문서 영상 처리 기술과 디지털 도서관

전북대학교 오일석*
전남대학교 김수형*
서해대학교 유태웅**
한국과학기술원 곽희규

1. 서 론

인류의 지식은 문자와 문서라는 형태를 통해 오랜 세월 축적되어 왔다. 최근 컴퓨터 기술과 인터넷의 발전으로 인해 문서의 제작과 제공 패러다임이 아날로그 방식에서 디지털 방식으로 급속히 바뀌고 있다. 이러한 과정에서 기존 컨텐츠는 아날로그 형태로 존재하나 새로 발생하는 컨텐츠는 디지털 형태를 갖게 되므로 존재 방식에 큰 간격이 발생하여 여러 문제를 야기할 수 있다. 앞으로 문서에의 접근과 검색 방식이 급속도로 디지털로 전환될 것이므로 기존 아날로그 컨텐츠에 대한 접근 고리를 상실할 수도 있다.

1.1 컨텐츠 패러다임의 변화

1990년대 들어 WWW이 등장함으로써 컨텐츠 제공 패러다임에 근본적인 변화를 가져왔다. 기존에는 자신이 가진 정보를 외부로부터 폐쇄하고 적절한 제어 하에 외부로 제공한 반면, 이제는 정보를 인터넷에 올려놓고 누구나 다운로드 할 수 있는 서비스가 주류를 이루는 개방성 패러다임으로 전환되었다[1, 2]. 또한 PC와 인터넷의 대중화로 문서 작성이 필기 위주에서 워드 프로세서로 바뀜에 따라 문서가 디지털 형태로 발생하는 비율이 급속도로 높아졌다.

그림 1에서는 컨텐츠를 2000년 기점으로 2000년-이전과 2000년-이후로 구분하고, 이들의 디지털 비율을 개념적으로 보여주고 있다. 또한 사용자들이 컨텐츠를 획득할 때 디지털 방식에 의존하는 비율도 같

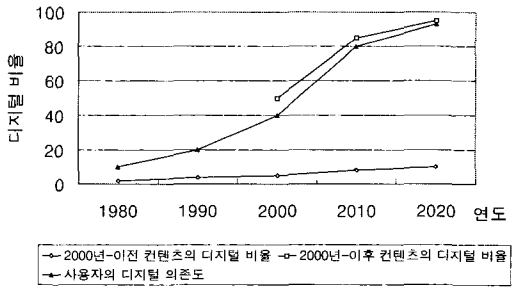


그림 1 시대적 디지털 격차를 설명하는 개념 그래프

이 보여준다. 2000년-이전 컨텐츠는 대부분 아날로그 형태로 작성되어 있으므로, 이를 스캔하여 디지털로 변환하고 검색이 가능한 형태로 가공해야 디지털 접근이 가능하다. 하지만 입력 비용이 엄청나기 때문에 극히 일부분만 변환되어 있고 앞으로도 획기적으로 늘어날 전망은 보이지 않는다. 물론 국가적인 디지털 도서관 사업 등이 수행되고 있어 디지털 비율이 소폭으로 증가할 것이지만, 기존 아날로그 컨텐츠 분량의 방대성과 배정된 예산의 한계를 감안하면 여전히 디지털 비율은 현저히 낮을 것이다. 반면 2000년-이후 컨텐츠 대부분은 발생 당시부터 컴퓨터 파일로 저장되므로 디지털 비율이 아주 높으며, 이들 파일은 여러 형태로 인터넷 서버에 올려져 사용자들에게 개방되고 그 비율은 꾸준히 증가할 것이다.

사용자의 자료 획득 방식을 크게 아날로그와 디지털로 구분해 보면, 아날로그 방식에서는 도서관을 방문하여 필요한 자료를 복사한다든지 하는 행위를 수반하는 반면, 디지털 방식에서는 자신의 컴퓨터에서 디지털 도서관 또는 e-book 서버 등에 접속한 후 관

* 종신회원
** 정 회 원

심 있는 키워드를 입력하여 원하는 자료를 검색하고 획득한다. 현재의 정보화 속도를 감안하면 사람들이 디지털 획득 방식으로 전면적으로 전환하는데는 그리 긴 시간이 걸리지 않을 것이다. 이를 달리 말하면 급속도로 아날로그 방식을 기피할 것이며, 굳이 도서관을 방문한다든지 종이 책에서 원하는 부분을 찾기 위한 시간과 이를 복사하는 수고를 지불하지 않으려 할 것이다. 이러한 상황이 굳어지면 사람들은 디지털 방식으로만 자료를 검색하고 획득하는 반면, 2000년-이전 컨텐츠는 극히 일부분만 디지털화 되어 있으므로 나머지 아날로그 형태로 남아있는 컨텐츠는 사람들로부터 단절되는 상황을 예측할 수 있고, 이러한 현상을 "시대적 디지털 격차"(chronological digital divide)라 부를 수 있다.

인류는 그 당시 발생한 지식을 문서 형태로 작성하여 후손에 전달하고, 후손은 이를 토대로 새로운 지식을 창출하여 인류 발전을 이룩하여 가는 사이클을 가지고 있다. 따라서 조상이 작성한 컨텐츠를 검색하고 획득하여 읽는 행위가 사이클의 핵심 요소로 포함되어 있다. 하지만 위에서 설명한 바와 같이 사람들의 디지털 획득 방식이 습관으로 굳어져 버리면 컨텐츠의 존재 형태와 접근하는 방식의 차이로 인해 2000년-이전 컨텐츠에의 접근 고리를 상실하고 이로 인해 지식의 시대적 단절이 발생할 것이다. 물론 컨텐츠의 나이에 따라 영향력(impact factor)이 줄어들어 절대적인 비율 차이보다는 적은 심각성으로 단절이 나타나겠지만 그 심각성은 어느 상황에서도 무시할 수 없을 것이다.

우리나라에서는 이미 1900년대에 언어가 한문에서 한글로 전환하는 과정에서 이전 컨텐츠에 대한 접근 고리를 상실한 역사가 있다. 이에 따라 우리 선조의 사상, 철학, 역사 등에 대한 전수가 제대로 이루어지지 않아 급속도로 서구화된 경험이 있다(그림 2 참조).

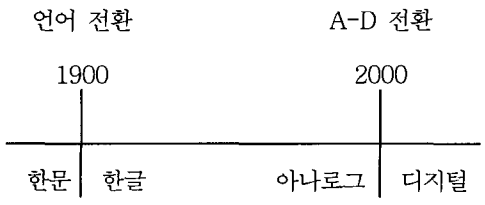


그림 2 지식의 시대적 단절 현상

1.2 문서 영상 처리 기술의 역할

디지털 도서관이 제공하는 문서 서비스에 대한 사용자 요구는 빠른 검색, 전문(full-text)을 대상으로 하고 적합도 랭킹(relevance ranking)이 가능한 검색, 문서 영상의 브라우징 편리성, 디스플레이/프린트 품질, 적은 파일 크기 등일 것이다.

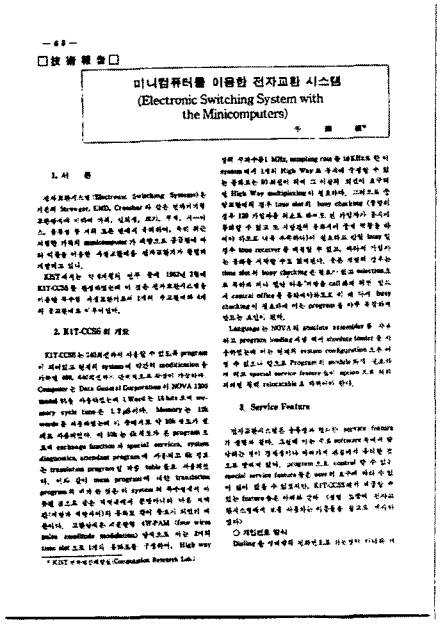
현재 우리나라에서 추진하고 있는 국가 전자도서관 사업에서 종이로 존재하는 문서의 디지털 작업 대부분은 스캔하여 저장하는 데 국한하고 있다. 1997~1998년에 걸쳐 수행한 국가 전자도서관 사업은 총 200만 쪽 정도를 입력해 놓았다[3]. 이미 입력되어 서비스되고 있는 문서는 우리가 보유하고 있는 전체 종이 문서의 극히 일부분이다. 정부기관으로서 특허청(특허문서 등), 법원(판결문 등), 지방 정부(조례 등), 대학 도서관의 석·박사 학위논문, 사기업의 기업문서 등 그 범위와 분량은 우리가 상상하는 것보다 훨씬 방대하다.

종이 문서로부터 스캔되어 디지털 도서관에 저장되어 있는 문서는 제목, 요약, 키워드만 텍스트로 저장되어 있고 전문(full-text)은 비트맵 영상 형태로 저장되어 있다. 따라서 제목으로 검색은 가능하나 전문을 대상으로 하는 검색은 불가능하다는 한계를 안고 있을 뿐만 아니라 적합도 랭킹 기능도 불가능하다. 일반 자연 영상을 위해 개발되어 있는 tiff 영상 포맷을 사용함으로써 영상 브라우징과 파일 크기 측면에서도 불리함을 안고 있을 뿐만 아니라 문서 영상에 대한 저작권 보호 장치도 마련되어 있지 않은 실정이다.

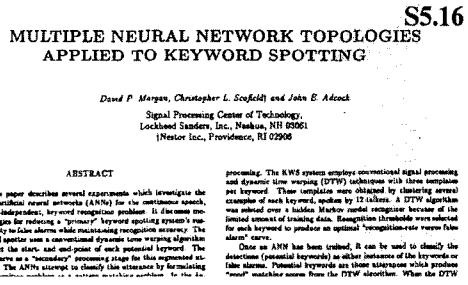
그림 3은 현재 운영되고 있는 디지털 도서관의 문서 영상 예이다. IEEE 디지털 도서관의 경우 예전에 발간된 Transactions, 그리고 대부분 Conference Proceedings가 비트맵 영상으로 저장되어 있다. 그림에서 나타나듯이 문서 영상은 발생 연도에 따른 원본 품질과 스캔 품질에 따라 다양성을 띠고 있다.

비트맵 문서 영상을 텍스트 정보로 바꾸기 위해서는 한정된 범위 내에서는 수 작업 입력을 사용할 수 있으나 대부분의 경우 비용과 시간 제약으로 자동 입력이 필수적이다. 이러한 필요에 따라 국내외에서 문서 영상 구조 분석(layout analysis), 인식(recognition), 그리고 이해(understanding)에 관한 연구가 수십 년간 이루어졌으며, 현재 특정 응용에서 상품화되어 성공적으로 사용되고 있는 사례가 여럿

(a) NDL(1945년 이전 신문을 대상으로 "김구"로 검색)



(b) KISTI 정보과학회 논문지



S5.16

MULTIPLE NEURAL NETWORK TOPOLOGIES APPLIED TO KEYWORD SPOTTING

David P. Morgan, Christopher L. Scofield and John E. Adcock

Signal Processing Center of Technology, Lockheed Sanders, Inc., Nashua, NH 03061 | Nestor Inc., Providence, RI 02906

ABSTRACT

(c) IEEE Explorer

그림 3 현재 운영되고 있는 디지털 도서관의 문서 영상 예제

있다. 디지털 도서관은 이러한 기술이 가장 효과적으로 적용될 수 있는 좋은 응용임에도 불구하고, 현재 단계에서는 디지털 도서관에 적용하여 서비스를 한 차원 끌어올리려는 노력은 부족한 것으로 보인다.

국내에서는 1990년대에 문자와 문서 영상 처리와 인식에 관한 연구가 활발히 수행되어 많은 기술이 여러 대학, 연구소, 그리고 기업에 축적되어 있다. 현재 PDA, 은행과 회사의 문서 관리, 국가의 고문서 입력 사업 등에서 이러한 기술이 부분적으로 활용되고 있는 사실이 이를 반증해 주고 있다. 따라서 디지털 도서관 응용에서 이러한 기술을 적극적으로 활용함으로써, 디지털 도서관에 저장되어 있는 컨텐츠에 생명력을 불어넣고 한 차원 높은 서비스로 끌어올릴 필요성이 있다.

1.3 논문 요약

2장에서는 문서 영상의 전처리에 대해 다룬다. 이 단계는 입력 장치로부터 받은 문서 영상에 대해 잡음 제거와 단어 또는 문자 단위 분할 등의 역할을 하며, 이 단계의 출력을 인식, 검색, 압축, 워터마킹 등에서 사용하므로 높은 성능의 처리가 매우 중요하다. 3장에서는 문서 영상의 검색 기능에 대해 다룬다. 디지털화 되었으나 아직 텍스트 코드로 변환되지 않은 문서 영상을 대상으로 질의 단어를 어떻게 검색할 지를 다룬다. OCR을 사용한 방법과 OCR을 사용하지 않는 방법을 모두 제시하며 두 방법의 장단점을 비교한다.

4장에서는 문서 영상의 특성을 고려한 문서 영상 표준 압축 방법에 대해 기술한다. 특히 문서 영상의 특성을 잘 이용하고 있는 DjVu 포맷을 중심으로 여러 가지 정보를 제공한다. 5장은 문서 영상 저작권 보호를 위한 워터마킹 기술을 소개한다. 문서 영상이 워터마킹 관점에서 일반 영상과 다른 점을 지적하며, 문서 영상 처리 프로그램의 필요성도 언급한다. 마지막으로 6장에서 결론을 제시한다.

2. 문서 영상 전처리

문서 영상의 인식 및 검색을 위해서는 영상을 인식의 대상이 되는 기본 단위로의 분할이 필수적이다. 문서 영상의 스캐닝 및 분석, 분할에 따른 모든 연산이 포함되는 영상 전처리 기술은 문서 영상을 처리(인식, 검색 등)하는 시스템의 성능을 좌우하는 매우 중요한 기술이다. 그럼에도 불구하고, 국내의 문서

영상 전처리에 대한 연구는 매우 소극적인 것이 사실이다. 이것은 문서 영상의 인식이나 검색에 비해 상대적으로 전처리 기술의 역할을 간과하고 있었던 것과 문서가 포함하는 요소들과 이들의 배치가 복잡하고 다양하다는 데도 그 이유를 찾을 수 있다. 또한, 적용 분야에 따른 지식(domain knowledge)이 요구됨으로써 범용의 전처리 기술을 구축하기가 쉽지 않다는 것이다[4, 5, 6].

고성능의 문서 영상 전처리 기술의 실현을 어렵게 하는 대표적인 장애요인에는 문서의 왜곡(종이 품질, 인쇄 상태, 보관 상태, 밑줄, 이물질 등), 기계적 왜곡(스캐닝, faxing 등에 의한 비틀림(skew), 잡영(noise) 등), 문서 구조의 복잡성(다단 편집, 텍스트와 비텍스트의 다양한 배치 등), 문서 내용의 복잡성(처리대상 언어, 비텍스트 부분에 대한 처리 요구) 등이 있다[4, 5, 6]. 따라서, 효율적인 문서 영상 전처리 기술은 다음과 같은 성질을 포함하도록 해야 한다.

- 처리 가능한 문서 종류에 대한 제약이 없어야 한다. 즉, 다단으로 복잡하게 배치된 잡지나 논문 문서 등의 처리와 표, 복잡한 모양의 그림 등을 갖는 임의의 문서에 대한 처리가 가능해야 한다. 이런 무제약의 성질을 만족시키지 못하면 문서 처리 시스템의 활용 범위가 제한될 수밖에 없다.
- 영상 입력 과정에서 어느 정도 기울어진 상태로 입력되는 문서의 처리가 가능해야 한다. 이것은 문서 인식의 신뢰성에 영향을 미치는 중요한 전처리 과정이다.
- 전체 문서 처리 속도의 향상을 위하여 영상 분할에 소요되는 처리 시간이 짧을수록 좋다.
- 문서 영상 분할과 동시에 다음 단계인 인식 과정에서 필요로 하는 각종 정보를 추출할 수 있으면 영상 인식 단계의 복잡도를 줄일 수 있다.

문서 영상의 분할은 크게 상향식(bottom-up)과 하향식(top-down) 접근 방법을 고려할 수 있다. 상향식 방법은 연결화소와 같은 개념을 이용하여 문자, 문자열, 문단, 문서와 같은 순서로 기본 추출단위로부터 보다 큰 복합 추출단위로 병합해 나가는 방식이다. 하향식 방법은 상향식과 반대로, 투영과 같은 방법을 이용하여 큰 추출단위로부터 작은 추출단위로 분리해 나가면서 영역을 분할하는 방식이다. 또한, 문서 영상 분할은 분할 단위를 자소 혹은 문자나 단어를 고려할 수 있는데, 이것은 인식 단위에 따라 다르다. 지금까지 대부분의 연구는 OCR 패키지의 활용을 위한 자소 혹은 문자 분리에만 집중되었고 정확한 단어 분리에 대한 체계적인 연구 발표는 이루어지지 않았다. 그나마 단어 단위의 분할을 전처리 단계로 고려하고 있는 몇몇 주제어 인식 연구에서도 최적의 단어 분할을 가정하고 있는 실정이다.

본 논문에서는 문서 영상의 전처리 모듈의 구축을 위해 반드시 필요한 네 가지 핵심 요소기술에 대해 언급하고자 한다(그림 4 참조).

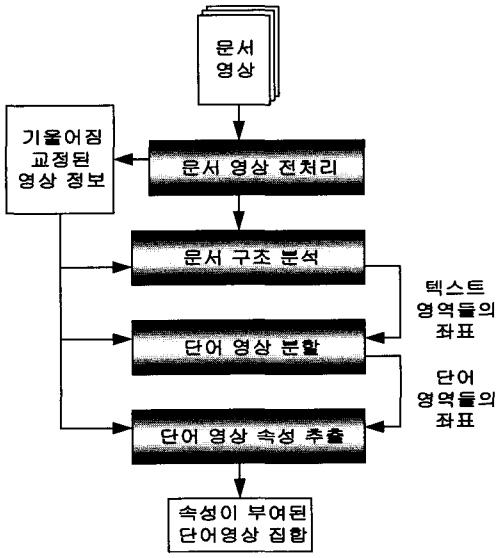


그림 4 문서영상 전처리를 위한 핵심 요소기술

2.1 기울기 교정

문서 영상의 전처리는 스캐닝 과정에서 문서를 잘못 놓거나 자동 급지 장치(ADF : Automatic Document Feeder)의 다양한 속도 변이에 의해 발생하는 문서 영상의 기울어짐을 추정하고 교정하는 과정을 말한다. 문서 영상의 기울어짐은 영상의 스캐닝 과정에서 흔히 발생하는 문제이며, 그림 5는 기울어진 문서 영상과 교정된 문서 영상의 예를 잘 나타내고 있다[6]. 기울어진 문서 영상은 문서 구조 분석 및 분할과 인식의 성능에 지대한 영향을 미치는 모듈로서, 처리 속도 및 정확도가 요구되는 알고리즘의 개발이 필요하다.

2.2 문서 구조 분석

문서 영상 전처리를 통해 문서 영상의 기울어짐이 교정된 후에 문서 영상을 의미있는 영역들로 분할하

그림 5 기울어진 영상과 교정된 영상의 예

게 되는데, 각 영역들은 텍스트 또는 그래픽 정보(그림, 테이블, 선 등)를 포함한다. 이것은 문서 영상으로부터 기하학적인 구조를 추출하여 문서를 다양한 영역의 집합으로 분할하는 단계와 각 영역의 특성을 조사하여 텍스트와 비텍스트 영역으로 분류하는 영역 분류 단계로 구성된다. 문서 영상의 기하학적인 구조를 추출하는 방법은 크게 상향식 접근 방법과 하향식 접근 방법으로 연구되어 왔는데, 분할의 정확도와 속도 측면에서 각각 장단점을 가지고 있다. 따라서 계산 비용과 분할 오류를 최소화하기 위해 두 접근 방법을 적절히 결합하여 적용하고, 다양한 해상도를 가지는 다단계 영상 정보를 생성하여 분할 검사에 저해상도와 고해상도 영상을 함께 사용하는 것이 효율적이다. 그림 6은 문서 영역 분할 후, 텍스트와 비텍스트 영역으로의 분류 결과를 잘 나타내고 있다[6].

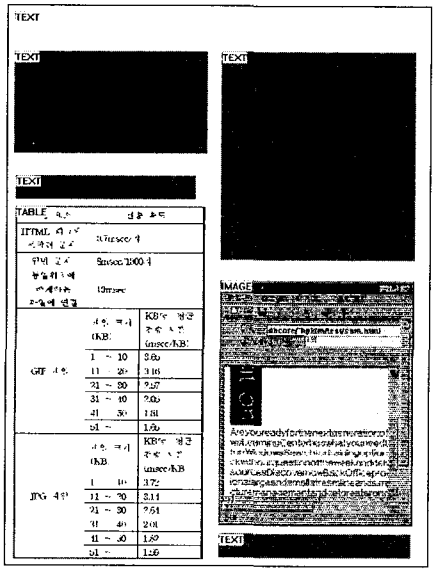


그림 6 문서 영상의 영역 분할과 분류의 결과

2.3 단어 영상 분할

문서 구조 분석을 통해 추출된 텍스트 영역은 인식의 대상인 단어 단위로 분할하게 되는데, 일반적으로 투영 프로파일(projection profile) 방법을 적용하거나 연결요소 분석(connected component analysis) 방법 등을 적용한다. 한편 두 방법은 처리 시간 및 정확도에 대한 장단점을 가지고 있기 때문에, 텍스트 영역에 대한 수평 투영 프로파일을 적용하여 텍스트 라인을 분할하고 각각의 텍스트 라인에 대한 연결요소 분석 방법을 적용하여 단어 단위 분할 지점을 찾을 수 있다면 매우 효율적이다. 또한, 텍스트 라인에는 다양한 특수기호(', ", ~, -, (,), [,], {, } 등)들이 존재하여 단어 분할에 대한 오류를 유발한다. 따라서 연결요소에서 추출한 휴리스틱(heuristic) 정보를 사용하여 특수기호들을 인식하고 단어 분할 지점으로 결정함으로써, 분할에 따른 정확도를 개선한다. 그림 7은 텍스트 영역에 존재하는 특수기호들을 인식함으로써 정확하게 단어 분할 지점을 찾는 결과를 잘 나타내고 있다[6].

노드의 확장 방법을 나타내는 d-배열의 초기 개체군을 생성한다. 2 단계에서는 해(solution)에 빠른 수렴을 위하여 초기 개체에 [10]에서 제안한 GroupSift-DTL 알고리즘을 적용한다. 3단계에서는 각 요소에 입력변수 연산자(Reproduction, PMX)를 적용한다. 그리고 나서 4단계로 입력변

그림 7 텍스트 영역의 단어 단위 분할 결과

2.4 단어 영상 속성 추출

문서 영상을 구성하는 텍스트 요소들은 다양한 인쇄상의 속성 정보들을 포함한다. 이러한 속성 정보들은 문서 인식뿐만 아니라 검색(retrieval) 및 요약(summary) 생성, 재생산(reproduction) 등의 다양한 분야에서 응용 가치를 찾을 수 있지만 실질적인 연구가 이루어지지 않았다. 최근에서야 비로소 문자가 가지고 있는 속성 정보의 유용성을 고려하여 문자의 속성을 인식(OFR: Optical Font Recognition)하는 연구가 활발히 진행되고 있고, OCR과의 결합을 통한 성능 향상을 꾀하고 있다.

텍스트 영역의 단어들이 포함하는 속성 정보들은 언어(script)의 종류, 문자의 크기(size), 스타일

(style), 문자 개수, 서체(typeface) 등을 고려할 수 있다. 이러한 단어 영상의 속성 정보는 문서 인식 및 검색 단계의 복잡도를 줄이거나 OCR 시스템의 인식률을 향상시킬 수 있다. 실제 그림 8과 같이 다섯 가지 속성들을 고려한 시스템에서 소수의 특징들과 계층적 분류기를 도입하면 우수한 성능의 속성 인식 모듈을 구성할 수 있다[6].

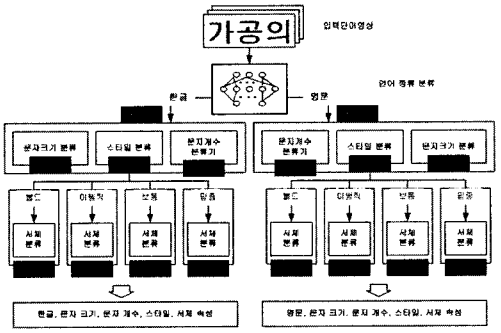


그림 8 단어 영상의 속성 인식 모듈 및 속성 분류 사례

3. 문서 영상 검색

문서 영상 검색은 웹 기반의 디지털 도서관 구축을 위한 주요한 요소 기술중의 하나이다. 디지털 도서관에서는 논문, 정기 간행물, 보고서 등의 일반 텍스트 문서와 영상으로 된 문서들의 검색서비스를 제공하는데, 이들 문서들은 포스트스크립트, 순수 텍스트, HTML 또는 영상형태로 저장되어 있다. 포스트스크립트, 순수 텍스트, HTML 형태로 저장된 자료들은 종래의 텍스트 기반의 색인 및 검색시스템을 그대로 적용하여 검색서비스를 제공할 수 있으나 영상형태로 저장되어 있는 경우에는 텍스트 형태로 완전히 변환하거나 또는 사람이 일일이 색인 작업을 하지 않는 한 종래의 색인 및 검색시스템을 적용할 수 없다[7].

문서 영상 분할과 인식 기술을 문서 영상의 자동 색인 및 검색에 응용하는 방법에는 두 가지 전략이 있을 수 있다. 첫 번째는 문자나 자소 단위의 분할 기술과 OCR 기술을 결합하여 문서 영상의 내용 전체를 텍스트 형태로 변환한 후, 형태소 분석 등의 과정을 거쳐 색인하고 검색하는 방법이다[8, 9]. 두 번째는 문서 영상을 띄어쓰기 단위인 단어 영상의 집합으로 분할한 후, 검색하려는 주제어와 분할된 단어 영상을 매칭하는 과정을 거쳐 문서 영상을 탐색하는 방식이다[10, 11]. 서베이 논문으로는 [12]와 [13]을 추천할 수 있다.

3.1 OCR 기반 검색

이 방법은 문서 영상에 OCR 소프트웨어를 적용하여 텍스트 변환을 시도한다. 성공적인 텍스트 변환이 보장되면 검색에 필요한 모든 정보가 얻어졌다고 말할 수 있다. 문자인식에 관련된 연구는 1960년대 초부터 꾸준히 진행되어 수작업 입력을 대체할 만큼의 인식 정확도를 갖는 제품들이 출시되고 있으나, 실제 응용분야에 적용하기에는 문제점이 없지 않다. OCR 기반의 문서 입력 방법에서 해결해야 할 몇 가지 문제점을 요약하면 다음과 같다.

- 낮은 인식률 : 현재까지 개발된 OCR 기술은 아직 사람의 정확도를 능가하지 못하고 있다. 한글의 경우 99% 이상의 성능을 보인다고 발표된 제품은 많지만 일반적 환경에서 95% 이상의 인식률을 보이는 제품이나 기술이 없다고 판단된다.
- 수작업에 의한 OCR 결과의 검증 : 인식 결과에 대해 별도의 검증 단계가 필요하게 되는데, 경우에 따라서는 사후 검증 단계가 문서를 처음부터 수작업으로 입력하는 기존의 방법보다 비효율적이 될 수 있다. 숙련된 전문인력이 검증하는 속도는 전문 타이피스트가 입력하는 속도의 1/5 정도라는 실험 결과는 인식률이 80%에 미치지 못하는 문서에 대해서는 OCR 기술을 이용한 입력이 수작업 입력보다 비효율적이라는 것을 의미한다.

문서 검색에 OCR 기술을 사용하는 데 있어 핵심적인 기술은 OCR 에러가 포함된 파일에서 에러를 견디는 알고리즘을 개발하는 것이다. 예를 들어 10%의 에러를 포함한 텍스트 파일에서 95% 이상의 정확도(precision)와 재현률(recall)을 제공하는 알고리즘을 설계하는 것을 목적으로 할 수 있다.

현재 운영되고 있는 디지털 도서관이 이미 방대한 양의 문서 영상을 보유하고 있고, OCR 소프트웨어도 사용 가능하므로 이러한 목적을 달성하는데 그다지 많은 비용이 들지 않을 것이다.

OCR 에러를 견디는 검색 방법으로는 영문 문서 내에서 오인식 빈도 정보를 갖는 혼동 행렬을 이용한 검색 방법이 제안되었다[14]. 한글 OCR 소프트웨어의 혼동 행렬을 기반으로 설계된 알고리즘과 실험 결

과를 제시한 사례도 있다[9]. Marukawa[15] 등은 일본어에 대해 혼동 행렬을 이용한 질의어 확장 검색 방법과 인식 과정에 있어 모호성을 가지는 문자에 대해 다중 후보 문자를 출력하여 검색하는 방법을 제안하고 있다. 한글 문서에 대해 형태소 단위 색인법과 2-gram 기반 색인법을 이용한 연구도 있다[16].

3.2 단어 매칭에 의한 검색

그림 3(a)의 문서 영상과 같이 품질이 아주 낮은 경우, OCR 인식률은 매우 낮을 것이고 따라서 거의 쓸모없는 성능을 보일 수 있다. OCR의 대안으로서 영상-기반 단어 매칭(word matching) 기법을 사용할 수 있다. 이 방법은 문서 영상에서 분할된 단어를 대상으로 적절한 특징을 미리 추출하여 저장해 놓는다. 사용자로부터 질의 단어가 들어오면 같은 종류의 특징을 추출하고 문서 안의 단어들과 매칭을 시도하여 검색을 수행한다. 이 방법을 주제어 스파팅(keyword spotting)이라고도 한다. 이와 같은 알고리즘은 다음 두 가지 요구사항을 모두 만족해야 한다.

- 디지털 도서관 데이터베이스에는 아주 방대한 양의 단어 영상들이 있기 때문에 매칭 알고리즘이 매우 빨라야 한다.
- 알고리즘은 재현율(recall)과 정확도(precision) 관점에서 신뢰성 높은 성능을 제공해야 한다.

Xerox PARC의 DID(Document Image Decoding) 팀은, NSF 주관으로 미국 6개 대학이 주도한 디지털 도서관 선도 사업[2]의 하나인 UC Berkeley의 Environmental Digital Library 프로젝트에 합류하여 DID 기술의 부분으로 단어 매칭에 의한 검색을 시도하였다[17]. 또한 상태가 좋지 않은 인쇄 문서들을 위한 HMM을 이용한 방법[18], 단순한 모양 코드(shape code)를 사용하는 기법[19], 가설-검증 기법을 사용한 기법[20], 단어 영상 특징과 N-gram을 결합한 기법[21] 등이 있다.

단어 매칭을 통한 문서 영상 검색시스템의 괄목할 만한 사례는 미국 SRI사에서 개발한 SCRIBBLE 시스템이다[11]. 이 시스템은 단어 영상 자체를 분할하여 문자인식을 적용하는 방식 대신에 문자나 단어 영상의 외형적인 특징을 사용하여 인식하는 방식이다. 단어 영상은 개별적인 문자 영상보다 많은 정보를 포함하고 있기 때문에, 문서 영상의 품질이 낮은 경우에 보다 높은 인식 정확도를 유지할 수 있고 처리 속도가 빠르다. 실제로 OCR에 의한 영상 검색과의 비교를 통해 이러한 사실을 입증하였는데, 문서 영상의 품질이 저조한 경우에 상용 OCR 패키지에 비해 6~10% 정도 우수하고 2~20배정도 빠른 처리 속도를 보였다고 한다.

4. 문서 영상 압축 포맷

현재 자연 영상을 대상으로 개발되어 널리 사용되고 있는 압축 표준으로 pcx, jpeg, gif, tiff 등이 있다. 이들은 압축 알고리즘을 채택하여 파일 크기를 줄이고 있으며, 각각 고유한 특성과 장단점을 지니고 있다. 손실 압축 기법을 사용하고 있는 jpeg의 경우 높은 압축률에 비해 비교적 좋은 품질의 영상을 보장하므로 web에서 널리 사용되고 있다. 국내 디지털 도서관의 경우 tiff를 많이 사용하고 있다. 하지만 이러한 표준을 문서 영상에 적용할 수는 있지만, 문서 영상이 갖는 고유한 특성을 이용하지 못하므로 여러 측면에서 비효율성이 나타난다.

문서 영상은 다음과 같은 고유한 특성을 가지므로 이를 충분히 이용하여 압축 알고리즘과 영상 포맷을 설계하는 것이 바람직하다.

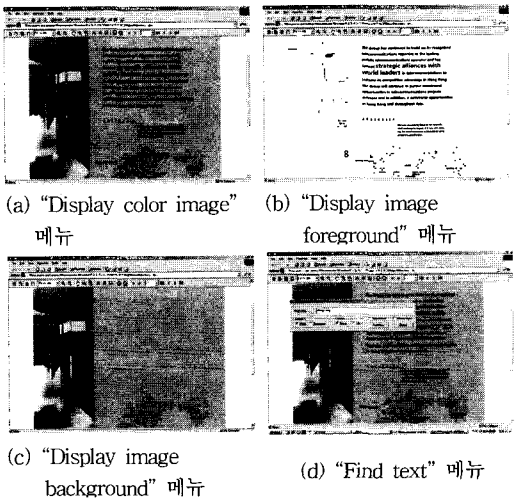
- 쪽/블럭/줄/단어/문자의 계층적 구조를 갖는다.
- 배경(background)과 전경(foreground)으로 구분할 수 있으며, 배경이 차지하는 비율이 높다.
- 전경의 대부분은 문자와 표로 구성되며, 문자는 길다란 획이 가로, 세로, 또는 대각선으로 규칙적으로 배열된다.

이러한 특성을 충분히 이용하면 디지털 도서관 환경에 아주 효과적인 여러 장점을 얻을 수 있다. 첫째는 높은 압축률이다. 배경은 비교적 균일한 색상과 명암을 가진 저주파 성분이 많다는 특성을 이용하여 높은 압축률을 얻을 수 있고, 전경의 대부분은 흰 영역이므로 또한 높은 압축률이 가능하다. 또한 영상을 브라우징할 때 전경과 배경을 따로 처리하여 고속으로 상하좌우로 이동할 수 있고 주밍(zooming)도 고속으로 처리할 수 있다. 전경에 나타나는 텍스트를 OCR 소프트웨어를 이용하여 인식하여 텍스트 검색도 가능하게 할 수 있다.

문서 영상 처리 연구에서는 이러한 아이디어를 실현하는데 필요한 기술을 연구해 왔다 [22]. 필요한 기술은 문서 영상 전처리, 전경/배경 분리, 표/그림/텍스트 분리, 텍스트 인식 등이다. 이러한 기술에 대한

연구는 국내외에서 오랫동안 수행되어 왔고, 아이디어 실현에 필요한 정도의 기술 수준이 되어 있다고 평가할 수 있다. 단지 이들 기술을 통합하여 하나의 문서 영상 포맷을 제정하고 구현하는 노력이 필요한 것이다.

여기에서는 상업화 단계에 접어든 사례 하나를 소개한다. 1996년 AT&T에서 개발된 DjVu는 문서 영상을 위해 설계된 압축 포맷이다[23, 24]. 위에서 말한 문서 영상의 특성을 충분히 이용하고 있으므로 위에 제시한 모든 장점을 갖고 있다. AT&T는 오랫동안 문서 영상 처리와 인식에 관한 연구를 수행해 왔는데, 이러한 꾸준한 투자와 노력이 하나의 세계적인 성과로 이어졌으며 이를 통해 문서 영상 포맷에서 선도 역할을 수행할 수 있게 되었다. 그림 9는 DjVu 브라우저 화면인데, 문서 영상의 전경/배경 분리, 텍스트 검색 기능 등을 볼 수 있다. 이 예제 영상의 경우 110쪽 full-color를 갖는 어느 기업의 연차 보고서인데, DjVu 포맷으로 2.87M 바이트인데 pdf는 147M 바이트로서 DjVu의 높은 압축률을 엿볼 수 있다.



(a) "Display color image" 메뉴 (b) "Display image foreground" 메뉴

(c) "Display image background" 메뉴 (d) "Find text" 메뉴

그림 9 DjVu 브라우저와 예제 영상

DjVu에 대해서는 AT&T의 공식 사이트를 참조하면 된다[25]. 특히 브라우저 프로그램의 소스 코드를 공개하고 있다. 또한 DjVu 관련 자료를 모아 디지털 도서관을 운영하고 있고 벤치마킹 자료를 제공하고 있으므로, 디지털 도서관을 구축하고자 할 때 문서 영상 압축 포맷 채택을 위한 기초 자료로 활용할 수 있을 것이다.

문서 영상의 특성을 이용한 압축 포맷 하나를 소개하였다. 현재 다른 연구진이 또 다른 포맷을 개발하고 있는지에 대한 정보는 없다. 단지 문서 영상 처리 기술과 압축 기술이 발전하고 있기 때문에 경쟁적 포맷이 출현할 가능성이 없진 않다.

국내 디지털 도서관이 보유하고 있는 문서 영상에 대해 DjVu가 압축률과 영상의 품질 면에서 어느 정도 성능을 보이는지 벤치마킹해 볼 필요가 있다. 또한 한글에 대한 구현이 제공되지 않으므로 한글 구현에 대한 노력이 필요하다.

5. 워터마킹을 통한 문서 영상 저작권 보호

디지털 도서관에 저장되어 있는 문서는 국가적인 공식 문서로서 가치가 높거나 또는 상업적인 가치가 높기 때문에 적절한 저작권 보호 장치가 마련되어야 한다. 최근 들어 멀티미디어 자료의 저작권 보호 목적으로 워터마킹에 대한 연구가 활발히 이루어지고 있는데, 자연 영상과 비디오 영상에 치우쳐 있다[26]. 상대적으로 문서 영상에 대한 연구는 적은 편인데 디지털 도서관 응용에서는 문서 영상을 중요하게 다룰 필요성이 높다.

문서 영상은 일반 영상과 다른 고유한 특성을 가지므로 이를 충분히 이용하여 워터마킹 알고리즘을 설계하는 것이 바람직하다. 워터마킹 기법은 2차원 영상 배열에 직접 신호를 삽입하는 공간-영역(spatial-domain) 방식과 영상을 변환한 후(DCT, wavelet 변환 등) 변환 공간에 신호를 삽입하는 변환 영역(transform-domain) 방식이 있다. 각 방식은 나름대로의 장점과 단점을 가지는데, 문서 영상의 경우 문서 영상의 특성을 활용하기 위해 공간 영역 알고리즘이 주류를 이룬다[27].

문서 영상을 위한 워터마킹 기법은 조작하는 요소(primitive)가 무엇이냐에 따라 세 가지로 구분할 수 있다.

- 화소 수준(pixel-level): 화소의 값을 변경
- 특징 수준(feature-level): 문자를 구성하는 획 특징을 추출한 후 획의 모양을 변경
- 문자 수준(character-level): 문자, 단어, 또는 줄을 약간 이동

숨길 수 있는 정보 양은 화소 수준이 가장 많으며, 특징, 문자 순으로 줄어든다. 반면에 강인성은 문자

수준이 가장 강하며, 특징, 화소 순으로 약해진다. Brassil 등은 문자 수준 알고리즘을 제시하였는데, 줄 이동(line-shift), 단어 이동(word-shift), 또는 문자 이동(character-shift) 방식을 사용하였다[28]. 이들은 복사, 팩스, 또는 프린트 등이 여러 번 반복된 문서에서 강인하게 워터마크를 탐지하기 위해 정교한 문서 분석 기술이 필요하다. 예를 들어 단어 이동에서는 한 줄에서 한 단어를 좌우로 1~2 화소 이동시켜 두는데, 문서 분석 모듈이 처리과정에서 1화소 이내의 정확도를 보장하지 못하면 이러한 이동 정보를 탐지하지 못한다.

문자 또는 단어의 명도를 조절하여 정보를 숨기는 방법도 제시되었다[29]. 이 방법에서는 문서 영상을 단어 단위로 분할한 후, 정보 인코딩 규칙에 따라 단어를 선택하여 해당 단어 영역의 명도를 조절한다. Amano는 특징 수준 알고리즘을 제시하였다[30]. 이 방법에서는 문서 영상을 문자 단위까지 분할한 후, 각 문자에 대해 획 추출을 수행한다. 정보 삽입 규칙에 따라 특정 획을 선택하고 make-fat 또는 make-thin 연산을 적용하여 획의 굵기를 조작한다.

문서 영상의 워터마킹은 정교한 문서 구조 분석과 단어 (또는 문자와 획) 분할 프로그램을 필요로 한다. 따라서 문서 영상 처리 기술과 연계하여 워터마킹 알고리즘을 설계할 필요가 있다. 다음 표는 위에 제시한 세 가지 방법이 문서 영상 처리 기술을 필요로 하는 정도를 보이고 있다.

표 1 워터마킹 기법에 따른 문서영상 처리 기술 필요 정도

문서 영상처리 연산	화소 수준	특징 수준	문자 수준		
			줄 이동	단어 이동	문자 이동
블 록 분 할	o	o	o	o	o
줄 분 할	x	o	o	o	o
단 어 분 할	x	o	x	o	o
문 자 분 할	x	o	x	x	o
획 추 출	x	o	x	x	x

(o : 필요함, x : 필요 없음)

대표적인 응용 사례 하나를 제시한다. IEEE Communications Society는 발간하는 잡지들을 웹을 통해 배포할 계획을 수립하고(이 사업의 일부분 SEPTEMBER라 부름), 1995년 10월 JSAC(Journal on Selected Areas Communications)에 대해 시범 서비스를 시작하였다[31]. 이 서비스는 IEEE이 디지털 도서관 서비스를 준비하며 추진한 계획으로 보이며, 저널을 인터넷을 통해 배부하는 첫 시도라는데 의미가 있고 워터마크를 이용하여 인터넷에서 문서 영상에 대한 저작권을 보호한 최초의 시도로서 큰 의미를 갖는다.

이 서비스에서 워터마킹은 크게 두 가지 목적을 갖는다. 첫째는 독자가 받은 논문에 대해 값어치를 높게 평가하도록 하는 것이고, 두 번째는 불법 재배포하려는 의욕을 상실시키는 것이다. 사용자는 등록을 해야 하고, 등록하면 곧바로 데이터베이스에 고유한 사용자 코드가 부여되고 코드북에 저장된다. 사용자가 논문을 다운로드 할 때 그의 고유한 코드를 비밀키로 하여 워터마크 신호가 삽입되고 워터마킹된 문서가 전달된다. 나중에 불법 재배포 문제가 발생하였을 때 워터마크 신호를 탐지하고 사용자 코드를 디코딩하여 어느 사용자에게 배포된 문서인지 파악할 수 있다.

Mintzer 등은 워터마킹 기술을 이용하여 디지털 도서관 컨텐츠의 저작권 보호 강도를 높이는 방안을 제시하고 있다[32]. 그들은 워터마킹을 불법 복제를 어렵게 만들기 위한 세 가지 기술(encryption, digital signature, digital watermarking) 중의 하나로 바라보고 있다. 즉 워터마킹이 디지털 도서관 컨텐츠의 저작권 보호를 위한 유일한 방법이라기보다는 저작권 보호 강도를 한 단계 높이는 하나의 방법인 것이다

6. 결 론

디지털 도서관은 ACM이나 IEEE과 같은 학술 단체와 상업적 출판사 등이 이미 경쟁적으로 서비스를 제공하고 있고, 국내에서도 KESLI 컨소시엄을 통해 대부분 대학이 이러한 서비스에 접근이 가능하다[33]. 이러한 서비스에 익숙한 사용자는 아날로그 컨텐츠에 대한 접근 고리를 머지않아 상실할 것이 분명해 보인다. 인류의 지식은 시간이 지남에 따라 조금씩 점진적으로 축적되는 연속성이 있다. 기술 발전에 따라 지식을 저장하는 매체와 이를 접근하는 방식이 급속히 아날로그에서 디지털로 전환되었다고 이러한 연속성이 효력을 잃는 것은 분명 아니다.

따라서 기존의 아날로그 문서를 디지털화하고 적

절한 접근 방법을 개발하여 현재 발생하고 있는 컨텐츠와의 차이를 줄이는 작업은 매우 중요하다 할 수 있다. 이러한 의미에서 문서 영상 처리 기술의 중요성을 강조하고 문서 영상 처리 기술이 이러한 간격을 줄이는데 어떻게 기여할 수 있는지를 살펴보았다.

현재 국내에 축적되어 있는 문서 영상 처리 기술을 디지털 도서관 서비스를 한 차원 끌어올리는데 어떻게 활용할 것인지를 진지하게 생각해보아야 할 것이다.

참고문헌

[1] CACM, Special issue on Digital Libraries, Vol.38, No.4, April 1995.

[2] IEEE Computers, Special issue on Digital Library Iinitiative, Vol.29, No.5, May 1996.

[3] 국가 전자 도서관, http://www.dlibrary.go.kr/.

[4] 김병기, 이창숙, "연결요소를 이용한 실제적 문서 영상 분할", 영상처리 및 이해에 관한 워크샵, pp.415-420, 1999.

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

[6] 곽희규, 문서 영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구, 전남대학교 박사학위논문, 2001.

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

[8] 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 6, No.4, pp.6-18, 1995.

[9] 안재철, OCR 소프트웨어를 이용한 한글 문서 검색 시스템, 전북대학교 석사학위 논문, 2002년 2월.

[10] J.L. DeCurtins and E.C. Chen, "Keyword spotting via word shape recognition," Proc. SPIE Document Recognition II, pp.270-277, 1995.

[11] "SCRIBBLE: SRI's keyword spotting system," http://www.erg.sri.com/projects/scribble, 1998.

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

[13] M. Mitra and B.B. Chaudhuri, "Information retrieval from documents: a survey," Information Retrieval, pp.141-163, 2000.

[14] M. Ohta, A. Takasu, and J. Adachi, "Retrieval methods for English-text with missrecognizied OCR characters," Proceedings of 4th International Conference on Document Analysis and Recognition, Vol.2. pp.950-955, 1997.

[15] 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.

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

[17] G.E. Kopec, "Document image decoding in the Berkeley digital library," Proceedings of ICIP, pp.769-772, 1996.

[18] S.-S. Kuo and O.E. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models," IEEE Trans. on Pattern analysis and Machine Intelligence, Vol.16, No.8, pp.842-848, August 1994.

[19] A.L. Spitz, "Shape-based word recognition," International Journal on Document Analysis and Recognition, Vol.1, No.4, pp.178-190, May 1999.

[20] J. Zhu, T. Hong, and J.J. Hull, "Image-based keyword recognition in Oriental language document images," Pattern Recognition, Vol.30, No.8, pp.1293-1300, 1997.

[21] C.L. Tan, et al., "Imaged document text retrieval without OCR," IEEE Tr. PAMI, Vol.24, No.6, pp.838-844, 2002.

[22] H. Bunke, P.S.P. Wang, and H.S. Baird,

Document Image Analysis, World Scientific, 1994.

[23] L. Bottou, P. Haffner, and Y. LeCun, "Conversion of Digital Documents to Multilayer Raster Formats," Proceedings of the International Conference on Document Analysis and Recognition, 2001.

[24] Y.L. Cun, L. Bottou, P. Haffner, and J. Triggs, "Overview of the DjVu Document Technology," Proc. SDIUT'01, Symposium on Document Image Understanding Technologies, Columbia MD, pp.119-122, April 2001.

[25] AT&T의 DjVu 공식 사이트, http://www.djvu.att.com/.

[26] F. Hartung and M. Kutter, "Multimedia watermarking techniques," Proceedings of the IEEE, Vol.87, No.7, pp.1079-1107, July 1999.

[27] 김영원, 문경애, 오일석, "텍스트 문서 영상의 화소 수준 워터마킹 알고리즘", 제14회 영상 처리 및 이해에 관한 워크샵, 제주 롯데, pp.31-36, 2002.

[28] J.T. Brassil, S. Low, and N.F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," Proceedings of the IEEE, Vol.87, No.7, pp.1181-1196, July 1999.

[29] A. Bhattacharjya and H. Ancin, "Data embedding in text for a copier system," Proceedings of the ICIP, Vol.2, pp.245-249, 1999.

[30] T. Amano and D. Misaki, "A feature calibration method for watermarking of document images," Proceedings of the ICDAR, pp.91-94, 1999.

[31] J. Brassil, et al., "SEPTEMBER: secure electronic publishing trial," IEEE Communications Magazine, pp.48-55, May 1996.

[32] F. Mintzer, J. Lotspiech, and N. Morimoto, "Safeguarding digital library contents and users: digital watermarking," D-Lib Magazine, December 1997(http://www.dlib.org/에서 얻을 수 있음).

[33] 국가과학기술 전자도서관, http://www.ndsl.or. kr/.

오 일 석

1984 서울대 컴퓨터공학과 학사
1992 KAIST 전산학과 박사
1992~현재 전북대학교 교수
관심분야:문서 영상 처리, 패턴인식, 유전 알고리즘의 패턴인식 응용
E-mail:isoh@moak.chonbuk.ac.kr

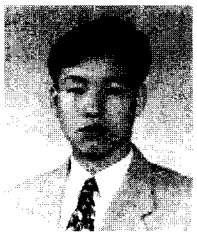
김 수 형

1986 서울대학교 컴퓨터공학과 학사
1988 KAIST 전산학과 석사
1993 KAIST 전산학과 박사
1993~1996 삼성전자 멀티미디어연구소 선임연구원
1997~현재 전남대학교 컴퓨터정보학부 부교수
관심분야:패턴인식, 문서영상처리, 필적 및 서명감정, WBI
E-mail:shkim@chonnam.chonnam.ac.kr

유 태 웅

1991 전북대학교 수학과 학사
1993 전북대학교 전산통계학과 석사
1998 전북대학교 전산통계학과 박사
1999~현재 서해대학 컴퓨터정보기술계열 조교수
관심분야:컴퓨터비젼, 문서영상처리, 패턴인식
E-mail:twyoo@sohae.ac.kr

곽 희 규



1996 전남대학교 전산학과 학사
1998 전남대학교 전산통계학과 석사
2001 전남대학교 전산통계학과 박사
2001~현재 한국과학기술원 박사후 연구원
관심분야:패턴인식, 영상처리, 필기한자 인식
E-mail:hkkwag@ai.kaist.ac.kr