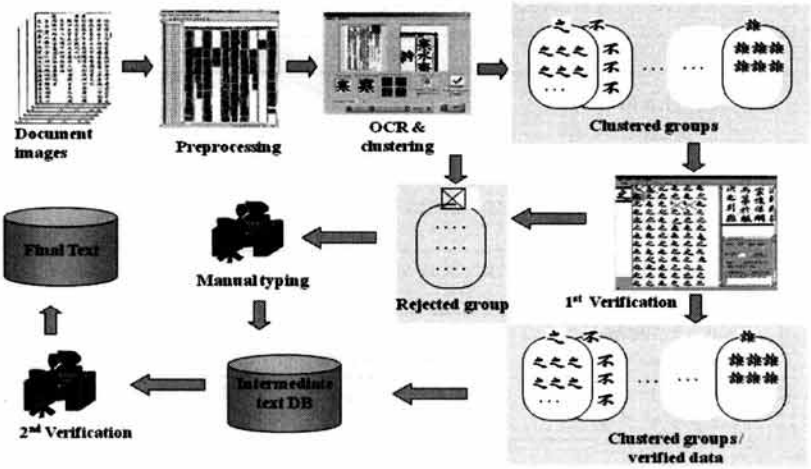
수작업에 의한 디지털화 방법은 문자의 비정형성 때문에 문자인식 기술의 성능이 현실성이 없기 때문이며, 실제로 2005년부터 활발히 진행된 국가지식 정보자원 관리 사업에서 역사, 문화적 한자사료들은 많은 시간과 비용, 인력을 동원하여 소장기관의 자료를 디지털화하고 대국민서비스하고 있다[15].

2.1.2 문자인식 기술을 활용한 방법

실질적으로 역사, 문화적인 고문서에 대해 문자인식 기술을 활용하여 처리가 불가능한 것은 고문서 이미지의 품질이 훼손 및 노후화에 의해 매우 낮다는 것과 고문서 한자에 대한 인식 성능이 매우 낮다는 것에서 기인한다. 또한, 고문서의 한자가 정형성이 없는 필사본 또는 목판본이기 때문이며,



(그림 1) 역사, 문화적인 사료 이미지의 샘플



단계	처리 내용	비고
Preprocessing	고문서 이미지의 품질 향상 및 구조분석 단계	자동화 단계
OCR & clustering	문자인식 및 동일한 클래스로 클러스터링 단계	자동화 단계
Clustered groups	문자인식에 의해 분류된 한자 그룹	
Rejected groups	문자인식 과정에서 인식 불가능으로 분류된 그룹	
1st Verification	전문인력에 의해 검증 및 교정하는 단계	수작업 처리
Manual typing	문자인식으로 처리 불가능한 한자 입력 단계	수작업 처리
2nd Verification	전문인력에 의한 최종 검증 및 검수 과정	수작업 처리
Final Text	최종 텍스트데이터 DB	

(그림 2) 문자인식 기술 및 수작업 검증 인터페이스가 결합된 디지털화 시스템

인식해야 하는 한자가 수천 클래스에 해당하기 때문이다. 반면, 고문서의 디지털화 과정에서 문자인식의 기술적인 한계를 극복하려는 연구가 있었다. 연구 [14, 15]는 고문서 디지털화에서 문자인식 기술과 수작업 검증 작업을 적절히 병행하여 전체적인 디지털화 처리율과 품질을 높이는 방향으로 시스템을 구성했다(그림 2). 본 연구에서는 수작업 디지털화 방법에 비해 문자인식 기술을 활용한 방법이 시간 및 비용 측면에서 효율적일 뿐만 아니라, 수작업 입력 방법에 비해 오류율이 현저히 줄어드는 효과를 기대할 수 있다. 따라서 고문서 자료에 대한 텍스트데이터 구축 과정이 훨씬 빨라지고 방대한 규모의 자료를 처리할 수 있으며, 검증 인력 활용 측면에서도 훨씬 효율적이었다.

문자인식 기술을 활용한 디지털화의 또 다른 방법은 완전하게 문자인식 기술만을 활용하는 방법으로 어느 정도 정형성이 있는 활자본 데이터 처리에 활용되는데, 문자인식에 의한 오류가 존재하더라도 고품질 텍스트데이터 획득보다는 검색이 가능한(searchable) 형태까지를 목표로 하는 방법이다[1-4]. 국내 공공기관 등에 활자본 형태로 소장되어 있는 고문서 자료의 경우, 한글, 한자, 영숫자가 복합적으로 존재하며 어느 정도 정형적인 문서구조로 인쇄되어 있는 것이 특징이다. 그러나 최근 컴퓨터에 의해 양산되는 문서와 비교하면 그 품질이나 형태가 문자인식의 성능을 저해하는 요소를 상당히 포함하고 있기 때문에 국내외 상용화 문자인식 기술 및 제품으로는 한계가 있다. 이것은 문자인식 기술이 이미 상용화 단계를 넘어섰지만, 모든 고문서 자료를 안정적으로 인식 처리할 수 있는 일반화된 기술 및 제품이 존재하지 않는 것에 기인한다. 국외 사례를 보면, 국가적인 지식(national knowledge)이나 기관의 소장 자료의 검색 및 정보 공유를 위한 디지털화 사업에서 기존 방법의 한계점을 극복하기 위한 방법론과 다양한 필드테스트에 의한 휴리스틱 노하우를 활용하고 있다[5, 8-13].

2.2 국외 사례 조사

고문서 자료를 디지털화하기 위해 사용되는 일반적인 국외 문자인식 기술의 경우, 영어, 프랑스어, 독일어 등 알파벳을 사용하는 언어에 대한 문자인식은 알파벳 자체의 단순성 및 소수의 인식 대상(영어의 경우 대소문자 52자 및 숫자 등), 수십 년에 걸친 연구개발 역사로 인해 99.8% 성능을 보유한 제품이 다수 상용화되어 있다. 이러한 인식 성능을 바탕으로 한 문자인식 기술은 전문검색, 전자 도서관 구축, 등의 고문서 처리 등에 폭넓게 활용되고 있다. 최근에는 한자문화권의 방대한 소장자료를 보유하고 있는 중국 및 일본 등에서 한자인식에 대한 연구가 활발히 진행되고 있으며, 소장자료의 디지털라이브러리(digital library) 구축에 대한 연구도 이루어지고 있다. 고문서 디지털화에 대한 또 다른 특징으로 문자인식 및 검색 기술을 활용하여 역사, 문화, 사회 전반의 고문서 자료를 디지털화하는 사례는 기업과 대학교의 연구기관 등이 협업하여 진행하는 경우가 많다. <표 1>은 국외 대학교, 도서관, 공공기관 등에서 진행한 고문서 자료 디지털화 사례를 대상자료 및 접근방법, 결과 측면에

〈표 1〉 국외 소장기관의 역사적인 사료 디지털화 방법 조사

연구 개발 내용	접근 방법(기술)	결과	국가(연구기관)	비고
중국 사고전서(四庫全書)의 디지털화[8]	한자인식(OCR) + 유니코드 기반 검색엔진	OCR 성능 91%. 전자판 배포 (www.skqs.com)	중국 (Digital Heritage Publishing Ltd. 등 7개 기관)	8억자 디지털화 과정에서 18년 소요, 4,000명의 OCR결과 검증 인력 투입
인도제국의 역사적 문서 영상에 대한 패턴인식 프로그램 개발[9]	필기문서의 주제어검색을 위한 OCR 기술	인식률 82.5%	미국 (Caltech: California Institute of Technology)	International Digital Archives Initiative 프로젝트
Senator John Heinz의 국회보고서 디지털화 시스템 개발[5]	OCR + 자연어처리기술 + 검색엔진	HelioScan v2.0 + CLARIT(검색)	미국 (CMU: Carnegie Mellon Univ.)	OCR 결과의 수작업 검증, 필기데이터는 수작업 입력
문화적 사료의 디지털화를 위한 핵심기술 개발 (Gamera)[10]	OCR + OMR(Music Recognition)	Gamera (http://dkc.jhu.edu/gamera/)	미국 (Milton S. Eisenhower Library)	전문가의 검증을 위한 GUI 결합
한자 필기문서의 자동인식을 위한 영상처리 기술 개발[11]	영상 전처리 기술 + 문서 구조분석 및 분할 기술	필기문서 영상처리 및 분할기술	일본 (National Institute of Japanese Literature)	University of California Berkeley와의 공동연구
티베트 언어 목판 불교문서의 디지털화[12]	자동 분할 기술+ OCR	OCR 성능 99%	일본 (Tohoku Institute of Technology)	적용 데이터: 자동 분할 141,988자, OCR 17,753자
명/청 왕조 궁중기록 디지털화[13]	OCR + 전문검색엔진	Transmission Text Retrieval System	대만(Transmission Information System Ltd.)	4년 동안 3백만 페이지의 31만 주제에 대한 디지털화

서 살펴본 것이다[5, 8-13]. 가장 특징적인 것은 문자인식 및 문서처리 기술을 활용한 접근방법에서 대상자료에 대한 분석을 통해 특화된 이미지처리 및 문서분석 기술, 문자인식 기술을 연구 개발하여 적용시키고 있다는 것이다.

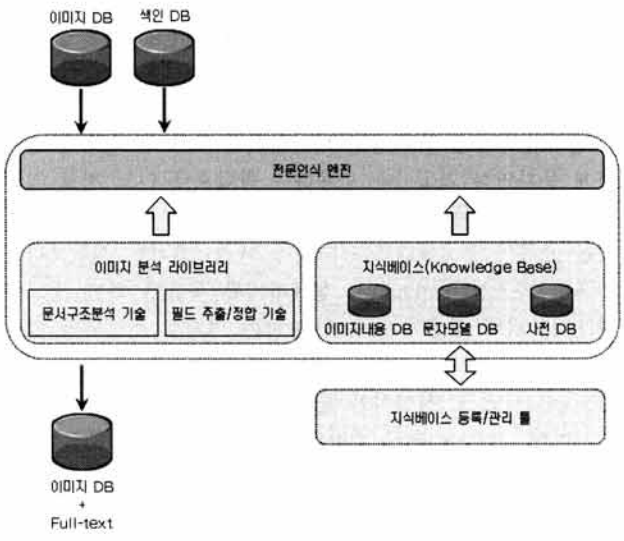
3. 연구 내용

본 연구의 목표는 공공기관이 소장하고 있는 역사, 문화, 사회 전반에 걸친 고문서 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 지식베이스(knowledge base)를 구성하는 것이다. 따라서 고문서 이미지데이터를 검색이 가능한 텍스트데이터로 변환하기 위한 문서인식시스템 구성 시 이미지분석 라이브러리 및 지식베이스를 활용하여 고품질의 안정적인 데이터를 생성할 수 있도록 제안한다. 또한, 향후 확장성을 고려하여 지식베이스를 생성하고 지속적으로 관리할 수 있는 툴을 개발하는 것이다(그림 3).

본 연구는 국가기록원의 소장자료를 대상으로 진행되었으며, 소장자료가 기본적으로 문서의 시대적인 특성, 작성 환경, 문서의 구조, 원본의 상태 등 다양한 특성을 가지고 있기 때문에 소장자료의 분석을 통해 특성을 파악한 후에 최적의 접근방법을 도출하였다. 본 연구의 결과로 국가기록원을 비롯하여 공공기관의 소장 이미지데이터의 검색은 기존 색인DB를 통한 정보검색 및 이미지조회 서비스에서 자료의 전문검색서비스까지 확대될 것이며, 그 핵심기술 기반인 특화된 문서인식 시스템 모델을 연구 개발하는 것이 본 연구의 목적이다.

고문서 이미지데이터 활용을 위한 특화된 문서인식시스템 모델 연구 개발은, 먼저 소장자료 분석을 통해 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 구조분석 방법 중 소장자료의 문서 구조분석에 가장 적합한 방법을 선택하여 라이브러리로 구성하고, 추출된 문서구조를 등록된 DB와 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. 문서 블록별로 선택적으로 사용할 수 있도록 한글, 한글과 영문, 한글과 한자 문자인식기를 구성하고, 용어사전은 소장자료의 분야에 따라 나누어서 구성한 후 인식 과정에서 선택된 용어사전을 참조하여 인식 결과를 최적화할 수 있도록 구성한다.

본 연구의 주요 내용은 다음과 같다.



(그림 3) 제안하는 문서인식시스템 모델

(1) 소장자료 조사 및 분석
(2) 문서영상 전처리 및 구조분석 라이브러리 개발
(3) 이미지내용DB 설계 및 관리도구 개발
(4) 문자모델DB 설계 및 관리도구 개발
(5) 용어사전DB 설계 및 관리도구 개발
(6) 지능형 인식라이브러리 결합 인식시스템 개발

4. 연구 및 개발 방법

4.1 소장자료 조사 및 분석

공공기관에서 소장하고 있는 이미지데이터는 크게 정기간행물과 내부문서의 부류로 나누어지며, 이러한 소장자료들은 연간 발생하는 규모나 형태가 유사하거나 점진적으로 변화 또는 확대되고 있다. 따라서 본 연구에서는 특정 년도를 기준으로 연구 개발에 필요한 데이터를 수집하기 위해 조사 및 분석 작업을 진행했다. 소장자료 조사를 통해 문서를 유형별로 분류하고 각 문서 유형에 따른 특징적인 요소를 도출하여 정리한다. 이러한 조사 및 분석 내용은 지능형 인식라이브러리의 문서영상 처리 기술 및 지식베이스를 설계하고 개발하는 것에 대한 방향을 수립하기 위한 것이다. 다음 (그림 4)는 정기간행물 및 내부문서를 조사 분석하는 과정에서 작성하는 분석 시트 일부분을 발췌한 예이다.

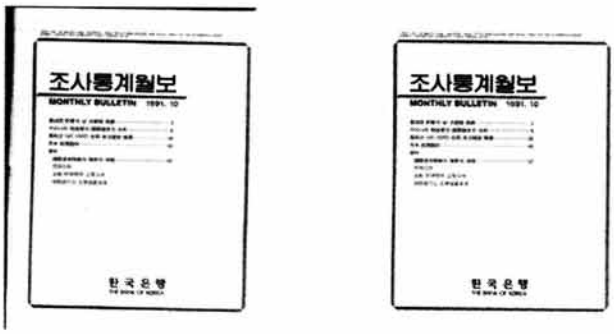
구분	현황	샘플수집	문서특징	개발방향수립
정기간행물	-2005년 기준 : 9779건 존재	정기간행물 → 56건(권)	-일반적인 서적형태 -종류 별로 통일된 형식 존재 -장기 발행된 문서의 초기본은 문서상태가 좋지 않음	-이미지 상태에 따른 전처리 모듈의 추가 ex) noise제거, 기울기보정 -통일된 형식을 갖는 문서의 인식 효율성을 높이기 위해 페이지구성 정보와 문서 내용구조 정보를 등록하고 관리할 수 있도록 함 -타자체전용 문자모델 필요
내부문서	-1990년 기준 :10만 건 존재 :원본확인 → 1500건 -1997년 기준 :3만 건 존재 :원본확인 → 1350건	-1990년도 → 25건(철) -1997년도 → 35건(철)	-<기안문+첨부>형태 -기안문은 정형성 존재 -첨부는 매우 다양한 형식을 갖고 있음 -1990년도 기준 문서는 타자체 존재	

문서제목	연도	스캔상태	구조현태	비고
국회보	1989.1-4	중	잡지	한자, 2단문서
	1995.10	중하	"	잡지스타일, 2단 3단 혼용
	2004.12	상	"	배경에 이미지 존재
건설기술정보	1993.10	하	2단	단 구분선 존재
	2004.10	중상	"	가끔 3단과 표 존재
	2005.09	상	"	
고층기상월보	1993.10	중하	표	시계반대방향 90도 회전된 표 존재
	2007.01	상	"	"
산업생산통계	1992.10	중하	표	한자
	2006.11	상	"	

(그림 4) 소장자료 조사 및 분석 시트 샘플

4.2 문서영상 전처리 및 구조분석 라이브러리 개발

소장자료 분석 과정에서 추출된 샘플 이미지데이터는 전처리(preprocessing) 단계를 통해 문서구조분석 및 분할, 인식 등의 성능을 저해하는 요인을 제거한 이미지로 재생산된다. 소장자료 분석 결과, 가장 기본적으로 고려되어야 하는 전처리 기술 항목은 문서영상 전반에 분포하는 잡영(noise) 제거와 문서영상의 기울어짐(skewness) 보정이었다(그림 5). 이것은 소장자료의 작성 및 보관이 오래되고 종이의 품질이 저급이어서 스캔과정에서 원본의 상태에 따라 이미지의 품질이 영향이 받는 것에서 기인한다. 또한, 문서영상의 외곽



원본 이미지 이미지 전처리 후

(그림 5) 이미지데이터 전처리 결과

검은 테두리 제거 및 180도 회전된 이미지데이터의 자동 보정 등의 기능도 추가한다.

문서구조분석 및 영역을 분할하는 방법에는 전체 영상을 프로파일(projection profile) 등을 이용해 잘게 쪼게 나가는 Top-down방식과 최소 연결요소(connected component)를 추출한 후 이를 동일한 속성을 가진 요소끼리 묶어 나가는 Bottom-up방식이 있다. 본 연구에서는 텍스트 블록의 누락이 적게 발생하고 문자열 추출이 용이한 Bottom-up방식을 이용하여 영역을 분할하였다. 일반적인 문서구조분석 방식에서는 그래픽 요소의 근처에 배치된 문자나 문자 블록이 누락되는 경우가 많이 발생하는데, 본 연구에서는 문서 구조 분석의 추출 단위를 단어나 문자열 규모로 축소하여 가급적 많은 양의 텍스트 변환 정보를 획득할 수 있도록 한다. 이것은 인식 과정에서 지식베이스를 이용한 문서구조분석의 후처리(postprocessing) 과정에서 보다 정확한 지식을 이용하여 검증 및 교정하는 단계로 연결되기 때문이다.

4.3 지식베이스 설계 및 관리도구 개발

본 논문에서 문서인식 시스템의 성능을 최적화하기 위해 제안하는 지식베이스는 이미지내용DB, 문자모델DB, 용어사전DB로 분류된다. 세가지 지식베이스를 결합하여 입력 이미지데이터에 적합한 이미지내용정보를 생성하여 인식 및 처리 과정에 활용한다. 각각의 지식베이스는 관리도구를 통해 생성되고 관리할 수 있다.

4.3.1 이미지내용DB 설계 및 관리도구

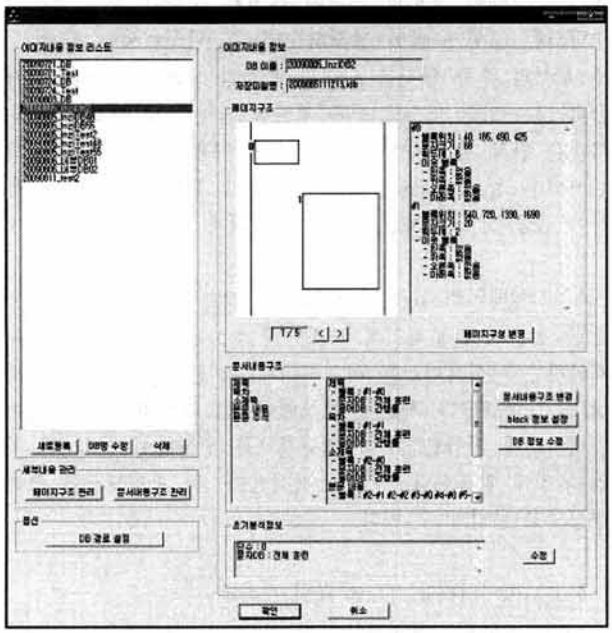
이미지내용DB는 문서영상의 물리적인 구조(physical structure), 논리적인 구조(logical structure) 정보를 포함하도록 설계되는데, 문서의 유형과 문서의 블록(block) 구조, 직선 등의 블록 구분자 등에 대한 정보도 포함한다. 또한, 블록의 속성정보도 포함하는 블록의 유형을 나타내는 지시어(목차, 요약, 작성자 등)와 문자모델DB에 대한 인덱스(index), 용어사전DB에 대한 인덱스가 포함된다. <표 2>는 이미지내용DB의 구성 및 속성을 나타낸다. 또한, 문서는 여러 개의 하위문서를 가질 수 있기 때문에 계층적(hierarchical) 모델을 기반으로 다양한 형태의 문서를 표현할 수 있도록 한다.

따라서 이미지내용DB 관리도구는 상기 정보를 생성, 수

〈표 2〉 이미지내용DB 정의 및 속성

구분	정의	속성
초기 분석용 정보	이미지 전처리 및 구조분석에 사용	단수 초벌 인식용 문자모델DB 인덱스
페이지 구성 정보	페이지 내에 존재 가능한 요소를 블록으로 정의, 한 건의 문서에 존재하는 페이지 유형을 모두 저장	블록 위치 정보 블록 정렬 정보 내부 라인의 정렬 정보 블록 간의 위치관계 정보
문서내용구조 정보	전체 문서의 논리적인 내용구조를 정의	내용 레이블 레이블 순서 정보 레이블에 해당하는 블록 인덱스
인식용 정보	문서내용구조 정보에 정의된 레이블에 대한 인식정보를 정의	문자모델DB 인덱스 용어사전DB 인덱스

정, 삭제할 수 있는 기능을 제공한다(그림 6). 이미지내용정보 리스트에서 이미지내용DB를 선택하면 인식에 사용할 이미지내용DB를 쉽게 선택할 수 있도록 페이지구조의 레이아웃과 문서내용구조의 레이블 구조를 상세히 볼 수 있다. 또한, 페이지구조 정보와 문서내용구조 정보는 일반적인 문서에서 사용 가능한 대표적인 정보를 미리 저장할 수 있게 되어 있고 새로운 이미지내용DB를 구성할 때 이러한 정보를 선택해서 사용할 수 있도록 되어 있다.



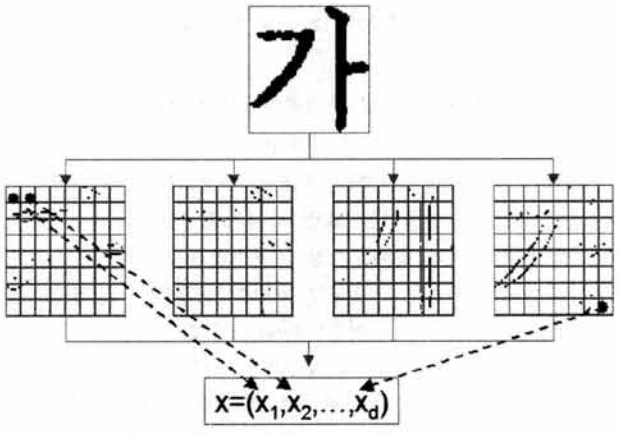
(그림 6) 이미지내용DB 관리도구 구성

4.3.2 문자모델DB 설계 및 관리도구

문자모델DB는 입력 이미지데이터에 가장 적합한 문자모델을 이용하여 인식을 수행함으로써 최적의 인식결과를 얻기 위해 문자 유형별로 샘플 문자영상들을 수집한 정보이다. 각 문자모델DB는 훈련을 통해 해당 문자영상들로부터 인식에 필요한 데이터파일을 생성할 수 있도록 설계되었다. 소장하고 있는 자료 분석결과 사용되는 대표적인 문자세트(set)는 한글 2,350자 이외에도 한자 4,881자, 영숫자 및 특수 문자가 있다. 본 연구에서는 7,000개 이상의 많은 문자클래스를 구분해야 하는 문자 세트의 특성으로 인해 확률 통계적 방법을 활용하며, 확률 통계적 방법 중 문자클래스와 문자영상 특징 벡터간의 거리를 측정하여 문자영상을 분류하는 최소거리 분류(Minimum Distance Classifier) 방식을 이용하여 문자모델을 구성하였다. 거리 측정 방법은 선형거리함수(LDF: Linear Distance Function), 수행 시간은 더 소요되나 나은 성능을 보이는 2차형거리함수(QDF: Quadratic Distance Function), 마할라노비스(Mahalanobis)거리함수, 2차형거리함수와 마할라노비스거리함수를 결합한 함수를 사용한다.

$$g_{QDF}(x,c) = (x-\mu_c)\Sigma_c^{-1}(x-\mu_c) + \log|\Sigma_c|$$

여기서 x는 입력특징 벡터이며 c는 클래스를 나타낸다. 그리고 μ_c는 클래스의 중심 벡터이고 Σ_c는 클래스의 공분산 매트릭스(covariance matrix)를 나타낸다. 또한, 매칭(matching)에 사용할 특징은 클래스에 대한 변별력이 크고, 변이에 대한 흡수력이 뛰어난 4방향(Horizontal, Vertical, Diagonal, Inverse Diagonal) 윤곽선 방향 특징(contour directional feature)을 사용하는데, 메쉬(mesh)의 크기는 실험을 통해 결정하였다(그림 7).



(그림 7) 4방향 윤곽선 방향 특징(8×8 mesh)

(그림 8) 문자모델DB 관리도구 구성

문자모델DB 관리도구는 사용자가 이미지내용 DB관리 프로그램에서 각 블록에 적합한 문자모델을 설정할 때 활용할 수 있도록 한다(그림 8). 사용자는 관리도구에서 인식대상 블록의 문자영상과 가장 유사한 샘플문자영상들을 훈련시켜 얻은 문자 모델을 선택할 수 있다. 문자모델DB 관리도구는 문자모델 보기, 문자모델 추가, 샘플 문자영상 추가 등의 기능을 제공한다.

또한, 문자모델DB 관리도구는 사용자가 활용하고자 하는 여러 개의 샘플문자영상들을 통합한 뒤, 직접 학습(training)시켜 새로운 문자모델을 생성하는 기능을 제공한다. 이 기능은 사용자가 새로운 문자 형식, 즉 새로운 폰트에 대한 적응력을 향상시키고 특정한 폰트만을 사용하는 문서에도 손쉽게 적합한 문자모델을 생성하는 것이 가능하게 된다. 문자모델DB 관리도구에서는 문자모델 학습기로 선형거리함수, 2차형거리함수, 마할라노비스거리함수, 2차형거리함수와 마할라노비스거리함수를 결합한 함수를 선택할 수 있다.

4.3.3 용어사전DB 설계 및 관리도구

용어사전DB는 소장자료에 적합한 용어사전을 구성하여 인식 결과에 대한 최종의 검증 절차에 활용할 수 있다. 소장자료로부터 추출한 용어를 저장하기 위한 구조로는 Trie (prefix tree)를 사용하였다. Trie는 주로 문자열 목록을 저장하는데 사용되는 순서트리 데이터구조(ordered tree data structure)로 검색이 빠르고, 단어의 추가, 삭제가 용이하다는 장점이 있다. 용어사전DB 관리도구는 용어DB 생성 및 관리가 가능하도록 개발하였는데, 저장되어 있는 용어DB 목록을 확인할 수 있고 선택된 용어DB에 대해서는 관련 정보가 출력되며, 용어DB의 생성, 수정, 삭제할 수 있도록 기능을 제공한다(그림 9).

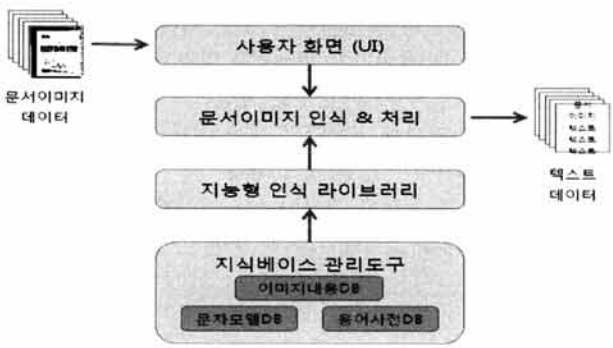
4.4 문서인식 프로토타입 시스템 개발

본 연구는 상기와 같이 소장자료에 대한 분석 결과로부터 도출된 이미지 전처리 및 구조분석 기술 개발과 이미지내용 DB, 문자모델DB, 용어사전DB의 지식베이스를 결합하여 지능형 인식라이브러리를 구성하였으며, 문서이미지 인식 및 처리를 위한 프로토타입(prototype) 시스템을 개발하였다.

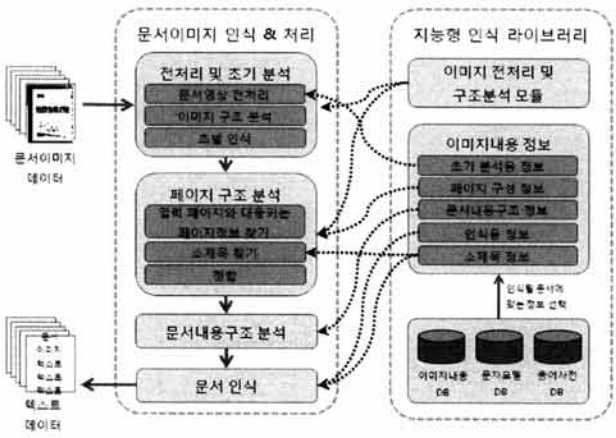
(그림 9) 용어사전DB 관리도구의 구성

(그림 10)과 (그림 11)은 프로토타입 시스템의 블록 다이어그램을 나타내고 있다.

본 연구의 문서이미지 인식 및 처리는 전처리 및 초기분석, 페이지구조 분석, 문서내용구조 분석, 문서인식의 4단계로 구성된다.



(그림 10) 문서이미지 인식 및 처리 시스템 다이어그램



(그림 11) 문서인식 부분과 지능형 인식라이브러리 세부 구성

4.4.1 전처리 및 초기분석 단계

전처리 및 초기분석 단계는 문서영상 전처리, 이미지 구조 분석, 초벌 인식의 3단계로 구성된다. 문서영상 전처리 단계에서는 소장 자료 분석을 통해 확인된 문서영상 전반에 분포하는 잡영(noise)을 제거하고 문서영상의 기울어짐(skewness) 보정을 수행한다. 이미지 구조 분석단계에서는 전처리를 통해 재생성된 이미지로부터 문서의 구조를 분석하여 텍스트라인 블록을 추출하고 텍스트라인 블록의 구성 정보로부터 문단 블록을 추출한다. 문단 블록은 이 후의 페이지구조 분석 단계에서 DB페이지와의 대응 관계를 결정하기 위해 기본요소로 사용된다. 초벌 인식 단계에서는 추출된 텍스트라인 블록들을 인식하여 인식된 문자열 결과와 폰트크기, 폰트두께와 같은 문자열 정보를 추출한다. 각 문단 블록은 문단블록을 구성하는 텍스트라인 블록의 폰트크기와 폰트두께를 평균한 값을 속성정보로 이용한다.

4.4.2 페이지구조 분석 단계

페이지구조 분석 단계에서는 입력 문서이미지의 초기 분석 결과에 대응하는 지식베이스의 대표 DB페이지를 찾는 과정과 해당 대표 DB페이지의 블록들과 입력된 문서이미지의 블록들간의 대응정보를 결정한다. 여기에서 대표 DB페이지는 이미지내용DB에 저장된 각 문서 건의 대표페이지로, 블록 구성 및 논리적인 배치 정보를 포함하고 있다(그림 12).

페이지구조 분석 단계는 입력된 문서이미지의 초기분석 결과로부터 블록들을 조합하여 가능한 모든 페이지 구성을 생성하고, 대표 DB페이지의 블록들과 입력된 문서이미지 페이지 내의 블록간의 대응 점수를 계산하여 대응 정보를 결정하며, 앞에서 계산한 블록간의 대응 점수를 통해 가장 잘 대응되는 DB페이지 구성을 찾게 된다.

블록 조합으로 생성된 문서이미지 페이지는 DB페이지의 블록과 대응관계를 결정하기 위해 블록간 유사도를 계산한다. 그리고 블록간 유사도를 이용하여 페이지 대응 매칭점수를 산출한다. 이 페이지 매칭점수가 가장 높은 블록 조합 페이지 구성이 DB페이지와 매칭되는 페이지로 결정된다.

두 페이지 내의 블록의 유사도는 아래의 식으로 계산된다.

$$S_{AiBj} = S_{PropertyAiBj} + S_{NeighborAiBj}$$

여기서 $S_{PropertyAiBj}$ 는 페이지 A의 i번째 블록과 페이지 B의 j번째 블록의 속성유사도이며, $S_{NeighborAiBj}$ 는 i와 j 블록의 이웃유사도를 나타낸다.

두 블록의 속성유사도는 블록의 기본속성인 폰트크기와 폰트두께를 이용하여 두 대응블록의 유사정도를 나타낸다. 속성유사도는 아래 식으로 계산된다.

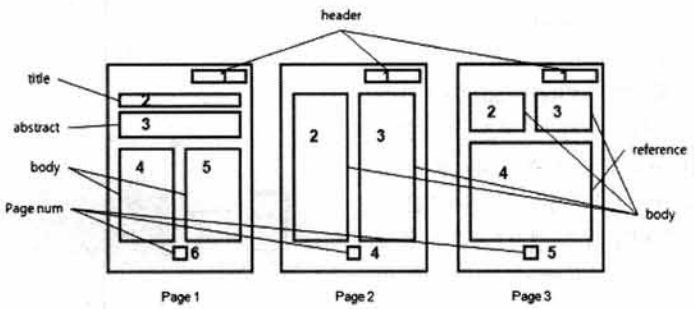
$$S_{PropertyAiBj} = 200 - \left| fs_{Ai} - fs_{Bj} \right| + \left| ft_{Ai} - ft_{Bj} \right|$$

여기서 fs_{Ai} 와 fs_{Bj} 는 각 블록의 대표폰트크기, ft_{Ai} 와 ft_{Bj} 는 각 블록의 대표폰트두께를 나타낸다.

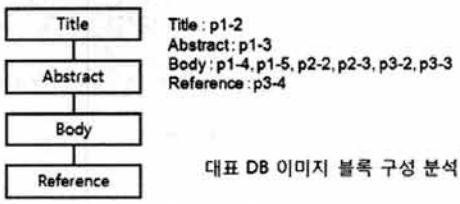
$$S_{AiBj} = S_{PropertyAiBj} + S_{NeighborAiBj}$$

두 블록의 이웃유사도는 대응되는 두 블록의 상하좌우 4방향의 이웃이 얼마나 유사하게 위치하고 있는가를 나타내며 아래의 식으로 계산된다.

$$S_{NeighborAiBj} = S_{NorthAiBj} + S_{SouthAiBj} + S_{EastAiBj} + S_{WestAiBj}$$



대표 DB 이미지 블록 구성



대표 DB 이미지 블록 구성 분석 및 배치

(그림 12) 이미지내용DB의 대표 DB페이지 구성 및 배치 정보

$$S_{NorthAiBj} = 100 \times \left(1 - \left|\frac{OverlapH(W_{Ai}, W_{AiNei_N})}{W_{Ai}}\right| - \left|\frac{OverlapH(W_{Bj}, W_{BjNei_N})}{W_{Bj}}\right|\right) \times Ps$$

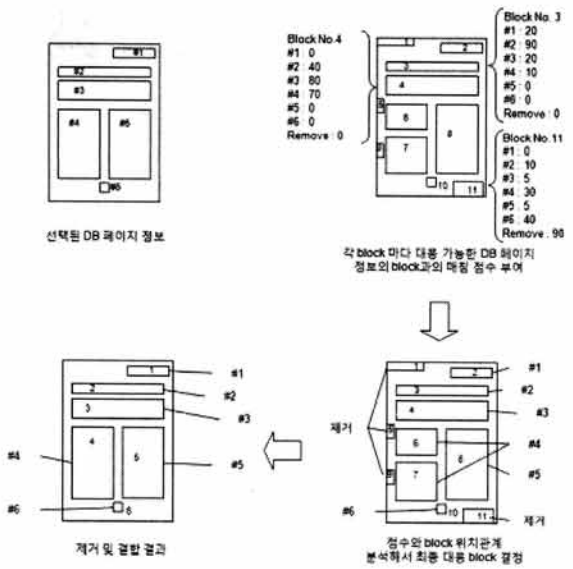
$$S_{SouthAiBj} = 100 \times \left(1 - \left|\frac{OverlapH(W_{Ai}, W_{AiNei_S})}{W_{Ai}}\right| - \left|\frac{OverlapH(W_{Bj}, W_{BjNei_S})}{W_{Bj}}\right|\right) \times Ps$$

$$S_{EastAiBj} = 100 \times \left(1 - \left|\frac{OverlapV(H_{Ai}, H_{AiNei_E})}{H_{Ai}}\right| - \left|\frac{OverlapV(H_{Bj}, H_{BjNei_E})}{H_{Bj}}\right|\right) \times Ps$$

$$S_{WestAiBj} = 100 \times \left(1 - \left|\frac{OverlapV(H_{Ai}, H_{AiNei_W})}{H_{Ai}}\right| - \left|\frac{OverlapV(H_{Bj}, H_{BjNei_W})}{H_{Bj}}\right|\right) \times Ps$$

여기서 $S_{NorthAiBj}$, $S_{SouthAiBj}$, $S_{EastAiBj}$, $S_{WestAiBj}$는 i와 j 블록을 중심으로 상하좌우 4방향의 유사도를 나타낸다. W_{Ai}와 H_{Ai}는 각각 페이지 A의 i번째 블록의 폭(width)과 높이(Height)를 나타내며, W_{AiNei_N}, W_{AiNei_S}, H_{AiNei_E}, H_{AiNei_W}는 각각 블록 i의 상하좌우 4방향의 이웃 블록의 폭과 높이를 나타낸다. 마찬가지로 W_{Bi}, H_{Bi}, W_{BjNei_N}, W_{BjNei_S}, H_{BjNei_E}, H_{BjNei_W}는 페이지 B의 j번째 블록의 정보를 나타낸다. $OverlapH()$와 $OverlapV()$는 각각 두 블록이 수평과 수직으로 중첩되는 구간의 길이를 계산하는 함수를 나타내며, Ps는 가중치(이웃블록이 중첩되는 형태의 유사성, 1 또는 0.8)를 나타낸다.

(그림 13)은 입력된 문서이미지의 블록 추출 결과와 DB 페이지 내의 블록들과 대응정보를 추출한 결과이며, (그림 14)는 대응정보를 통해 입력된 문서이미지와 가장 잘 부합하는 DB페이지를 찾은 결과를 나타낸다. 페이지 매칭점수는



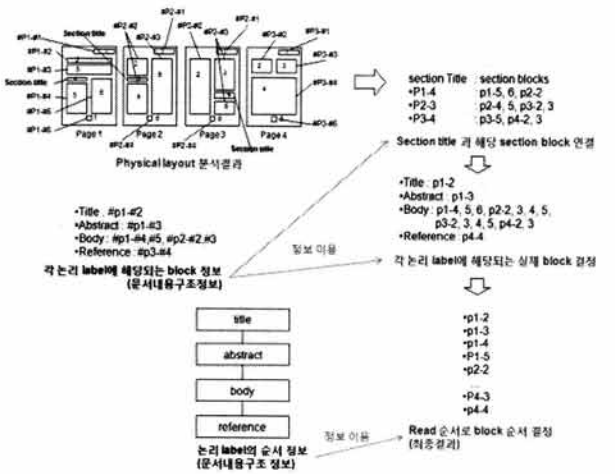
(그림 13) 입력 문서이미지의 블록 추출 및 DB페이지와의 대응정보 추출

(그림 14) 문서이미지에 잘 부합하는 DB페이지 결정 결과

각 블록의 유사도 점수를 모두 합산하고 페이지 내의 블록의 총 면적으로 나눠서 정규화 한다. 페이지 매칭점수가 가장 높은 블록조합 페이지가 최종적으로 DB페이지와 매칭되는 페이지로 결정이 된다.

4.4.3 문서내용구조 분석 단계

문서내용구조 분석 단계는 입력 문서이미지의 각 페이지에 해당하는 블록 추출 결과에 대한 논리적인 순서를 부여하는 단계인데, 이미지내용DB의 문서내용구조 정보를 활용한다. (그림 15)는 입력 문서이미지의 블록이 추출된 결과에 대해 이미지내용정보의 문서내용구조 정보를 이용하여 블록의 논리적인 순서를 추출하는 과정을 나타내고, (그림 16)은 실제 문서 이미지를 처리한 결과를 예시로 나타내고 있다.



(그림 15) 문서이미지의 블록에 대한 논리적인 순서 부여 과정

(그림 16) 실제 문서처리 과정에서 블록에 대한 논리적인 순서 부여 예시

4.4.4 문서인식 단계

문서인식 단계는 논리적인 순서가 부여된 모든 블록에 대해서 인식을 수행하는 단계이다. 블록내의 모든 라인은 이미지내용정보의 인식용 정보를 이용하여 인식되고 인식결과는 논리순서에 맞게 출력된다. 각 블록은 최적의 인식결과를 얻기 위해 가장 적합한 인식용 정보를 사용하게 된다. 인식용 정보는 문서내용구조 정보의 각 논리레이블 별로 지정되어 있기 때문에 각 블록에 사용할 인식용 정보는 문서내용구조 분석 단계에서 결정된 논리레이블의 인식용 정보를 사용하게 된다. 본 연구의 문서인식은 인식용 정보의 문자모델DB를 이용하여 인식하는 단계와 용어사전DB를 이용하여 인식결과를 보정하는 단계로 구성된다.

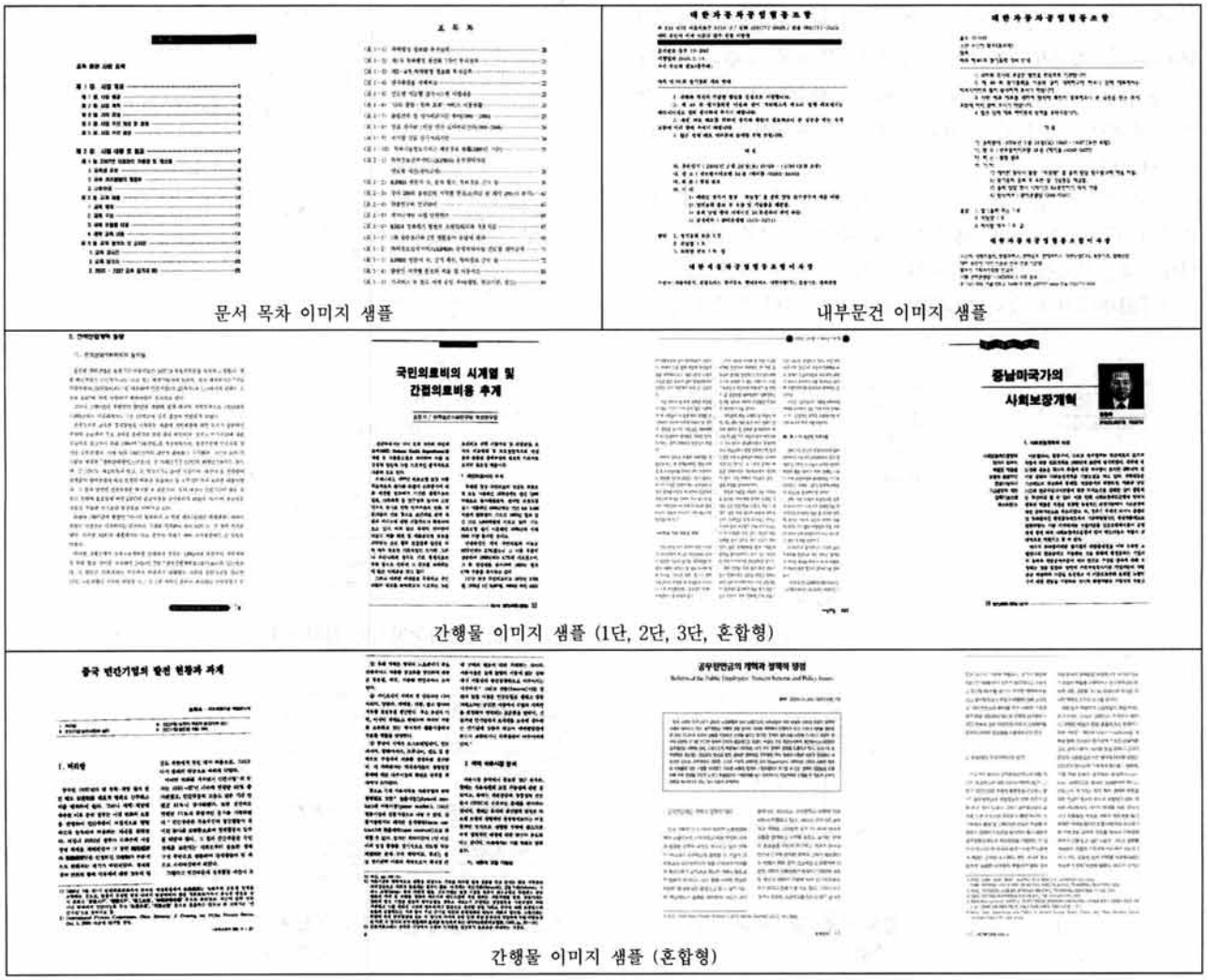
5. 성능 평가

본 장에서는 본 논문에서 제안하는 시스템의 성능평가 결과를 기술한다. 본 논문에서는 다양한 구조의 문서를 테스트 하기 위해서 소장자료 단계에서 취득한 샘플 데이터를 아래와 같이 분류하고 테스트에 이용하였다<표 3, 4>.

〈표 3〉 테스트에 사용한 샘플 문서이미지의 구성

구분	샘플 구성
목차	간행물로부터 목차부분만 추출 521개 영상
내부문건	내부문건은 보안상 많은 샘플을 활용하기는 어려움 대표성 있는 내부 서식을 선별하여 테스트데이터 선정 80개 영상(4종, 각20건) 4종 중 2종은 타자체 폰트(대부분 과거 문서)를 사용한 샘플 선정
간행물	수집된 문서를 구조유형별로 분류 546개 영상 • 1단 - 5종, 총 237개 영상 - 1996년, 2000년, 2005년, 2006년 문서 • 2단 - 5종, 총 245개 영상 - 1989년, 1996년, 2000년, 2006년 문서 • 3단 - 1종, 총 32개 영상 - 2006년 문서 • 혼합 - 3종, 총 134개 영상 - 1996년, 2000년 문서 - 주석 존재, 1단과 2단이 동시에 존재하는 문서들

〈표 4〉 테스트에 사용한 문서이미지 샘플



문서 목차 이미지 샘플

내부문건 이미지 샘플

간행물 이미지 샘플 (1단, 2단, 3단, 혼합형)

간행물 이미지 샘플 (혼합형)

〈표 5〉 성능평가에 사용한 상용 전문인식 제품

구분	설명
A사	외산 전문인식 제품 (다국어 문서인식)
B사	국산 전문인식 제품 (한글, 한자, 영숫자 등 문서인식)

테스트에 사용한 이미지내용 DB는 각 문서 종 마다 별도의 DB를 작성하였고 인식모델은 일반폰트용 모델과 타자체 폰트용 모델을 이용하였다. 그리고 용어DB는 내부문건 테스트에서만 적용하였으며 내부문건에서 사용되는 키워드와 빈번하게 사용되는 단어 위주로 구성을 하였다. 그리고 본 논문에서 제안하는 시스템의 객관적인 비교평가를 위해서 타사 전문인식 제품의 인식결과와 비교하였다.

5.1 인식률 테스트

인식률 테스트를 위해서 각 샘플문서에 대한 정답을 작성하고 각 인식기의 인식결과를 정답과 비교하였다. 두 인식결과를 비교하기 위해서 LCS(Longest common subsequence) 알고리즘을 사용하였다. 그리고 인식기의 구조분석 결과가 정답의 형태와 달라서 인식결과의 텍스트 위치가 정답과 맞지 않는 경우에는 인식결과를 정답의 형태와 유사하게 수정을 하고 결과를 비교하였다.

전체 샘플의 인식률은 <표 6>과 같다. 본 논문에서 제안하는 지능형 문서인식시스템은 목차, 내부문건, 간행물 모두에 대해서 타사의 인식률 결과보다 안정적인 인식률을 유지하고 있는 것을 볼 수 있다. <표 6>과 <표 8>에서는 일반문서의 구조와 다른 형태를 갖는 목차 샘플문서와 타자체 폰트로 구성된 내부문건의 샘플문서에 대해서 타사 인식기의 인식률이 눈에 띄게 저하되는 것을 볼 수 있다. 특히 B사 인식기의 경우는 목차구조를 제대로 처리하지 못하고 일부 텍스트라인 블록이 추출이 안 되는 경우가 빈번하게 발생하여서 인식결과를 정답과 유사하게 수정을 하였는데도 불구하고 인식률 저하가 두드러지게 나타나고 있다. <표 7>의 언어별 인식률 비교결과에서도 본 논문에서 제안하는 시스템은 한글, 한자, 숫자, 영문자의 모든 언어영역에서 인식률이 안정적으로 유지되고 있는 것을 볼 수 있다.

〈표 6〉 문서이미지 종류별 전체 인식률 비교

구분	목차	내부문건	간행물
제안하는 시스템	95.99	92.75	96.09
A사	91.93	80.69	95.88
B사	83.97	86.92	94.18

〈표 7〉 언어별 인식률 비교

구분	한글	한자	숫자	영문자
제안하는 시스템	97.15	75.68	95.88	90.16
A사	92.60	55.05	93.90	77.95
B사	90.55	57.41	84.49	62.89

〈표 8〉 내부문건 폰트별 인식률 비교

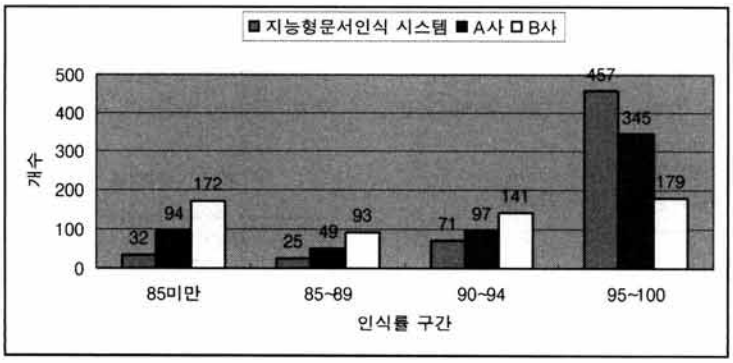
구분	일반폰트	타자체폰트
제안하는 시스템	97.01	89.50
A사	96.45	64.94
B사	96.63	77.23

(그림 17)은 인식률 구간별 문서이미지 개수를 나타낸 도표이다. 본 논문에서 제안하는 지능형문서인식 시스템의 경우 95%이상 인식률 구간에 존재하는 문서이미지가 타사 인식기에 비해서 상대적으로 많이 분포하고 있으며 반대로 낮은 인식률 구간에서는 적게 분포하고 있음을 볼 수 있다.

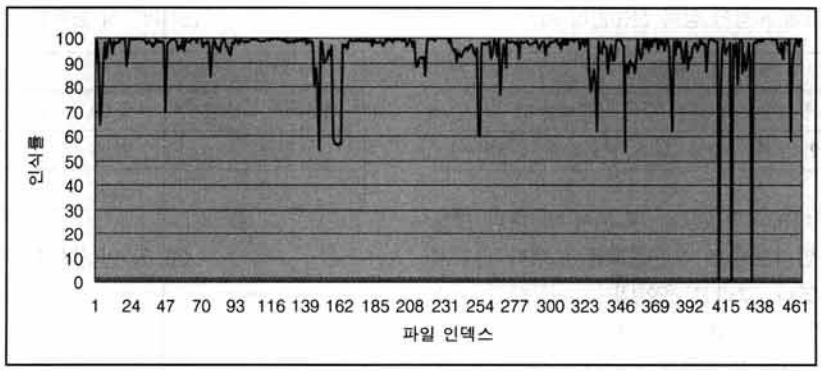
(그림 18)은 목차 샘플데이터에 대해서 각 파일 별로 인식률을 나타낸 그래프이다. 인식률이 0%로 나타나는 부분은 문서구조분석에 실패했거나 분석결과가 비정상적이기 때문에 처리하지 못한 데이터를 나타낸다. 타사 인식기의 그래프에서는 문서이미지에 따라서 인식률 변화가 심하게 나타나고 있다. 반면에 본 논문에서 제안하는 시스템은 인식률 변화가 크지 않고 상대적으로 안정적인 성능을 보이고 있다.

5.2 문서구조분석 테스트

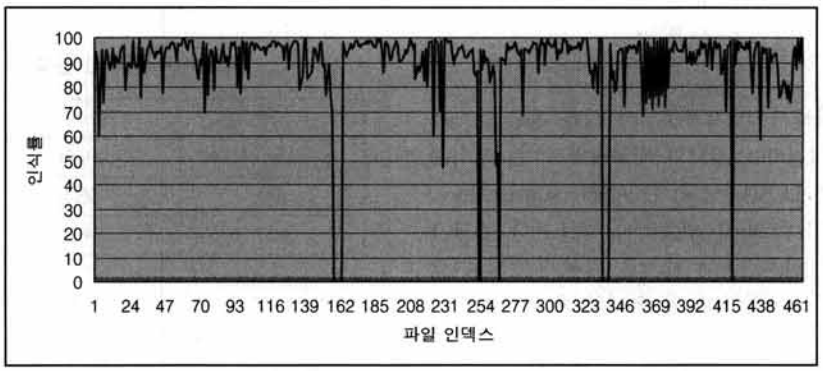
문서구조분석 테스트는 각 문서의 문단 블록의 구조, 문단 블록의 순서, 텍스트라인 블록 추출이 정상적인 문서이



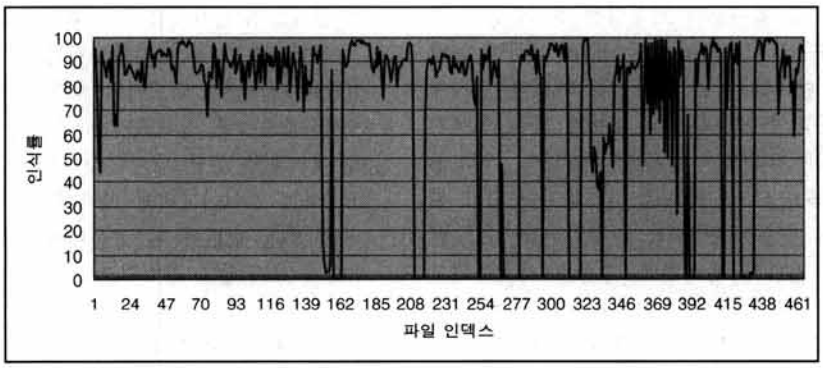
(그림 17) 인식률 구간별 문서이미지 개수 비교



(a) 제안하는 시스템



(b) A사 전문인식 제품



(c) B사 전문인식 제품

(그림 18) 목차 샘플이미지 데이터별 인식률 분포 비교

미지의 비율을 측정한 것이다. 테스트는 각 인식기에서 제공하는 문단블록과 텍스트라인 블록 추출결과를 측정자가 직접 육안으로 확인하고 위의 세가지 요소에 부합하는 문서 이미지를 기록하는 방법으로 진행하였다. <표 9>와 <표 10>은 각각 목차와 간행물 샘플데이터에 대한 구조분석 성공률을 나타낸다. 간행물 1단의 경우에는 비교적 간단한 구조의 문서이므로 모든 인식 방법의 구조분석 성공률 차이가 크지 않지만, 조금 더 복잡한 구조를 갖는 2단, 3단, 혼합 그

〈표 9〉 목차 문서이미지의 구조분석 성공률

구분	목차
제안하는 시스템	91.83
A사	86.24
B사	55.05

〈표 10〉 간행물 문서이미지의 구조분석 성공률

구분	간행물 1단	간행물 2단	간행물 혼합 (3단 포함)
제안하는 시스템	99.02	98.48	93.15
A사	99.51	80.81	63.01
B사	92.68	89.36	86.30

리고 목차 구조에 대해서는 타사 인식 제품의 구조분석 성공률이 현저히 떨어지고 있다. 반면에 본 논문에서 제안하는 시스템은 문서구조의 복잡도와 관계없이 일정하게 구조분석 성공률이 유지되는 것을 볼 수 있다.

6. 결 론

본 연구는 공공기관이 소장한 고문서 이미지데이터의 검색 및 열람 등의 활용성을 높이기 위한 전문검색서비스 구현 시 필수적인 문서인식시스템 성능을 안정적으로 유지하는 것을 목표로 한다. 이를 위해 소장한 문서이미지 데이터의 특징을 사전에 분석하고, 그 결과로부터 문서이미지 전처리 및 문서구조분석 기술을 개발하였고 소장자료에 특화된 지식베이스를 구축하여 문서인식 과정에서 활용토록 하였다. 구축한 지식베이스는 문서의 구조적, 논리적 특징을 표현하는 이미지내용DB, 문자 및 용어의 특징을 표현하는 문자모델DB과 용어사전DB로 각각 구성하였으며, 지식베이스 관리도구를 개발하여 정보를 생성, 관리하며 향후 다양한 형태의 문서이미지로의 확장을 가능하게 하였다. 고문서 자료의 다양한 인쇄 특징 및 소장기간이 오래되어 원본의 노후화에 의한 문제로부터 야기되는 문서인식 성능 저하를 해소하기 위해, 문서인식시스템은 본 연구의 결과인 이미지 분석 라이브러리 및 특화된 지식베이스를 활용하여 고품질의 안정적인 데이터를 생성할 수 있도록 구축하였다. 국가기록원이 소장하고 있는 문서이미지데이터를 대상으로 국내외 상용 전문인식 제품과 제안하는 시스템을 정량적인 평가요소에 따라 실험하였으며, 제안하는 시스템의 문서분석 및 인식 성능이 상대적으로 높고 안정적으로 유지됨에 따라 향후 전문검색시스템 연계 시 보다 효율적일 것으로 기대한다.

참 고 문 헌

[1] 김두식, 김상엽, 이성환, "한글문서 분석 및 인식기술의 최근 연구동향", 전자공학회지, 제24권, 제9호, pp.1058-1070, 1997.

[2] 이준호, 이충식, 한선화, 김진형, "문자 인식에 의해 구축된 한글 문서 데이터베이스에 대한 정보 검색", 한국정보처리논문지, 제6권, 제4호, pp.833-840, 1999.

[3] 정규식, 권희웅, "내용기반의 인쇄체 영문 문서 영상 검색을 위한 특징기반 단어 검색", 한국정보과학논문지, 제26권, 제10호, pp.1204-1218, 1999.

[4] 오일석, 김수형, 유태웅, 곽희규, "문서 영상 처리 기술과 디지털 도서관", 한국정보과학회지, 제20권, 제8호, pp.24-34, 2002.

[5] E. A. Galloway and G. V. Michalek, "The Heinz Electronic Library Interactive Online System(HELIOS): Building a digital archive using imaging, OCR, and natural language processing technologies," The Public-Access Computer Systems Review, Vol.6, No.4, pp.6-18, 1995.

[6] K. Marukawa, T. Hu, H. Fujisawa and Y. Shima, "Document retrieval tolerating character recognition errors-evaluation and application," Pattern Recognition, Vol.30, No.8, pp.1361-1371, 1997.

[7] D. Doermann, "The indexing and retrieval of document images: A survey," Computer Vision and Image Understanding, Vol.70, No.3, pp.287-298, 1998.

[8] Digital Heritage Publishing Ltd., "The electronic version of Siku Quanshu," http://www.skqs.com.

[9] T. Keaton, H. Greenspan and R. Goodman, "Keyword spotting for cursive document retrieval," Proceedings of the workshop on Document Image Analysis, pp.74-81, 1997.

[10] M. Droettboom, I. Fujinaga, K. MacMilan, G. S. Chouhury, T. DiLauro, M. Patton and T. Anderson, "Using the Gamera framework for the recognition of cultural heritage materials," Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital Libraries, pp.11-17, 2002.

[11] S. Hara, "OCR for CJK classical texts preliminary examination," Proc. Pacific Neighborhood Consortium(PNC) Annual Meeting, Taipei, Taiwan, pp.11-17, 2000.

[12] M. Kojima, Y. Kawazoe and M. Kimura, "Automatic Tibetan Script Recognition by Computer," Proceeding of the 7th Seminar of the International Association for Tibetan Studies, Graz, 1995, edited by Ernst Steinkellner, Vol.1, pp.527-533, 1997.

[13] T. Shih, "Transformation of palace archives of Ming and Ching Dynasties onto CD-ROM and Internet," Proc. Pacific Neighborhood Consortium(PNC) Annual Meeting, Taipei, Taiwan, 2000.

[14] Minsoo Kim, Kyutae Cho, Heegue Kwag, Jin Hyung Kim, "Segmentation Method of Handwritten Characters for Digitalizing Korean Historical Documents," The 6th international Conference on Document Analysis Systems, Florence, pp.114-124, 2004.

[15] M. S. Kim, S. Ryu, K. T. Cho, T. H. Rhee, H. I. Choi, J. H. Kim, "Recognition-based Digitalization of Korean Historical Archives," Asian Information Retrieval Symposium(AIRS2004), Beijing, China, pp.186-189, 2004.

[16] J. Beusekom, D. Keysers, F. Shafait, T. M. Breuel, "Example-Based Logical labeling of Document Title Page Images," 2007, Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), pp.919-923.

[17] F. Shafait, J. Beusekom, D. Keysers, T. M. Breuel, "Structural Mixtures for Statistical layout Analysis," 2008, Proc. 8th Int. Workshop on Document Analysis Systems (DAS) Accepted for publication.

곽 희 규

e-mail : hkkwag@inzisoft.com

1996년 전남대학교 전산학과(학사)
1998년 전남대학교 전산학과(이학석사)
2001년 전남대학교 전산학과(이학박사)
2001년~2002년 KAIST 박사후연구원 (PostDoc)
2002년~2004년 (주)동방라이텍 연구소장
2004년~2006년 (주)유씨티코리아 연구소장
2006년~현 재 (주)인지소프트 책임연구원
관심분야 : 영상처리, 패턴인식, 정보검색, 멀티미디어 등