문서 영상의 정교한 기하적 구조 분석을 위한 지식베이스 시스템

(A Knowledge-based System for Analyzing Sophisticated Geometric Structure of Document Images)

이 경 호 † 최 윤 철 †† 조 성 배 ††
(Kyong-Ho Lee) (Yoon-Chul Choy) (Sung-Bae Cho)

요 약 문서 영상으로부터 논리적인 구성 요소를 추출하여 전자 문서를 생성하기 위해서는 정교한 수준의 기하적인 구조 분석이 선행되어야 한다. 본 논문은 과학기술 논문을 대상으로 정교한 수준의 기하적인 구조 분석을 지원하기 위하여 지식베이스에 기반한 방법을 제안한다. 제안된 지식베이스는 과학기술 논문 유형이 공통적으로 갖는 기하적인 특성은 물론이고 출판물 특유의 특성에 대한 지식을 규칙 형태로 표현한다. 제안된 방법은 상향식과 하향식의 복합 기법을 사용하며 영역 분할과 식별의 두 단계로 구성된다. 일반적으로 영역 분할에 의하여 분할된 영역과 레이아웃을 구성하는 복합 객체사이에는 일-대-일의 대응 관계가 존재하지 않는다. 따라서 제안된 방법은 분할된 영역을 추가로 분할하거나 통합하면서 이미지, 드로잉, 그리고 테이블 등의 비 텍스트 객체는 물론이고 텍스트 라인이나 수식과 같은 텍스트 객체를 식별한다. 제안된 방법의 성능을 평가하기 위하여 IEEE Transactions on Pattern Analysis and Machine Intelligence로부터 스캐닝한 372개의 논문 영상으로 실험한 결과, 제안된 방법은 99% 이상의 실험 영상에 대한 기하적인 구조 분석에 성공하여 기존 방법에 비해 정교한 수준의 성능을 보였다.

Abstract Sophisticated geometric structure analysis must be preceded to create electronic document from logical components extracted from document image. This paper presents a knowledge-based method for sophisticated geometric structure analysis of technical journal pages. The proposed knowledge base encodes geometric characteristics that are not only common in technical journals but also publication-specific in the form of rules. The method takes the hybrid of top down and bottom-up techniques and consists of two phases' region segmentation and identification. Generally, the result of segmentation process does not have a one-to-one matching with composite layout components. Therefore, the proposed method identifies non-text objects such as image, drawing and table, as well as text objects such as text line and equation by splitting or grouping segmented regions into composite layout components. Experimental results with 372 images scanned from the IEEE Transactions on Pattern Analysis and Machine Intelligence show that the proposed method has performed geometrical structure analysis successfully on more than 99% of the test images, resulting in sophisticated performance compared with previous works.

1. 서 론

최근 들어 전자 문서의 활발한 보급에도 불구하고 이와 더불어 종이 문서의 양도 급속도로 증가하고 있다. 종이 문서는 전자 문서와 비교하여 저장과 검색 및 편집 등의 다양한 문서 처리에 비 효율적이다. 따라서 종이 문서를 전자 문서로 변환하기 위한 문서 영상의 이해(document image understanding)[1,2,3,4,5,6,7]에 관한 연구가 활발히 진행 중이다.

한편 SGML(standard generalized markup language)[8]과 XML(extensible markup language)[9]은 논리

† 비 회 원 : NIST 연구원
kyongho@nist.gov
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
ycchoy@rainbow.yonsei.ac.kr
sbcho@csai.yonsei.ac.kr
논문접수 2000년 5월 12일
심사완료 : 2001년 8월 4일

적인 구조 정보를 표현할 수 있으며 이기종간의 호환이 가능하다는 장점 때문에 전자도서관은 물론이고 전자상거래 등의 다양한 분야에서 전자 문서의 표준 형식으로 널리 사용되고 있다. 따라서 논리적인 구조 정보를 포함하는 문서 영상으로부터 논리적인 구성 요소를 추출하여 SGML/XML에 기반한 전자 문서를 생성하는 방법의 개발이 절실히 요구된다[10,11].

인간은 문서의 기하적인 특성에 대한 일반적인 지식은 물론이고 문서 유형과 출판물 특유의 기하적인 특성에 대한 다양한 지식을 바탕으로 문서 영상으로부터 논리적인 구성 요소를 식별한다. 따라서 문서 영상으로부터 논리적인 구성 요소의 효과적인 추출을 위하여 정교한 수준의 기하적인 구조 분석이 선행되어야 한다.

예를 들어, 논리적인 구성 요소에 해당하는 단락은 텍스트 라인과 수식의 집합으로 구성되며 텍스트 라인의 정렬 방식 등에 따라 다른 단락과 구별된다. 따라서 단락을 식별하기 위해서는 문서 영상으로부터 텍스트 라인과 수식의 정확한 식별은 물론이고, 각각의 기하적인 특성을 추출하여야 한다. 그러나 기하적인 구조 분석 방법에 관한 기존 연구[12,13,14,15,16,17]의 대부분은 기하적인 특성이 유사한 인접한 영역을 통합하고, 통합된 영역을 단순히 텍스트와 비 텍스트 영역으로 분류하기 때문에 정교한 수준의 기하적인 분석을 지원하지 않는다.

한편 문서 영상의 기하적인 특성은 문서의 유형에 따라 상이하며 동일한 유형에 속하는 문서간에도 차이가 존재할 수 있다. 예를 들어, 과학기술 논문의 경우, 일반적으로 출판물에 따라 서로 다른 종류의 포맷팅 방식이 사용된다. 따라서 정교한 수준의 기하적인 구조 분석을 지원하기 위하여 문서 유형의 일반적인 특성은 물론이고 출판물 고유의 기하적인 특성이 반영되어야 한다.

본 논문에서는 과학기술 논문을 대상으로 정교한 수준의 기하적인 구조 분석을 지원하기 위하여 지식베이스에 기반한 시스템을 제안한다. 제안된 지식베이스는 논문 유형의 공통적인 특성은 물론이고 출판물 고유의 기하적인 특성에 대한 지식을 규칙 형태로 표현한다. 제안된 방법은 영역 분할과 식별의 두 단계로 구성되며 지식베이스를 구성하는 지식 규칙 역시 적용되는 처리 단계에 따라 영역 분할 규칙과 영역 식별 규칙으로 구분된다. 한편 추론 엔진 역시 기하적인 구조 분석 과정의 효율적인 제어를 위하여 계층적인 구조로 구성된 규칙의 집합에 기반한다.

일반적으로 영역 분할에 의하여 분할된 영역과 레이아웃을 구성하는 복합 객체 사이에는 일-대-일의 대응관계가 성립되지 않는다. 예를 들어, 이미지 객체는 이미지 또는 드로잉에 해당하는 다수의 작은 영역을 포함한다. 또한 텍스트 라인의 경우, 서로 중첩하는 아래 첨자와 위 첨자에 의하여 인접한 두 영역이 통합되며, 위 첨자 등에 의하여 분할되는 경우가 존재한다.

따라서 제안된 방법은 상향식과 하향식의 복합 기법을 적용하여 분할된 영역을 추가로 분할하거나 재 통합하면서 문서 영상으로부터 이미지, 드로잉, 그리고 테이블 등의 비 텍스트 객체는 물론이고 텍스트 객체로써 텍스트 라인과 수식을 식별한다. 제안된 방법의 성능을 평가하기 위하여 IEEE Trans. Pattern Analysis and Machine Intelligence(TPAMI)로부터 스캐닝한 372개의 논문 영상으로 실험한 결과, 제안된 방법은 실험 영상 중 99% 이상의 객체에 대한 기하적인 구조 분석에 성공하여 기존 방법에 비해 정교한 수준의 성능을 보였다.

본 논문의 구성은 다음과 같다. 2절에서는 관련 연구를 통하여 지식베이스에 기반한 문서 영상의 기하적인 구조 분석 방법에 대한 기존의 연구 결과를 간략히 기술한다. 3절에서는 제안된 시스템의 개요를 기술하고, 지식베이스와 추론 엔진을 구성하는 규칙을 자세히 설명한다. 4절에서는 제안된 방법을 영역 분할과 식별의 두 단계로 구분하고, 각 단계에 대한 설명을 이에 적용되는 규칙과 더불어 자세히 기술한다. 5절에서는 실험 결과를 통하여 제안된 방법의 성능을 기존 연구와 비교 및 분석한다. 마지막으로 6절에서는 결론 및 향후 연구 방향을 기술한다.

2. 관련 연구

문서 영상의 기하적인 구조 분석과 관련한 연구는 하향식[2,3,18,19,20], 상향식[4,10,21,22] 그리고 하향식과 상향식이 혼합된 복합 기법[22,23,24,25,26]의 세가지 범주로 구분된다[1,2]. 특히 최근에 발표된 Jain과 Yu의 논문[4]은 문서 영상의 기하적인 구조 분석과 관련한 기존 연구를 연대기 순으로 간략히 기술하였다.

기존 연구의 대부분은 기하적인 구조 분석을 위하여 표현 범위와 방법은 다르지만 문서 영상의 기하적인 특성에 대한 지식을 이용한다[27]. 일반적으로 상향식과 하향식 기법에서 사용되는 지식의 특성은 서로 구별되는데, 하향식에서 사용되는 지식이 출판물의 기하적인 특성에 보다 의존적이다. 기존 연구에서 사용되는 지식의 유형은 문서에 대한 일반적인 지식(generic knowledge), 유형 특유의 지식(class-specific knowledge), 그리고 출판물 특유의 지식(publication-specific knowledge)의 세 가지로 분류할 수 있다[28]. 본 절에서는 <표 1>과 같이 기하적인 구조 분석에 관한 기존 연구 중에서 지식

베이스에 기반한 방법에 대한 간략한 설명과 문제점을 기술한다.

일반적으로 지식베이스에 기반한 하향식 방법은 출판물 특유의 지식을 이용하여 문서 영상의 기하적과 논리적인 구조 분석을 동시에 수행한다. 예를 들어, Nagy등[13,28]과 Krishnamoorthy 등[14]은 특정 출판물의 페이지 영상에 대한 기하적인 특성을 페이지 문법(page grammar)으로 표현하고 이를 기반으로 문서 영상의 분할과 논리적인 구조 분석을 수행한다. 한편 Dengel 등[29,30]은 상용 편지 형식을 대상으로 레이아웃과 논리적인 구조에 대한 지식을 기하 트리(geometric tree)에 기반한 계층적인 문서 모델로 표현한다. 또한 Higashino 등[31]은 문서의 기하적인 특성에 대한 모델을 형식 정의 언어(form definition language)로 기술한다.

표 1 지식베이스에 기반한 구조 분석 방법

관련 연구	연도	접근 방식	사용된 지식의 특징
Higashino 등 [31]	1986	하향식	배치 구조에 대한 모델을 형식 정의 언어로 기술
Nagy 등[28]	1988	하향식, 상향식	출판물 특유의 문법에 기반한 하향식 기법과 레이아웃 구성 요소에 대한 일반적인 지식에 기반한 상향식 기법
Dengel과 Barth [29]	1988	하향식	계층적인 문서 모델로 기하 트리 정의
Fisher 등[32]	1990	상향식	연결요소에 대한 지식을 규칙으로 표현
Nagy 등[13]	1992	하향식	출판물 특유의 문법 정의
Dengel 등[30]	1992	하향식	계층적인 문서 모델로 기하 트리 정의
Krishnamoorthy 등 [14]	1993	하향식	출판물 특유의 문법 정의
Esposito 등[27]	1995	복합 방식	특정 문서 유형에 독립적인 일반적인 지식을 적용
Niyogi와 Srihari [33]	1996	상향식	지식 규칙과 제어 규칙 및 전략 규칙으로 구성된 규칙 모델 정의
Sauvola 등[34]	1997	상향식	일반적인 지식과 유형 특유의 지식을 규칙 형태로 표현

상향식 방식의 경우, Nagy 등[28]은 영역 분할 결과를 트리 구조로 표현하고 단말 노드로부터 영역을 병합하면서 레이아웃 구성 요소를 생성한다. 특히 영역 병합을 위하여 사용되는 지식의 대부분은 레이아웃 구성 요소에 대한 일반적인 지식에 해당한다. 또한 지식을 표현하기 위하여 술어 논리(predicate logic)에 기반한 언어를 제안한다. Fisher 등[32]은 지식을 규칙 형태로 표현하고 이를 기반으로 문서 영상으로부터 텍스트와 비 텍스트 영역을 추출하는 방법을 제안한다. 제안된 지식베이스는 연결요소의 크기와 밀도 등의 기하적인 특성에 대한 지식을 14개의 규칙으로 표현한다.

Niyogi와Srihari[33]는 신문 영상의 전자화를 위하여 규칙 모델에 기반한 구조 분석 시스템인 DeLoS(DErivation of LOgical Structure)를 제안하였다. DeLos는 연결요소에 기반한 상향식 기법을 적용하여 영역을 분할하고, 지식베이스에 기반하여 영역을 병합 및 식별한다. 지식베이스를 구성하는 지식 규칙(knowledge rule)은 신문 영상의 논리적인 구성 요소에 대한 기하적인 특성을 기술하며, DeLos의 추론 엔진 역시 계층적인 구조로 구성된 제어 규칙(control rule)과 전략 규칙(strategy rule)에 기반한다. Sauvola 등[34]은 문서 영상을 일정한 크기의 윈도우 영역으로 분할하고 문서 영상에 대한 일반적인 지식과 유형 특유의 지식을 규칙 형태로 표현한 지식베이스를 기반으로 분할된 영역의 유형을 식별한다.

한편 Esposito 등[27]은 복합 기법과 지식베이스를 적용하여 문서 영상으로부터 텍스트, 이미지, 드로잉, 그리고 구분선(ruler)을 식별한다. 특히 지식베이스는 포맷팅에 관한 일반적인 지식을 포함한다.

기존 연구는 문서 영상에 대한 다양한 수준의 지식을 여러 가지 형태의 지식베이스로 표현하지만 정교한 수준의 기하적인 구조 분석을 지원하지 않는다.[13,14,28, 29,30,31,32]는 비 텍스트 영역에 대한 식별 방법을 지원하지 않으며 [27,33,34]는 이미지와 드로잉 등의 비 텍스트 영역에 대한 단순한 수준의 구조 분석을 수행한다.

일반적으로 문서 영상의 전자화를 효율적으로 수행하기 위해서는 기하적인 구조 분석의 결과로 식별된 영역과 레이아웃 객체 사이에 대응 관계가 높아야 한다. 또한 텍스트 영역으로부터 논리적인 구성 요소를 정확히 추출하기 위해서는 텍스트 라인을 제목, 단락의 첫번째와 마지막 라인, 또는 리스트 등으로 분류하여야 한다. 따라서 텍스트 영역으로부터 텍스트 라인의 정확한 추출은 물론이고 수식 영역을 식별할 수 있는 방법이 요구된다.

3. 시스템 개요

제안된 시스템은 <그림 1>과 같이 영상 분석 모듈, 규칙 모델, 그리고 규칙 기반 시스템의 세 가지로 구성된다. 또한 영역 분할과 식별의 두 단계로 구성된 기하적 구조 분석 방법을 지원하기 위하여 영상 분석 모듈은 영역 분할과 영역 식별 모듈로 이루어진다.

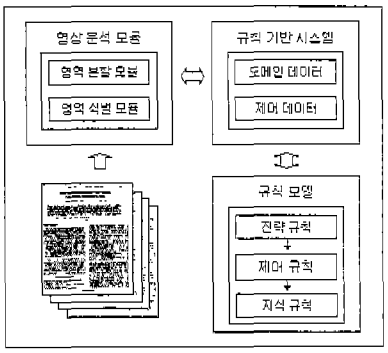


그림 1 시스템 개요

규칙 기반 시스템은 영상 분석의 중간 결과에 따라 규칙 모델을 기반으로 적절한 구조 분석 모듈을 선택 및 실행함으로써 기하적인 구조 분석 과정을 제어한다. 특히 규칙 기반 시스템은 구조 분석 과정에서 발생하는 다양한 정보를 저장하기 위하여 도메인 데이타와 제어 데이타 영역을 포함한다.

도메인 데이타 영역은 처리 대상에 대한 정보인 영상 분석 모듈을 적용하여 생성 및 변경된 영역에 관한 정보를 저장한다. 한편 제어 데이타 영역은 영상 분석 모듈의 실행 결과와 적용된 규칙 등의 처리 과정에 대한 제어 정보를 유지한다. 특히 규칙 기반 시스템은 영상 분석의 처리 결과에 따라 가장 적합한 규칙을 선택함으로써 기하적인 구조를 효율적으로 분석한다.

제안된 방법은 논문 영상의 기하적인 구조 분석을 위하여 세 단계의 계층적인 구조로 구성된 규칙 모델을 이용한다. 지식베이스를 구성하는 지식 규칙은 논문 유형이 공통적으로 갖는 기하적인 특성은 물론이고 출판물 고유의 기하적인 특성에 대한 지식을 표현한다. 특히 지식 규칙은 적용되는 처리 모듈에 따라 영역 분할 규칙과 영역 식별 규칙으로 구분된다.

추론 엔진을 구성하는 제어 규칙은 처리 대상 영역과 이에 적용할 처리 과정을 결정하는 역할을 한다. 즉, 지식 규칙이 적용될 영역을 선택하거나 현재 영역에 적용될 처리 과정을 결정함으로써 이에 해당하는 지식 규칙의 집합을 선택한다. 한편 전략 규칙은 문서 분석 과정의 효율적인 통제를 위하여 적절한 제어 규칙을 결정하는 역할을 한다.

한편 기존 연구는 주로 자연 영상의 영역 분할[35]이나 신문 영상의 논리적인 구조 분석[33]을 위하여 본 논문과 마찬가지로 계층 구조의 규칙 모델을 적용하였다. 또한 실험을 통하여 이와 같은 규칙 모델이 영상 분석의 모듈 단위 처리와 유연한 추론 메커니즘을 제공한다는 점을 보였다. 지식베이스와 추론 엔진을 구성하는 규칙에 대한 보다 자세한 사항은 다음과 같다.

3.1 지식베이스

제안된 시스템의 지식베이스를 구성하는 지식 규칙은 {IF 조건AND …AND조건 THEN 적용될 방법} 형태로 표현된다. 따라서 특정 규칙이 포함하는 조건을 모두 만족하는 경우에 해당 방법의 적용이 가능하다.

일반적으로 논문 영상은 출판물에 따라 다양한 종류의 포맷팅 방식이 사용되며, 논문의 시작 페이지와 나머지 페이지의 레이아웃이 다른 경우가 많다. 예를 들어, 시작 페이지의 일단 영역은 논문 제목, 저자 성명, 소속, 요약, 그리고 키워드 등의 텍스트 영역으로 구성되지만, 나머지 페이지의 일단 영역은 주로 그림과 테이블 등의 비 텍스트 영역으로 구성된다.

따라서 제안된 지식 규칙은 논문 영상이 속하는 출판물과 페이지의 종류에 따라 유형 규칙(class-specific rule), 출판물 규칙(publication-specific rule), 제목 페이지 규칙(title-page rule), 그리고 본문 페이지 규칙(body-page rule)으로 구분된다. 특히 본 논문에서는 시작 페이지를 제목 페이지, 그리고 나머지 페이지를 본문 페이지로 정의한다.

유형 규칙과 출판물 규칙은 각각 논문 유형에 속하는 문서 영상이 공통적으로 갖는 특성과 출판물 특유의 기하적인 특성을 기술한다. 또한 제목 페이지 규칙과 본문 페이지 규칙은 각각 시작 페이지와 나머지 페이지 특유의 기하적 특성을 표현한다. 따라서 출판물 규칙은 유형 규칙에 속하는 지식 규칙을 상속 받으며 제목 페이지 규칙과 본문 페이지 규칙은 출판물 규칙에 속하는 지식 규칙을 상속 받는다.

본 논문은 제안된 방법은 구현하기 위하여 TPAMI의 정규 논문(regular paper)을 대상으로 지식베이스를 구축하였다. 이를 위하여 문서 영상을 제목 페이지와 본문 페이지로 구분하고, 각 레이아웃의 특성을 조사하였다. 또한 논문 영상의 영역을 텍스트 라인, 수식, 이미지, 드로잉, 그리고 테이블 등의 복합 객체로 구분하고 각각의 기하적인 특성을 조사하여 규칙 형태의 지식베이스를 구축하였다. 구축된 지식베이스는 91개의 지식 규칙으로 구성된다.

제안된 방법은 논문 영상의 정교한 기하적인 구조 분석을 위하여 영역 식별은 물론이고 영역 분할을 위하여 논문 영상의 기하적인 특성에 대한 지식을 이용한다. 따라서 제안된 지식 규칙은 규칙이 적용되는 처리 모듈에 따라 영역 분할 규칙과 영역 식별 규칙으로 구분된다.

규칙 (10)은 이단 영역의 분할 규칙을 기술하는 영역 분할 규칙이다. 즉, 현재 영역이 이단 영역의 기하적인 특성을 기술한 8개의 조건을 모두 만족한다면 수직 방향의 공백을 기준으로 영역을 분할한다.

규칙 (10):

IF : (1) 현재 영역의 단 유형이 이단 영역이 아니다.
(2) 현재 영역의 넓이는 인쇄 영역의 넓이와 유사하다.
(3) 수평 방향의 흰 공백이 존재한다.
(4) 수직 방향의 흰 공백이 존재한다.
(5) 수직 방향의 흰 공백의 넓이는 수평 방향의 흰 공백의 높이보다 크다.
(6) 수직 방향의 흰 공백의 넓이는 임계값 ThWidth1보다 작다.
(7) 수직 방향의 흰 공백의 넓이는 임계값 ThWidth2보다 크다.
(8) 수직 방향의 흰 공백은 인쇄 영역의 중앙에 위치한다.

THEN: (1) 현재 영역을 수직 방향으로 분할한다.
(2) 분할된 영역의 단 유형을 이단 영역으로 설정한다.

3.2 추론 엔진

제안된 시스템의 추론 엔진은 제어 규칙과 전략 규칙에 기반한다. 제어 규칙과 전략 규칙은 영역을 직접적으로 수정하지 않는다는 점에서 지식 규칙과 구별된다. 제어 규칙은 각각 지식 규칙이 적용될 영역과 이에 적용될 지식 규칙의 집합을 결정하는 집중점 규칙(focus-of attention rule)과 메타 규칙(meta rule)으로 이루어진다. 집중점 규칙은 구조 분석 모듈의 처리 대상으로서 지식 규칙이 적용될 가장 적절한 영역을 선택한다. 메타 규칙은 구조 분석 과정의 상태에 따라 적절한 처리 모듈을 선택한다. 이는 특정한 처리 모듈과 연관이 있는 지식 규칙의 집합을 결정하는 역할을 한다.

예를 들어, 집중점 규칙 (1)은 현재 영역에 대하여 수평 방향으로 더 이상 분할 과정을 적용할 수 없는 경우에 영역 분할의 대상으로 인접한 영역을 선택하는 역할을 한다. 메타 규칙 (1)은 영역 분할을 수행하기 위하여 단 영역의 기하적인 특성에 해당하는 적절한 매개 변수의 설정과 더불어 해당 지식 규칙의 집합을 선택한다.

집중점 규칙 (1):

IF: (1) 현재 처리 상태가 "영역 분할 과정"에 해당한다.
(2) 현재 영역의 단 유형이 이단 영역이 아니다.
(3) 수평 방향의 흰 공백이 존재하지 않는다.

THEN : (1) 현재 영역의 단 유형을 일단 영역으로 설정한다.
(2) 현재 영역의 유형을 일반 텍스트 라인으로 설정한다.
(3) 유형이 일반 텍스트 라인으로 설정되지 않은 인접한 영역을 선택한다.

메타 규칙 (1):

IF : (1) 영역 분할 모듈의 시작

THEN: (1) 단 구조의 기하적인 특성에 해당하는 매개 변수를 초기화한다.
(2) 입력 영상을 영역으로 설정하고 이에 "영역 분할 규칙"을 적용한다.

한편 전략 규칙은 영역 분할과 식별 과정에서 제어 규칙의 종류와 실행 순서를 결정한다. 또한 각각의 처리 과정이 올바르게 수행되었는지의 여부 등을 검사한다. 따라서 제어 규칙과 전략 규칙은 문서 영상의 기하적인 구조 분석 과정을 통제하는 역할을 수행한다. 전략 규칙 (1)은 비 텍스트 영역의 식별 과정이 성공적으로 종결될 때까지 관련된 제어 규칙을 적용한다.

전략 규칙 (1):

IF : (1) 유형이 식별되지 않은 비 텍스트 라인이 존재한다.

THEN : (1) 유형이 식별되지 않은 영역이 존재하지 않을 때까지 각각의 영역에 영역 식별에 필요한 제어 규칙을 적용한다.

4. 지식베이스에 기반한 기하적인 구조 분석 시스템

본 절에서는 논문 영상의 기하적인 구조 분석을 수행하기 위해 제안된 시스템에 대하여 기술한다. 제안된 시스템의 구조 분석 과정은 <그림 2>와 같이 영역 분할과 식별의 두 단계로 구성된다.

4.1 영역 분할

제안된 방법은 논문 영상의 기하적인 특성에 대한 지식을 이용하여 단 영역을 추출하고, 각각의 단 영역을 수평 방향으로 분할한다. 이를 위하여 먼저 문서 영상에 런(run)에 기반한 방법[4]을 적용하여 연결요소를 추출한 후, 연결요소에 대한 투영윤곽도(projection profile) 분석 방법[36]과 영역 분할 규칙을 적용하여 영역을 분할한다.

문서 영상은 수직과 수평 방향의 투영윤곽도에 의하여 수직 또는 수평 방향으로 분할된다. 수직 방향의 투영윤곽도로부터 선택된 흰 영역의 넓이와 수평 방향의

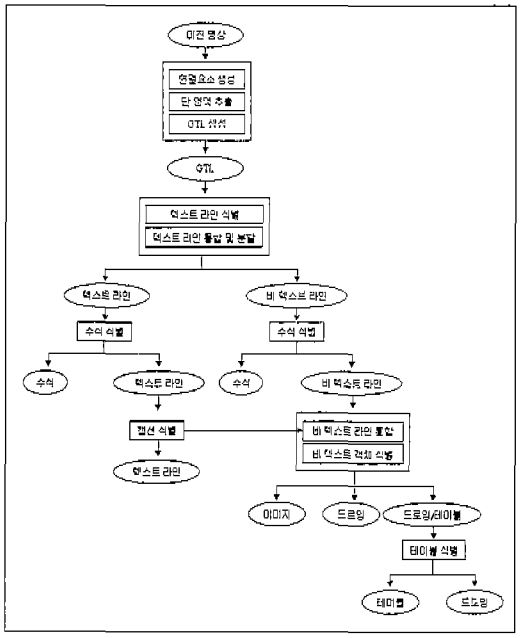


그림 2 제안된 기하적인 구조 분석 과정의 흐름도

투영윤곽도로부터 선택된 흰 영역의 높이를 비교하여 크기가 가장 큰 영역을 기준으로 영역을 분할한다. 지식 규칙에 해당하는 규칙 (10)과 (11)은 이러한 사실을 표현한다. 특히 단 영역은 영역을 수직 방향으로 분할함으로써 생성되는데, 수직 방향의 흰 영역은 논문 영상의 단 구조에 대한 기하적인 특성을 만족하여야 한다.

규칙 (11):

IF : (1) 현재 영역의 단 유형이 "이단 영역"이 아니다.
(2) 수평 방향의 흰 공백이 존재한다.
(3) 수직 방향의 흰 공백이 존재한다.
(4) 수평 방향의 흰 공백의 최대 높이는 수직 방향의 흰 공백의 최대 넓이보다 크다.

THEN : (1) 현재 영역을 수평 방향으로 분할한다.

수직 분할이 적용되지 않은 영역에 대해서는 규칙 (10)과 (11)을 반복적으로 적용하지만 수직 분할에 의하여 생성된 영역의 경우, 수평 방향의 투영윤곽도를 구성하는 흰 영역을 기준으로 수평 방향으로 분할한다. 제안된 영역 분할 방법은 규칙 (12)와 (13)과 같이, 각각의 단 영역을 수평 방향으로 분할한 영역의 집합을 생성한다. 본 논문에서는 이를 일반 텍스트 라인(GTL: generalized text line) [4]이라고 정의한다.

규칙 (12):

IF : (1) 현재 영역의 단 유형이 "이단 영역"이다.
(2) 수평 방향의 흰 공백이 존재한다.

THEN : (1) 높이가 가장 큰 수평 방향의 흰 공백을 선택한다.
(2) 흰 공백을 기준으로 현재 영역을 수평 방향으로 분할한다.

규칙 (13):

IF : (1) 현재 영역의 단 유형이 "이단 영역"이다.
(2) 수평 방향의 흰 공백이 존재하지 않는다.

THEN : (1) 현재 영역의 유형을 "일반 텍스트 라인"으로 설정한다.

4.2 영역 식별

전술한 바와 같이 과학기술 논문의 제목 페이지와 본문 페이지의 레이아웃은 서로 상이하다. 예를 들어, 제목 페이지는 머리말, 꼬리말, 그리고 본문 영역으로 구성되며 본문 페이지는 머리말과 본문 영역을 포함한다. 여기서 본문 영역이란 제목 페이지와 본문 페이지에서 머리말과 꼬리말을 제외한 나머지 영역을 의미한다. 따라서 제안된 영역 식별 방법은 논문 영상을 제목 페이지와 나머지 페이지로 구분하고 각각의 기하적인 특성을 기술한 지식 규칙을 이용하여 영역을 식별한다.

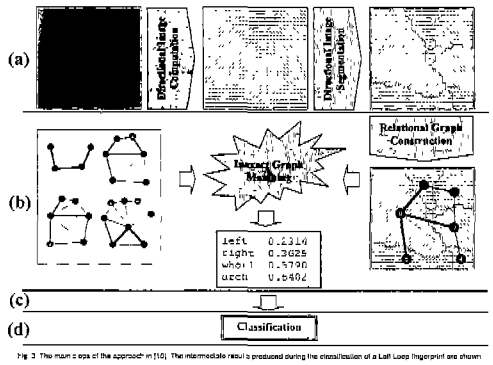


그림 3 그림 객체의 영역 분할 결과

제안된 영역 분할 방법은 단 영역을 단순히 수평 방향으로 분할하여 일반 텍스트 라인의 집합을 생성한다. 따라서 일반적으로 분할된 영역과 문서 영상의 레이아웃을 구성하는 객체 사이에는 일-대 일의 대응관계가 성립되지 않는다. 예를 들어, <그림 3>의 경우, 영역 분할 방법을 적용하면 그림 객체는 네 개의 세부 영역 (a) ~ (d)로 분할된다. 따라서 먼저 분할된 영역을 통합하고, 통합된 영역의 유형을 식별하는 방법이 요구된다. 제안된 방법이 영역 식별 규칙을 적용하여 분할된 영역

을 통합하거나 추가로 분할하면서 텍스트 객체와 비 텍스트 객체를 식별하는 과정에 대한 자세한 설명은 다음과 같다.

4.2.1 텍스트 영역

일반적으로 문서 영상을 구성하는 텍스트 영역은 논문 제목과 요약 등의 다양한 종류의 논리적인 객체로 구성된다. 또한 각각의 논리적인 객체는 이를 구성하는 텍스트 라인의 기하적인 특성에 의하여 서로 구별된다. 한편 과학기술 논문을 구성하는 단락은 텍스트 라인은 물론이고 수식을 다수 포함한다. 텍스트 영역으로부터 텍스트 라인과 수식을 식별하는 방법에 대한 자세한 기술은 다음과 같다.

(1) 텍스트 라인의 추출

본 절에서는 영역 분할의 결과인 일반 텍스트 라인의 집합으로부터 텍스트 라인과 비 텍스트 라인을 식별하는 방법을 기술한다. 일반적으로 비 텍스트 객체를 구성하는 연결요소의 크기는 텍스트 객체를 구성하는 연결요소보다 크기 때문에 먼저 규칙 (20) 등을 적용하여 일반 텍스트 라인을 텍스트 라인과 비 텍스트 라인으로 구별한다.

규칙 (20):

IF : (1) 영역의 단 유형이 이단 영역이다.
(2) 영역의 높이는 임계값 ThHeight5보다 크다.
(3) 영역의 높이가 임계값 ThHeight6보다 작다.
(4) 영역의 넓이는 임계값 ThWidth5보다 크거나 같다.
(5) 영역의 밀도는 임계값 ThDensity1보다 크다.
(6) 영역의 밀도는 임계값 ThDensity2보다 작다.
(7) 영역을 구성하는 연결요소의 밀도는 임계값 ThCCDensity1보다 크다.
(8) 영역을 구성하는 연결요소의 밀도는 임계값 ThCCDensity2보다 작다.

THEN : (1) 유형을 텍스트 라인으로 설정한다.

한편 <그림 4>와 <그림 5>와 같이, 제안된 영역 분할 방법은 단 영역을 단순히 수평 분할하기 때문에 인접한 텍스트 라인이 한 개의 영역으로 통합되거나 텍스트 라인이 두개의 영역으로 분할되는 경우가 존재한다. 예를 들어, 과학기술 논문의 경우, 텍스트 라인을 구성하는 첨자와 수식의 일부에 의하여 한 개의 텍스트 라인이 두 개의 영역으로 분할될 수 있다. 또한 인접하는 텍스트 라인을 구성하는 아래 첨자와 위 첨자의 하행자(descender)와 상행자(ascender)가 서로 중첩하여 인접한 텍스트 라인이 단일의 영역으로 통합될 수 있다.

본 논문에서는 TPAMI의 논문 26편을 대상으로 제

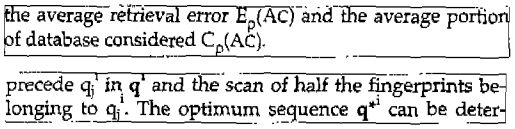


그림 4 아래 첨자 또는 위 첨자에 의하여 잘못 통합된 텍스트 라인의 예

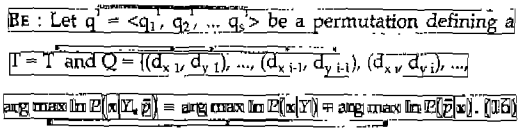


그림 5 첨자 또는 수식에 의하여 분할된 텍스트 라인의 예

안된 영역 분할 방법을 적용한 후 통합된 텍스트 라인의 빈도수를 조사하였다. 조사 결과, 총 319개의 텍스트 라인이 통합되어 논문 1편당 평균적으로 12.27개의 텍스트 라인이 통합된 것을 알 수 있었다.

제안된 방법은 텍스트 라인의 기하적인 특성에 대한 지식 규칙을 적용하여 통합된 텍스트 라인을 식별하고 이를 분할한다. Kanai[37]는 문자의 종류를 높이에 따라 구분하고, 실험을 통하여 문서 내에서 가장 빈도가 높은 문자의 종류가 존재한다는 사실을 입증하였다. 한편 본 논문은 문서 영상을 구성하는 연결요소 중에서 가장 빈도수가 높은 것은 문자이고, 가장 빈도수가 높은 문자와 텍스트 라인의 높이는 서로 비례한다는 가정에 기반한다.

본 연구에서는 가장 빈도수가 높은 문자의 높이와 텍스트 라인의 높이 사이의 상대적인 비율을 조사하기 위하여 다양한 문서 영상을 대상으로 실험하였다. 실험 결과, 가장 빈도수가 높은 문자와 텍스트 라인의 높이 사이에는 수식 (1)과 같은 상대적인 비율이 존재함을 알 수 있었다.

텍스트 라인의 평균 높이
=가장 빈도수가 높은 문자의 평균 높이 $\times \alpha$ (1)

따라서 영역 분할 결과로 생성된 연결요소의 높이 분포를 분석하여 텍스트 라인의 평균적인 높이를 계산하고, 이를 기반으로 통합된 텍스트 라인의 후보 영역을 선택한다. 수식 (2)는 후보 영역의 높이로부터 통합된 텍스트 라인의 개수를 계산한다.

통합된 텍스트 라인의 개수
=후보 영역의 높이/텍스트 라인의 평균 높이 (2)

본 논문에서는 통합된 문자열을 정확히 식별하기 위하여 후보 영역에 대한 검증 과정을 적용한다. 이를 위하여 분할된 후보 영역의 신뢰도를 계산하는데, 일반적으로 문서 내에서 가장 빈도수가 높은 문자의 밑변은 해당 텍스트 라인의 기준선(base line)이 된다[3]. 따라서 가장 빈도수가 높은 문자의 높이에 해당하는 연결요소를 기준으로 분할될 텍스트 라인의 기준선을 계산하고 이를 기준으로 후보 영역을 분할한다.

만일 분할된 영역이 텍스트 라인에 해당한다면 후보 영역에 속하는 연결요소의 대부분은 분할된 영역에 완전히 포함될 것이다. 따라서 수식 (3)을 적용하여 분할된 영역의 신뢰도를 검증한다. 본 논문에서는 실험을 통하여 통합된 텍스트 라인의 경우, 신뢰도가 0.9이상임을 알 수 있었다.

신뢰도

$$= \frac{\sum \text{분할된영역에완전히포함된 } ConnectedComponents}{\sum ConnectedComponents} \quad (3)$$

한편 분할된 텍스트 라인을 통합하는 방법은 다음과 같다. 일반적으로 텍스트 라인을 구성하는 첨자 또는 수식의 일부에 의하여 분할된 영역의 높이는 상대적으로 낮고, 분할된 영역은 서로 인접한다. 따라서 높이가 작은 일반 텍스트 라인의 경우, 분할된 영역으로 간주하여 이웃 하는 영역에 통합한다. 제안된 방법을 <그림 4>와 <그림 5>에 적용한 결과는 각각 <그림 6>과 <그림 7>과 같다.

the average retrieval error $E_p(AC)$ and the average portion of database considered $C_p(AC)$.

precede q_i' in $\mathbf{q}'$ and the scan of half the fingerprints belonging to q_j'. The optimum sequence $\mathbf{q}^{*'}$ can be deter-

그림 6 <그림 4>의 통합 영역을 분할한 결과

BE : Let $\mathbf{q}' = <q_1', q_2', \ldots q_b'>$ be a permutation defining a

$T = T^*$ and $Q = \{(d_{x\,1}, d_{y\,1}), \ldots, (d_{x\,i-1}, d_{y\,i-1}), (d_{x\,i}, d_{y\,i}), \ldots,$

$$\arg\max \ln P(\mathbf{x}|Y, \vec{p}) \equiv \arg\max \ln P(\mathbf{x}|Y) + \arg\max \ln P(\vec{p}|\mathbf{x}). \quad (16)$$

그림 7 <그림 5>의 분할된 문자열을 통합한 결과

특히 제안된 시스템은 규칙 (16)과 (20) 등을 적용하여 일반 텍스트 라인을 통합 및 분할하고, 일반 텍스트 라인으로부터 텍스트 라인과 비 텍스트 라인을 식별한다.

규칙 (16):

IF : (1) 일반 텍스트 라인의 높이가 임계값 ThHeight2의 두 배 이상이다.

(2) 대부분 문자의 크기에 해당하는 연결요소로 구성된다.

(3) 분할된 영역의 신뢰도가 ThConfidence보다 크거나 같다.

THEN : (1) 영역을 분할하여 새로운 영역을 생성한다.

(2) 영역의 유형을 일반 텍스트 라인으로 설정한다.

규칙 (20):

IF : (1) 영역의 단 유형이 이단 영역이다.

(2) 영역의 높이는 임계값 ThHeight5보다 크다.

(3) 영역의 높이가 임계값 ThHeight6보다 작다.

(4) 영역의 넓이는 임계값 ThWidth5보다 크거나 같다.

(5) 영역의 밀도는 임계값 ThDensity1보다 크다.

(6) 영역의 밀도는 임계값 ThDensity2보다 작다.

(7) 영역을 구성하는 연결요소의 밀도는 임계값 ThCCDensity1보다 크다.

(8) 영역을 구성하는 연결요소의 밀도는 임계값 ThCCDensity2보다 작다.

THEN : (1) 유형을 텍스트 라인으로 설정한다.

(2) 텍스트 라인으로부터 수식의 식별

제안된 방법은 영역의 크기, 밀도, 정렬 방식 등의 기하적인 특성을 이용하여 텍스트 라인의 집합으로부터 수식을 구별한다.

The game is being played out by a set of decision makers (or players) which, for our case, will correspond to the two segmentation modules that we want to integrate. Let $N = 1, 2$ be the player set, P^1 and P^2 be the strategy spaces of the first and second decision makers (players), respectively, and let

$$F^1 : P^1 \times P^2 \to R\,;\; F^2 : P^1 \times P^2 \to R$$

$$F^1(\vec{p}^1, \vec{p}^2) \le F^1(p^1, \vec{p}^2); F^2(\vec{p}^1, \vec{p}^2) \le F^2(\vec{p}^1, p^2). \quad (1)$$

be the cost functions for the decision makers 1 and 2, respectively.

그림 8 텍스트 라인과 수식의 예

<그림 8>과 같이, 수식과 텍스트 라인은 주로 문자와 숫자로 구성된다. 한편 수식과 비교하여 텍스트 라인을 구성하는 연결요소는 전체 영역에 고르게 분포한다. 따라서 텍스트 라인 영역의 검은 화소의 밀도는 수식보다 비교적 높다고 가정할 수 있다. 또한 <그림 8>에서와 같이, 일반적으로 텍스트 라인과 수식의 정렬방식은 서로 상이하다.

한편 <그림 9>와 같이, 영역 분할 결과 단일 수식이 여러 개의 영역으로 분할되는 경우가 있다. 이와 같이

segmented to simple uniform regions, then, of course, we simply have

$$u = u_{i,j} = \sum_{(i,j)\in A_{ij}} x_{i,j}$$

and

$$v = v_{i,j} = \sum_{(i,j)\in A_{ij}} x_{i,j}$$

However, more complex modeling of the intensity infor-

그림 9 텍스트 영역의 분할 결과

분할된 세부 영역의 경우, 밀도 분포와 정렬방식 등은 해당 객체의 기하적인 특성을 만족하지 않는다. 따라서 기하적인 특성에 대한 지식을 만족하지 않는 영역의 경우, 인접한 영역의 유형으로 설정하는 것이 바람직하다. 제안된 방법이 텍스트 라인으로부터 수식을 식별하기 위하여 사용하는 지식 규칙의 예는 규칙 (25)와 같다.

규칙 (25);

IF : (1) 텍스트 라인의 밀도가 임계값 ThDensity3보다 크다.
(2) 텍스트 라인의 밀도가 임계값 ThDensity4보다 작다.
(3) 텍스트 라인을 구성하는 연결요소의 밀도가 임계값 ThCCDensity3보다 크다.
(4) 텍스트 라인을 구성하는 연결요소의 밀도가 임계값 ThCCDensity4보다 작다.

THEN : (1) 영역의 유형을 "수식"으로 설정한다.

4.2.2 비 텍스트 영역

전단계에서 제안된 방법은 크기가 큰 영역을 비 텍스트 라인으로 분류하기 때문에 크기가 큰 수식 영역은 비 텍스트 라인으로 분류된다. 따라서 제안된 방법은 먼저 비 텍스트 라인으로부터 텍스트 객체에 해당하는 수식 영역을 식별한 후 나머지 비 텍스트 라인을 재 통합하여 이미지, 드로잉, 그리고 테이블 등의 비 텍스트 객체를 식별한다.

(1) 비 텍스트 라인으로부터 수식 영역의 식별

일반적으로 수식 영역은 주로 문자나 숫자 등의 비교적 크기가 작은 연결요소를 포함한다. 반면에 이미지, 드로잉, 그리고 테이블에 해당하는 영역의 경우, 상대적으로 크기가 큰 연결요소의 사각형 영역(해당 연결요소를 포함하는 최소 사각형 영역)이 작은 연결요소를 포함한다.

본 논문에서는 비 텍스트 라인의 집합으로부터 수식을 구별할 수 있는 기하적인 특성을 습득하기 위하여 영역을 구성하는 독립 연결요소(independent connected-component)의 기하적인 특성을 조사하였다. 본 논문은

표 2 비 텍스트 라인을 구성하는 독립 연결요소의 특성 분포

종류	독립 연결요소			
	개수	면적		
		평균	최소	최대
수식	43.507	421.350	225.333	705.260
이미지	10	9,903.800	9,903.800	9,903.800
드로잉	25	39,575.020	2591.800	62180
테이블	1	46,585.504	46,585.504	46,585.504

다른 연결요소의 사각형 영역에 포함되지 않는 연결요소를 독립 연결요소라고 정의한다. 실험 결과, <표 2>와 같이 수식 영역을 구성하는 독립 연결요소의 면적의 평균값은 비 텍스트 객체의 영역과 구별됨을 알 수 있다. 제안된 방법은 비 텍스트 라인에 포함되는 수식 영역의 기하적인 특성에 대한 지식을 규칙으로 정의하고, 이를 기반으로 비 텍스트 라인으로부터 수식을 식별한다. 규칙 (27)은 비 텍스트 라인으로부터 수식을 식별하기 위하여 적용되는 규칙의 예이다.

규칙 (27) :

IF : (1) 영역의 단 유형이 이단 영역이다.
(2) 현재 영역의 유형이 비 텍스트 라인이다.
(3) 영역을 구성하는 독립 연결요소의 면적이 임계값 ThArea1보다 크다.
(4) 영역을 구성하는 독립 연결요소의 면적이 임계값 ThArea2보다 작다.
(5) 영역의 높이가 임계값 ThHeight6보다 작다.
(6) 영역은 수식의 정렬 방식을 갖는다.

THEN : (1) 영역의 유형을 수식으로 설정한다.
(2) 이미지, 드로잉, 그리고 테이블 객체의 식별

본 절에서는 수식을 식별하고 남은 비 텍스트 라인으로부터 이미지, 드로잉, 그리고 테이블을 식별하는 방법을 기술한다. 일반적으로 그림을 구성하는 비 텍스트 라인은 이미지 또는 드로잉에 해당하는 세부 영역을 모두 포함할 수 있다. 따라서 본 논문에서는 비 텍스트 라인의 유형을 이미지, 이미지와 드로잉이 혼합된 영역(이미지+드로잉), 드로잉, 그리고 테이블 영역으로 세분한다.

일반적으로 문서 영상을 구성하는 비 텍스트 객체와 비 텍스트 라인 사이에는 일-대-다의 대응관계가 성립한다. 따라서 비 텍스트 객체를 식별하기 위하여 인접한 비 텍스트 라인을 통합하고, 영역을 구성하는 비 텍스트 라인의 유형에 따라 통합된 영역의 종류를 식별한다.

비 텍스트 영역의 유형 =
1(이미지)
if 밀도 > ThDensity10
2(이미지 또는 이미지+드로잉)
else if ThDensity11 ≤ 밀도 ≤ ThDensity10
3(이미지+드로잉)
else if ThDensity12 < 밀도 < ThDensity11
4(이미지+드로잉 또는 드로잉)
else if ThDensity13 < 밀도 ≤ ThDensity12
5(이미지+드로잉 또는 드로잉 또는 테이블)
else if ThDensity14 ≤ 밀도 ≤ ThDensity13
6(테이블 또는 드로잉)
else if ThDensity15 ≤ 밀도 < ThDensity14
7(드로잉)
else if 밀도 < ThDensity15 (4)

일반적으로 이미지 영역의 밀도는 높으며 드로잉 영역은 비교적 낮은 밀도 분포를 갖는다. 본 논문에서는 샘플 영상의 비 텍스트 라인을 이미지, 드로잉, 이미지+드로잉, 그리고 테이블로 구분하고 각각의 밀도 분포를 조사하였다. 수식 (4)의 규칙은 밀도 분포에 따라 비 텍스트 라인의 유형을 7가지로 분류한 결과이다. 규칙 (51)은 밀도를 기준으로 비 텍스트 라인의 유형을 식별하는 규칙의 예이다.

규칙 (51):
IF : (1) 영역의 유형이 비 텍스트 라인이다.
(2) 영역의 밀도가 임계값 ThDensity10보다 작다.
(3) 영역의 밀도가 임계값 ThDensity11보다 크거나 같다.
THEN : (1) 영역의 유형을 "이미지 또는 이미지+드로잉"으로 설정한다.

한편 제안된 방법은 완전한 비 텍스트 객체를 추출하기 위하여 인접한 비 텍스트 라인을 통합하고, 통합 영역의 유형을 식별한다. 이를 위하여 제안된 방법은 수식 (5)에 기반한 규칙 (58) 등을 적용하여 통합된 영역의 종류를 식별한다.

통합 영역의 종류 =
8(이미지) if 유형 1, 2, 3, 또는 4를 포함
9(드로잉) else if 유형 7을 포함
10(드로잉 또는 테이블)
else if 유형 5와 6만을 포함 (5)

예를 들어, 제안된 방법은 수식 (4)의 규칙을 적용하여 <그림 13>의 비 텍스트 라인 (a) ~ (d)의 유형을 각각 유형 2, 유형 5, 유형 5, 그리고 유형 4로 분류한다. 또한 인접한 영역을 통합하고, 통합된 영역에 수식 (5)의 규칙을 적용하여 통합 영역의 유형을 이미지로 식별한다.

규칙 (58):
IF : (1) 영역의 유형이 "이미지"에 해당한다.
(2) 인접한 영역의 유형이 "이미지 또는 이미지+드로잉"에 해당한다.
THEN : (1) 두 영역을 통합한다.
(2) 통합된 영역의 유형을 "이미지"로 설정한다.

한편 통합 영역을 구성하는 비 텍스트 라인의 유형이 5 또는 6에 해당한다면 통합 영역의 유형은 "드로잉 또는 테이블"로 분류된다. 제안된 방법은 테이블의 기하적인 특성에 대한 일반적인 지식을 이용하여 테이블의 여부를 식별한다. 제안된 방법은 테이블로 식별되는 않은 영역은 드로잉으로 간주한다. 이를 위하여 적용되는 지식 규칙의 예는 규칙 (107)과 같다.

규칙 (107):
IF : (1) 수평과 수직 방향의 직선 요소를 한 개 이상 포함한다.
(2) 수평과 수직 방향의 직선 요소는 서로 직교한다.
(3) 텍스트 영역을 구성하는 연결요소는 수직 또는 수평 방향으로 정렬되어 있다.
THEN : (1) 영역의 유형을 "테이블"로 설정한다.

일반적으로 테이블을 구성하는 프레임은 수평 또는 수직 방향의 직선 요소로 구성되며 이들은 서로 직교한다. 또한 테이블의 텍스트 영역을 구성하는 연결요소는 수직 또는 수평 방향으로 정렬되어 있다. 제안된 방법은 테이블을 구성하는 프레임과 텍스트의 기하적인 특성을 이용하여 테이블을 식별한다. 여기에서 테이블의 프레임은 각각 영역의 넓이와 높이와 크기가 유사한 수평과 수직 방향의 직선 요소를 한 개 이상 포함한다고 가정한다.

테이블의 여부를 조사하기 위하여 통합 영역으로부터 수평과 수직 방향의 직선 요소를 추출하는 방법은 다음과 같다. 먼저 연결요소를 구성하는 수평 방향의 런의 집합으로부터 수평 방향의 직선 요소를 추출한다. 프레임 영역을 구성하는 수평 방향의 런 중에서 길이가 큰 런을 병합하여 수평 방향의 직선 영역을 추출하고, 이로부터 수평 방향의 직선 선분을 근사한다.

특히 직선 영역을 구성하는 런을 선택하기 위하여 사용된 임계값은 연결요소의 생성단계에서 결정된 가장 빈도수가 높은 문자의 높이로 설정하였다. 가장 빈도수가 높은 문자의 높이는 연결요소의 생성 단계에서 동적

으로 결정된다. 한편 수직 방향의 직선 선분을 추출하기 위하여 먼저 수평 방향의 런으로부터 수직 방향의 런을 생성한다. 수직 방향의 직선 선분을 추출하는 방법은 수평 방향과 동일하다.

프레임 영역으로부터 추출된 직선 요소가 테이블의 기하적인 특성을 만족한다면 텍스트 영역의 기하적인 특성을 분석한다. 먼저 텍스트 영역을 구성하는 연결요소의 크기가 균일한지의 여부를 조사한다. 또한 연결요소의 투영윤곽도를 분석하여 연결요소가 수평과 수직 방향으로 정렬되어 있는지의 여부를 조사한다.

5. 실험 결과 및 성능 분석

제안된 방법의 성능을 평가하기 위하여 1999년 1월부터 6월 사이에 발행된 TPAMI의 정규 논문 26편으로부터 스캐닝한 372개의 논문 영상을 대상으로 실험하였다. 실험 결과, <그림 10> ~ <그림 12>와 같이 제안된 방법은 문서 영상으로부터 이미지, 드로잉, 그리고 테이블 등의 비 텍스트 객체는 물론이고 텍스트 객체로써 텍스트 라인과 수식을 식별하였다.

표 3 성능 평가

객체의 종류	식별된 객체의 개수	식별되지 않은 경우	정확도(%)
텍스트 라인	22,193	53	99.8
수식	900	97	90.3
이미지	158	6	96.3
드로잉	192	2	99.0
테이블	71	6	92.2
구분선	80	3	96.4
합계	23,594	167	99.3

제안된 방법의 성능을 정량적으로 평가한 결과는 <표 3>과 같다. 제안된 방법은 평균적으로 99% 이상의 객체를 식별한 반면에 수식과 테이블 객체에 대하여 비교적 낮은 식별률을 보였다. 이에 대한 보다 자세한 분석은 다음과 같다

5.1 오류 분석

일반적으로 비 텍스트 객체는 캡션 영역을 수반한다. 제안된 시스템은 캡션 영역을 비 텍스트 영역의 일부분으로 간주하고, 이를 포함하는 비 텍스트 객체를 추출한다. 실험 결과로 식별된 텍스트 및 비 텍스트 객체에 대한 오류 분석은 다음과 같다.

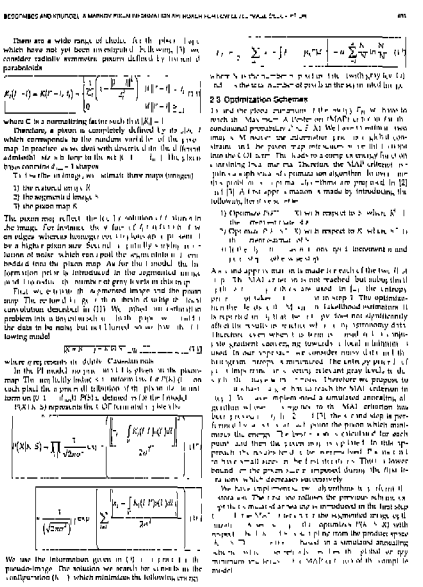
텍스트 라인의 경우, 식별 오류의 대부분은 수식 또는 캡션 영역으로 분류된 경우이다. 예를 들어, <그림 13(a)-①>과 같이, 영역 내부에 수식을 다수 포함하는 텍스트 라인은 수식 객체로 식별되었다. 한편 텍스트 라인 또는 수식 영역의 경우, 기하적인 특성만으로는 해당 영역의 유형을 결정할 수 없는 경우가 존재한다. 제안된 방법은 해당 영역의 유형을 식별하기 위하여 인접한 영역의 유형을 고려한다. 따라서 텍스트 라인과 수식 영역 사이에 위치하며 기하적인 특성이 모호한 텍스트 라인이 실제로 수식 영역에 보다 가까이 위치하여 수식으로 식별되는 경우가 발생하였다. 한편 <그림 13(a)-②>와 같이, 알고리즘 등을 기술하는 사각형 영역에 인접한 텍스트 라인의 경우, 비 텍스트 객체의 캡션 영역으로 분류되었다.

수식은 타 객체와 비교하여 식별률이 비교적 낮았다. 실제로 식별 오류의 대부분은 기하적인 속성이 텍스트 라인의 특성을 만족하여 텍스트 라인으로 분류되는 경우였다. 한편 수식 객체는 일반적으로 수식의 기하적인 특성을 만족하는 영역은 물론이고 만족하지 않는 영역을 다수 포함한다. 실제로 <그림 13(b)>는 수식에 해당되지만 기하적인 특성이 모호하며 인접한 텍스트 라인에 보다 가까이 위치하여 텍스트 라인으로 식별된 경우이다.

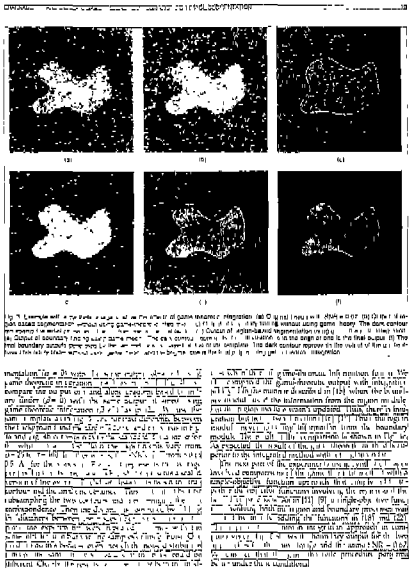
본 논문은 서로 교차하며 수평과 수직 방향의 직선 요소를 한 개 이상 포함하는 테이블을 대상으로 한다. 따라서 <그림 13(c)-①>과 같이, 수평 방향의 구분선만으로 구성된 연결되지 않은 테이블은 식별할 수 없다. 한편 <그림 13(c)-②>와 같이 테이블의 내용으로 문자가 아닌 크기가 큰 그림 등이 포함되는 경우, 드로잉으로 분류되었다.

제안된 방법은 그림 객체를 밀도 분포에 따라 드로잉과 이미지로 구분한다. 일반적으로 그림 객체는 이미지 또는 드로잉에 해당하는 여러 개의 작은 영역을 포함한다. 따라서 그림을 이미지와 드로잉의 두 가지 유형으로 구분하기 모호한 경우가 존재한다. 일반적으로 드로잉과 이미지로 식별된 영역에 각각 벡터라이징[38]과 이미지 압축 과정을 적용한다.

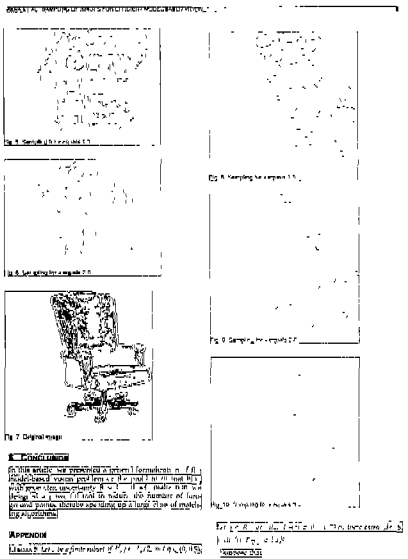
따라서 본 논문에서는 이미지와 드로잉으로 구분하기 모호한 객체의 경우, 벡터라이징에 적합한 영역은 드로잉으로, 압축이 보다 바람직한 영역은 이미지로 간주한 후에 실험하였다. 여기에서 식별 오류는 이미지와 드로잉 영역이 밀도 분포에 의하여 각각 드로잉과 이미지 객체로 분류된 경우이다. <그림 13(d)>는 드로잉으로 분류된 이미지 객체의 예이다. 그밖에 인접한 텍스트 라인에 중첩하는 구분선은 식별되지 않았으며, 또한 <그림 13(a)-②>와 같이 알고리즘 등을 기술하는 사각형 영역은 테이블 객체로 분류되었다.



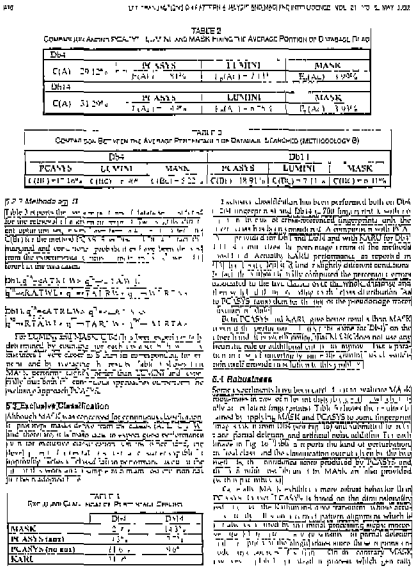
(a) 텍스트 영역으로부터 수식 객체의 식별 결과



(b) 이미지 객체와 텍스트 라인의 식별 결과

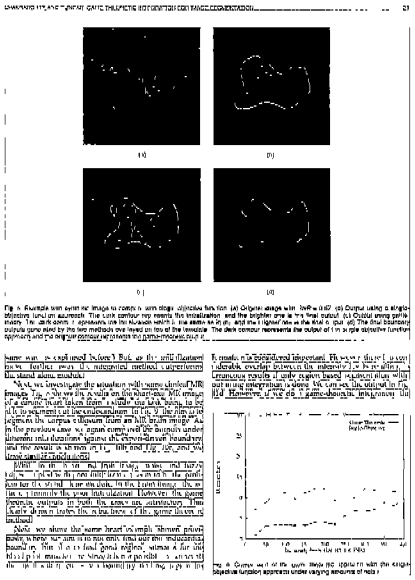


(c) 드로잉 객체와 텍스트 라인의 식별 결과

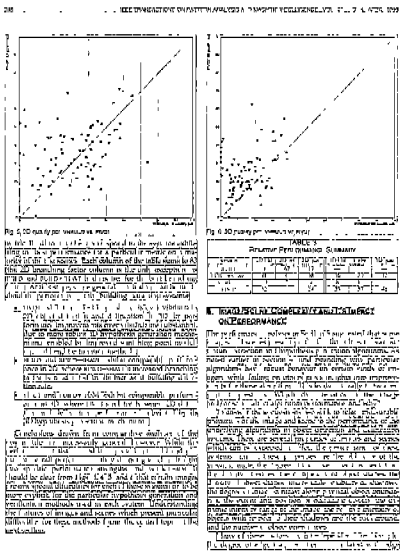


(d) 테이블 객체와 텍스트 라인의 식별 결과

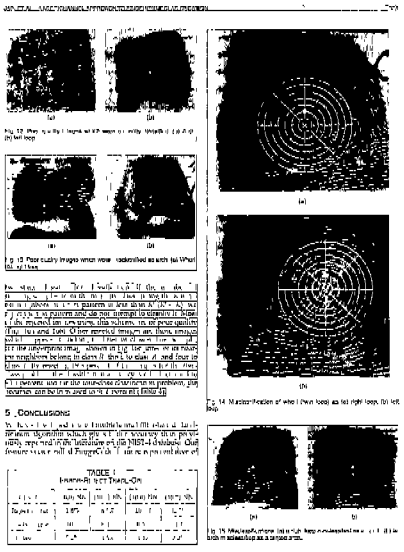
그림 10 실험 결과 (1)



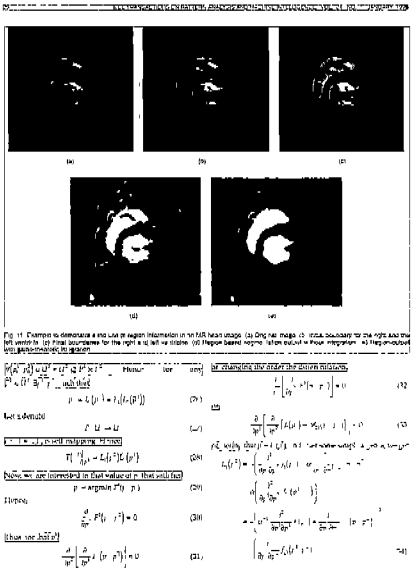
(a) 이미지, 드로잉, 텍스트 라인의 식별 결과



(b) 드로잉, 테이블, 텍스트 라인의 식별 결과

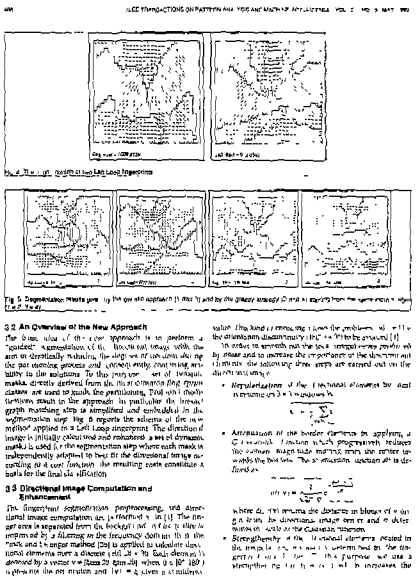


(c) 이미지, 테이블, 텍스트 라인의 식별 결과



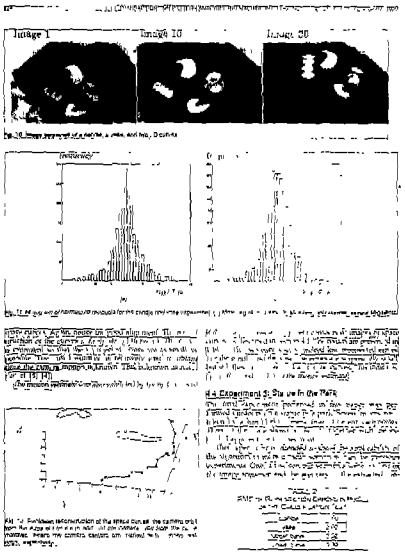
(d) 이미지와 텍스트 라인의 식별 결과

그림 11 실험 결과 (2)

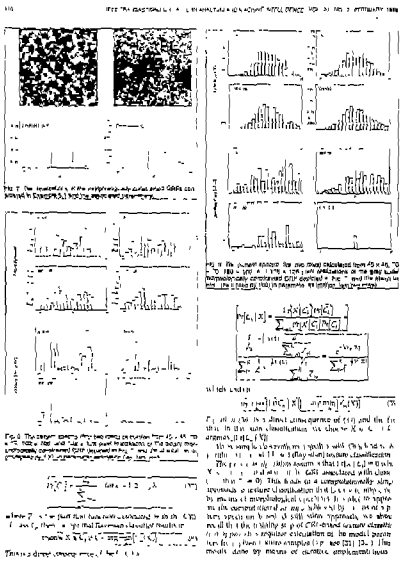


(a) 드로잉과 수식 객체의 식별 결과

(b) 테이블, 텍스트 라인 객체의 식별 결과



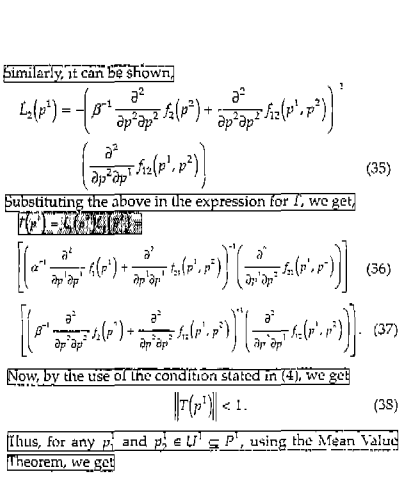
(c) 이미지,드로잉,테이블,텍스트 라인의 식별 결과



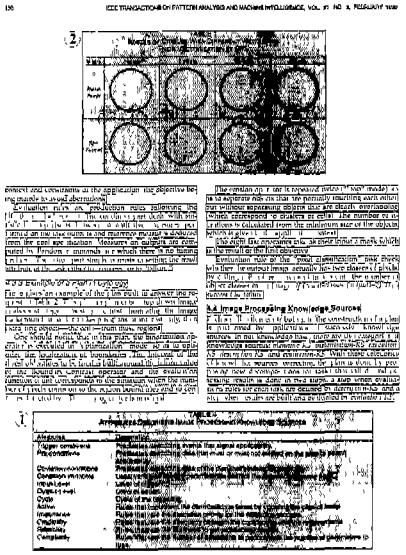
(d) 이미지, 드로잉, 수식의 식별 결과

그림 12 실험 결과 (3)

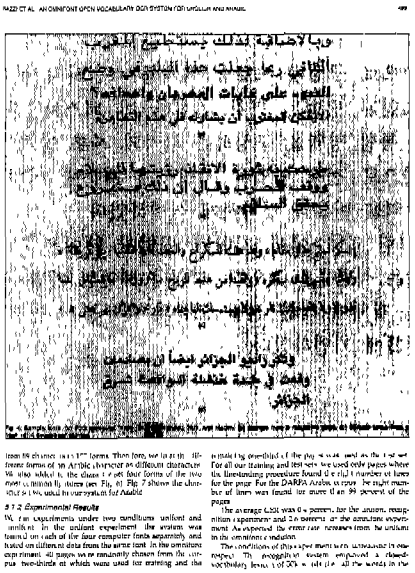
(a) 수식 또는 캡션으로 식별된 텍스트 라인의 예



(b) 수식이 텍스트 라인으로 잘못 인식된 경우



(c) 테이블의 식별 오류



(d) 드로잉으로 분류된 이미지 객체

그림 13 잘못된 기하적인 구조 분석 결과

5.2 관련 연구와의 비교

전술한 바와 같이 기존 연구의 대부분은 정교한 수준의 기하적인 구조 분석을 지원하지 않는다. 한편 최근에 Jain과 Yu[4]는 기하적인 구조 분석 방법을 제안하고 과학 기술 논문을 대상으로 체계적인 실험 결과를 제시하였다. 동일한 실험 데이타에 기반하지 않기 때문에 본 논문과 정량적으로 실험 결과를 비교할 수는 없다. 그러나 [4]가 과학 기술 논문 영상을 대상으로 시험 결과를 제시한 거의 유일한 연구이기에 <그림 14>와 같이 제안된 방법과 [4]의 실험 결과를 제시하고 이를 방법론 등의 정성적인 측면에서 비교하였다.

전술한 바와 같이 정교한 수준의 구조 분석을 위해서는 텍스트 영역으로부터 텍스트 라인과 수식을 정확히 분류하여야 한다. 그러나 [4]는 단순히 텍스트 라인을 추출하고 이를 통합하여 텍스트 영역을 추출하기 때문에 통합 및 분할된 텍스트 라인에 대한 처리를 지원하지 않는다. 또한 텍스트 영역의 구성 요소로써 텍스트 라인과 수식을 구별하지 않으며 수식을 크기에 따라 단순히 텍스트와 드로잉으로 분류하였다.

[4]는 흐리게 인쇄되어 스캐닝 결과 다수의 작은 영역으로 부서지거나 드문드문하게 분포하는 여러 개의 작은 영역으로 구성된 드로잉 객체를 식별하지 못한다. 그러나 제안된 방법은 먼저 인접한 비 텍스트 영역을 통합하고, 통합 영역을 구성하는 각각의 세부 영역의 유형을 기반으로 전체 영역을 식별하기 때문에 이미지와 드로잉 영역을 비교적 정확하게 식별하였다

특히 그림 <14>에서 제안된 방법이 테이블 객체에 대하여 낮은 식별률을 보였다. 그러나 제안된 방법과 [4]의 테이블 객체에 대한 식별 오류의 대부분은 연결되지 않은 테이블에 기인하였다.

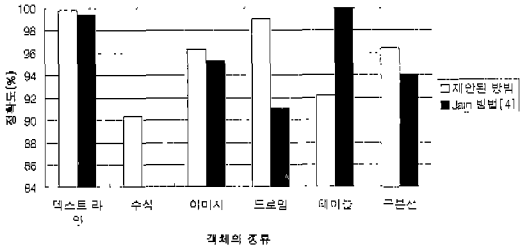


그림 14 제안된 방법과 [4]의 성능 비교

한편 제안된 방법의 성능을 상용 시스템과 비교하기 위하여 실험에 사용된 논문 영상을 상용 패키지인 TextBridge[39]와 InIt Reader[40]에 적용하였다. <표 4>와 같이 두 패키지 모두 텍스트 객체로서 텍스트 라인 그리고 비 텍스트 객체로서 그림과 테이블을 식별하였다. 그러나 그림 객체로부터 이미지와 드로잉을 식별하지는 않는다.

표 4 TextBridge와 InItReader의 성능 평가

S/W / 종류	TextBridge			InIT Reader		
	식별	비식별	정확도 (%)	식별	비식별	정확도 (%)
텍스트 라인	22,068	178	99.2	21,799	447	98.0
그림	233	125	65.1	235	123	65.6
테이블	64	13	83.1	8	69	10.4
합계	22,365	316	98.6	22,042	639	97.2

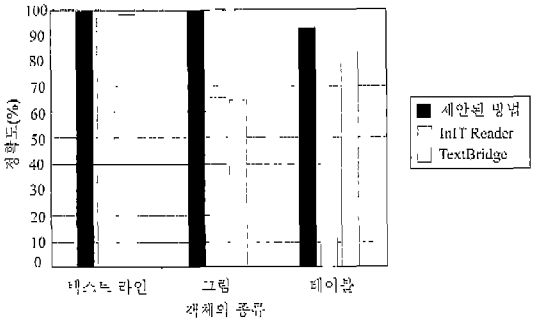


그림 15 제안된 방법과 상용 시스템의 성능 비교

TextBridge는 기본적인 수학 기호를 인식할 수 있으며 새로운 기호를 학습할 수 있는 기능을 제공한다. 그러나 출력 결과만을 가지고는 텍스트 영역으로부터 텍스트 라인과 수식을 정확히 구별하는지의 여부를 확인할 수는 없었다. 한편 InIT Reader는 수식을 식별하는 기능을 제공하지 않으며 테이블에 대하여 매우 낮은 식별률을 보였다. <그림 15>와 같이 제안된 방법은 두 상용 패키지보다 높은 식별률을 보였다.

6. 결론 및 향후 연구 방향

본 논문은 과학기술 논문을 대상으로 지식베이스에 기반한 영역 분할 및 식별 시스템을 제안하였다. 제안된 시스템의 지식베이스는 정교한 수준의 기하적인 구조 분석을 지원하기 위하여 논문 유형의 공통적인 특성은 물론이고 출판물 고유의 기하적인 특성에 관한 지식을 규칙 형태로 표현한다. 한편 실험 영상의 종류에 따라 구조 분석 과정의 제어 방법 역시 변경이 가능해야 하기 때문에 제안된 시스템의 추론 엔진 역시 계층적인 구조로 구성된 규칙에 기반하여 효율적인 구조 분석을

지원한다.

기하적인 구조 분석 결과로 식별된 영역과 레이아웃 객체 사이에 대응 관계가 높을수록 논리적인 구조 분석이 용이하다. 그러나 일반적으로 이들간에는 일-대-일의 대응 관계가 존재하지 않는다. 따라서 제안된 방법은 복합적인 기법을 적용하여 분할된 영역을 통합 및 재분할하면서 이미지, 드로잉, 그리고 테이블 등의 비 텍스트 객체는 물론이고 텍스트 영역으로부터 텍스트 라인과 수식 객체를 식별한다. 실험 결과, 제안된 방법은 실험 영상을 구성하는 99% 이상의 객체에 대한 기하적인 구조 분석에 성공하여, 기존 연구에 비교하여 정교한 수준의 성능을 보였다.

한편 보다 정교한 수준의 기하적인 구조 분석을 지원하기 위하여 개선이 요구되는 부분은 다음과 같다. 논리적인 구성 요소의 정확한 추출을 위하여 텍스트 라인의 정확한 추출이 요구된다. 따라서 각각 텍스트 라인과 수식으로 식별된 수식과 텍스트 라인에 문자 인식 등의 후처리 과정을 적용하여 오류를 수정할 필요가 있다. 또한 알고리즘 등을 설명하는 사각형 영역의 경우, 프레임을 구성하는 직선 요소의 기하적인 특성을 고려하여 이를 테이블 객체와 구별할 필요가 있다.

최근 들어 인터넷과 SGML/XML의 급속한 보급에 힘입어 전자도서관 구축에 관한 연구가 활발히 진행중이다. 따라서 기존의 종이 문서를 전자화하는 방법의 개발이 절실히 요구된다. 본 연구에서는 제안된 시스템의 결과로부터 논리적인 구성 요소를 추출하고 전자 문서를 자동으로 생성하기 위하여 논리적인 구조 분석 방법을 개발할 예정이다. 일반적으로 논리적인 구조를 포함하는 문서는 다수의 페이지를 포함한다. 따라서 문서 영상으로부터 논리적인 구조를 추출하기 위하여 다수의 페이지 영상에 대한 논리적인 구조 분석 방법이 요구된다.

참 고 문 헌

[1] L. O'Gorman and R. Kasturi, *Document Image Analysis*, IEEE Computer Society, 1995.

[2] G. Nagy, Twenty Years of Document Image Analysis in PAMI, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 38 ~ 62, Jan. 2000.

[3] A. Yamashita, T. Amano, Y. Hirayama, N. Itoh, S. Katho, T. Mano, and K. Toyokawa, A Document Recognition System and Its Application, *IBM Journal of Research and Development*, Vol. 40, No. 3, pp. 341-352, May 1996.

[4] A. K. Jain and B. Yu, Document Representation and Its Application to Page Decomposition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp. 294-308, Mar. 1998.

[5] R. M. Haralick, Document Image Understanding: Geometric and Logical Layout, *In Proc. Conf. Computer Vision and Pattern Recognition*, pp. 385-390, 1994.

[6] M. Worring and A. W. M. Smeulders, "Content based Internet Access to Paper Documents," *Int'l Journal on Document Analysis and Recognition*, Vol. 1, No. 4, pp. 209-220, 1999.

[7] Y. Y. Tang, S. W. Lee, and C. Y. Suen, Automatic Document Processing-A Survey, *Pattern Recognition*, Vol. 29, No. 12, pp.1931-1952, 1996.

[8] International Organization for Standardization, Information Processing-Text and Office Systems -Standard Generalized Markup Language (SGML), *ISO/IEC 8879*, 1986.

[9] World Wide Web Consortium, Extensible Markup Language (XML) 1.0, http://www.w3.org/TR/2000/REC-xml-20001006, 2000.

[10] O. Hitz, L. Robadey, and R. Ingold, Analysis of synthetic document images, *In Proc. Fifth Int'l Conf. Document Analysis and Recognition*, pp.374-377, Bangalore, India, Sep. 1999.

[11] P. Lefevre and F. Reynaud, ODIL: an SGML Description Language of the Layout Structure of Documents, *In Proc. Third Int'l Conf. Document Analysis and Recognition*, pp.480~487, 1995.

[12] T. Pavlidis and J. Zhou, Page Segmentation and Classification, *CVGIP: Graphical Models and Image Processing*, Vol. 54, No. 6, pp. 484-496, Nov. 1992.

[13] G. Nagy, S. Seth, and M. Viswanathan, A Prototype Document Image Analysis System for Technical Journals, *IEEE Computer*, Vol. 25, No. 7, pp. 10-22, July 1992.

[14] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan, Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals, *IEEE Trans. Pattern Analysis and Machine Intelligence*. Vol. 15, No. 7, pp. 737-747, July 1993.

[15] S. Tsujimoto and H. Asada, Major Components of a Complete Text Reading System, *Proc. IEEE*, Vol. 80, No. 7, pp. 1133-1149, July 1992.

[16] K. C. Fan, C. H. Liu, and Y. K. Wang, Segmentation and Classification of Mixed Text/Graphics/Image Documents, *Pattern Recognition Letters*, Vol. 15, pp.1201-1209, 1994.

[17] T. Saitoh, T. Yamaai, and M. Tachikawa, Document Image Segmentation and Layout Analysis, *IEICE Trans. Information and Systems*, Vol. E77-D, No.7, pp.778-784, July 1994.

[18] L. O'Gorman, The Document Spectrum for Page

Layout Analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 11, pp. 1162-1173, Nov. 1993.

[19] D. Wang and S. N. Srihari, Classification of Newspaper Image Blocks Using Texture Analysis, *Computer Vision, Graphics, and Image Processing*, Vol. 47, pp.327-352, 1989.

[20] F. Cesarini, M. Gori, S. Marinai, and G. Soda, "Structured Document Segmentation and Representation by the Modified X-Y tree," *In Proc. Fifth Int'l Conf. Document Analysis and Recognition*, pp. 563-566, IEEE Computer Society, Bangalore, India, Sep. 1999.

[21] L. A. Fletcher and R. Kasturi, A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 10, No. 6, pp. 910-918, Nov. 1988.

[22] A. Antonacopoulos and R. T. Ritchings, "Flexible Page Segmentation Using the Background," *In Proc. Int. Conf. Pattern Recognition*, 1994.

[23] A. K. Jain and S. Bhattacharjee, "Text Segmentation Using Gabor Filters for Automatic Document Processing," *Machine Vision and App.*, 5, pp.169-184,1992.

[24] K. Etemad, R. Chellappa, and D. Doermann, "Page Segmentation Using Wavelet Packets and Decision Integration," *In Proc. of Int. Conf. Pattern Recognition*, pp. 345-349, 1994.

[26] A. Antonacopoulos, Page Segmentation Using the Description of the Background, *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp. 350-369, June 1998

[27] F. Esposito, D. Malerba, and G. Semeraro, "A Knowledge-Based Approach to the Layout Analysis," *In Proc. Third Int'l Conf. Document Analysis and Recognition*, pp. 466-471, 1995.

[28] G. Nagy, J. Kanai, M. Krishnamoorthy, M. Thomas, and M. Viswanathan, "Two Complementary Techniques for Digitized Document Analysis," *In Proc. ACM Conf. Document Processing Systems*, pp. 169-176, 1988.

[29] A. Dengel and G. Barth, High Level Document Analysis Guided By Geometric Aspects, *Int'l Journal of Pattern Recognition and Artificial Intelligence*, Vol. 2, No. 4, pp. 641-655, 1988.

[30] A. Dengel, R. Bleisinger, R. Hoch, F. Fein, and F. Hnes, From Paper to Office Document Standard Representation, *IEEE Computer*, Vol. 25, No. 7, pp. 63-67, July 1992.

[31] J. Higashino, H. Fujisawa, Y. Nakano, and M. Ejiri, "A Knowledge-based Segmentation Method for Document Understanding," *In Proc. Eighth Int'l Conf. Pattern Recognition*, 745-748,1986.

[32] J. L. Fisher, S. C. Hinds, and D. P. D'Amato, "A Rule-based System for Document Image Segmentation," *In Proc. Tenth Int'l Conf. Pattern Recognition*, pp. 567-572, 1990.

[33] D. Niyogi and S. N. Srihari, An Integrated Approach to Document Decomposition and Structural Analysis, *Int'l Journal of Imaging Systems and Technology*, Vol. 7, pp. 330-342, 1996.

[34] J. Sauvola, M. Pietikainen, and M. Koivusaari, "Predictive Coding for Document Layout Characterization," *In Proc. Workshop on Document Image Analysis*, pp. 44-50, IEEE Computer Society, June 1997.

[35] A. M. Nazif and M. D. Levine, Low Level Image Segmentation: An Expert System, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 6, No. 5, pp. 555-577, Sep. 1984.

[36] J. K Ha, R. M. Haralick, and I. T, Phillips, "Document Page Decomposition By The Bounding-box Projection Technique," *In Proc. Third Int'l Conf. Document Analysis and Recognition*, Vol. 2, pp. 1119-1122, Montreal, Canada, Aug. 1995

[37] J. Kanai, "Text Line Extraction and Baseline Detection," *In Proc. Conf. Intelligent Text and Image Handling(RIAO'91)*, pp. 194-210, Barcelona, Spain, Apr. 1991.

[38] K. H. Lee, S. B. Cho, and Y. C. Choy, "Automated Vectorization of Cartographic Maps by a Knowledge-based System," *Engineering Applications of Artificial Intelligence*, Vol. 13, No. 2, pp. 165-178, Apr. 2000.

[39] *TextBridge Pro Millennium*. Peabody, MA: Scansoft Inc., 2000. http://www.scansoft.com.

[40] *InIT Reader*. Seoul, Korea: InIT Co., 2000. http://www.init.co.kr.

이 경 호
정보과학회논문지 : 소프트웨어 및 응용
제 28 권 제 7 호 참조

최 윤 철
정보과학회논문지 : 소프트웨어 및 응용
제 28 권 제 4 호 참조

조 성 배
정보과학회논문지 : 소프트웨어 및 응용
제 28 권 제 6 호 참조