문서 영상 내 테이블 영역에서의 단어 추출

정 창 부[†] · 김 수 형[††]

요 약

문서 영상은 문서 구조 분석을 통하여 텍스트, 그림, 테이블 등의 세부 영역으로 분할 및 분류되는데, 테이블 영역에 있는 단어는 다른 영역의 단어보다 의미가 있기 때문에 주제어 검색과 같은 응용 분야에서 중요한 역할을 한다. 본 논문에서는 문서 영상의 테이블 영역에 존재하는 문자 성분을 단어단위로 추출하는 방법을 제안한다.

테이블 영역에서의 단어 추출은 실질적으로 테이블을 구성하는 셀 영역에서 단어를 추출하는 것이기 때문에 정확한 셀 추출 과정이 필요하다. 셀 추출은 연결 요소를 분석하여 테이블 프레임을 찾아내고, 교차점 검출은 전체가 아닌 테이블 프레임에 대해서만 수행한다. 잘못 검출된 교차점은 이웃하는 교차점과의 관계를 이용하여 수정하고, 최종 교차점 정보를 이용하여 셀을 추출한다. 추출된 셀 내부에 있는 텍스트 영역은 셀 추출 과정에서 분석한 문자성분의 연결 요소 정보를 재사용하여 결정하고, 결정된 텍스트 영역은 투영 프로파일을 분석하여 문자열로 분리된다. 마지막으로 분리된 문자열에 대하여 갭 군집화와 특수 기호 검출을 수행함으로써 단어 분리를 수행한다. 제안 방법의 성능 평가를 위하여 한글 논문 영상으로부터 추출한 총 100개의 테이블 영상에 대해 실험한 결과, 99.16%의 단어 추출 성공률을 얻을 수 있었다.

키워드 : 문서 영상 검색, 문서 영상 전처리, OCR, 단어 분리

Word Extraction from Table Regions in Document Images

Chang Bu Jeong[†] · Soo Hyung Kim[††]

ABSTRACT

Document image is segmented and classified into text, picture, or table by a document layout analysis, and the words in table regions are significant for keyword spotting because they are more meaningful than the words in other regions. This paper proposes a method to extract words from table regions in document images.

As word extraction from table regions is practically regarded extracting words from cell regions composing the table, it is necessary to extract the cell correctly. In the cell extraction module, table frame is extracted first by analyzing connected components, and then the intersection points are extracted from the table frame. We modify the false intersections using the correlation between the neighboring intersections, and extract the cells using the information of intersections. Text regions in the individual cells are located by using the connected components information that was obtained during the cell extraction module, and they are segmented into text lines by using projection profiles. Finally we divide the segmented lines into words using gap clustering and special symbol detection. The experiment performed on 100 table images that are extracted from Korean documents, and shows 99.16% accuracy of word extraction.

Key Words : Document Image Retrieval, Document Image Preprocessing, OCR, Word Segmentation

1. 서 론

컴퓨터 기술과 인터넷 환경의 지속적인 발달로 문서 영상의 변환, 저장 및 전송 등의 처리들이 효과적으로 수행될 수 있게 되었다. 현재 인터넷의 원문 검색 서비스나 디지털 도서관에서는 일반적으로 문서를 영상의 형태로 제공하는데, 이런 문서에 대한 정보 검색은 영상 기반이 아닌 기존에 사용하던 텍스트 기반의 검색 엔진으로는 불가능하다. 따라서 문서 영상에 대한 검색은 사전에 텍스트로 입력되는 도서 목록의 정보(제목, 저자, 초록 등)만을 고려하여 수행되고, 사용자가 도서 목록에 없는 정보를 이용하여 관련 문서를 검색하려면 문서 영상의 일부 또는 전부를 다운로드하여 확인하여야 한다. 이러한 단점을 보완하기 위하여 문서 영상 처리(DIP : Document Image Processing) 기술을 사용하고 있으며, 이에 대한 연구는 접근 방식이 다른 두 가지 방법으로 진행되고 있다. 첫 번째는 광학 문자 인식(OCR : Optical Character Recognition)을 이용한 전문 검색(Full Text Searching)이며, 두 번째는 문서 영상의 자동 색인(Indexing)을 통한 키워드 탐색(Keyword Spotting)기법이다. 전자는 많은 연구에도 불구하고 OCR의 결과에 대한 매뉴얼

※ 이 논문은 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임. (R05-2003-000-10396-0)
† 준 회 원 : 호남대학교 정보통신대학 인터넷소프트웨어학과 전임강사
†† 정 회 원 : 전남대학교 공과대학 전자컴퓨터정보통신공학부 부교수
논문접수 : 2005년 1월 18일, 심사완료 : 2005년 6월 8일

한 수정과 검증 등의 후처리가 필요하며 그 비용이 경제적이지 못하다는 단점이 있다. 반면에 후자는 단어 영상의 특징 정보를 이용하기 때문에 문자 분할의 오류를 고려하지 않아도 되며, 문서의 언어와 상관없이 검색이 가능하기 때문에 문서 영상 검색 방법으로 연구가 활발히 진행되고 있다[1-6].

문서 영상의 자동 색인을 통한 키워드 탐색 방법에서 사용되는 문서 영상 처리 기술은 문서영상의 전처리(기울어짐 교정, 잡음 제거 등)와 문서 구조 분석, 텍스트 영역의 단어 단위 분할 등으로 나열할 수 있다. 최근에 발표된 Jeong 등의 시스템[7]은 키워드 탐색을 위한 문서 영상의 전처리 시스템을 제안하고 있다. 제안된 시스템은 문서 구조 분석을 이용하여 문서 영상을 텍스트 영역과 비텍스트 영역(테이블, 그림 등)으로 분할하고, 분할된 텍스트 영역에 대해서 단어 분리를 수행하여 단어단위 영상을 추출하였다. 그러나 제안된 시스템은 텍스트 영역에 대해서만 단어 분리를 수행하기 때문에 시스템 성능의 한계를 지니고 있다. 즉 비텍스트 영역 중, 테이블처럼 키워드 탐색의 대상으로 적합한 단어영상이 있는 영역에 대해서도 단어단위 영상을 추출할 필요가 있다. 하지만 문서 영상내의 테이블에 대한 연구는 폼 문서 영상에 대한 연구의 일부로 인식되어 별도로 발표된 사례가 거의 없다.

폼 문서 영상에 대한 기존 연구를 살펴보면, 테이블의 선 성분(수평선과 수직선 등)을 추출하여 폼의 구조나 셀을 분석하는 방법, 선 성분을 제거하고 문자 성분만을 추출하는 방법 등이 제안되었다. 그러나 이러한 방법은 선 성분만을 고려하거나 문자 성분의 추출만 고려하기 때문에 단어단위의 영상을 추출하는 데에는 적합하지 않다[8-16].

본 논문에서는 문서 영상 내의 테이블 영역으로부터 단어 단위 영상을 추출하는 방법을 제안한다. 제안한 방법은 테이블의 물리적 구조를 파악하여 셀의 위치를 구하고, 추출된 셀의 내부에 존재하는 문자열을 단어단위로 분리하는 두 단계의 과정으로 수행된다. 즉, 첫 번째 단계는 테이블 영상으로부터 테이블 프레임에 해당하는 가장 큰 연결 요소를 찾고, 그 연결 요소에 마스크 연산을 이용하여 셀을 구성하는 교차점의 정보(형태, 위치)를 추출한다. 추출된 교차점 정보는 이웃하는 교차점들을 분석하여 올바르게 수정되고, 테이블의 셀 위치를 결정하는데 사용된다. 두 번째 단계에서는 추출된 셀 정보를 이용하여 셀 내부에 있는 문자열을 추출하고, 갭 군집화와 특수 기호 검출을 이용하여 단어 분리를 실행함으로써 최종적인 단어단위 영상을 추출한다. 각 단계에 대한 실험을 통하여 본 논문에서 제안한 방법이 우수함을 입증하였다.

2. 관련 연구

문서 영상 내의 테이블 영역에 관한 연구는 선(수평선과 수직선, 대각선 등) 성분으로 구성된 테이블 분석과 선 성분 없이 내용이 가로와 세로로 정렬되어 테이블처럼 표현되는 문서 영상의 인식 등으로 분류된다[17, 18]. 본 논문의 처리 대상은 선 성분으로 구성된 테이블 영상으로써 폼 문서 영상의 인식에 대한 연구와 유사한 문제를 다루고 있지만, 기존 연구의 대부분이 폼 문서 영상에 관련된 것이다. 두 분야에서 공통적으로 다루어지는 문제는 셀과 관련 있는 선 성분을 분석하여 테이블이나 폼의 구조를 파악하는 것이다. 문서 영상 내의 테이블 영역에 관한 연구는 테이블을 구성하는 선 성분을 분석하여 테이블 영상을 벡터화하고 문자 성분을 분리하는 것이었다. 한편 폼 문서 영상의 인식에 관한 연구는 추출된 폼의 구조를 이용하여 데이터베이스에 미리 저장된 샘플용 폼 문서 중에서 일치하는 것을 선택하고 샘플용 폼 문서와 다른 내용(차후 기입된 것)을 추출하는 방법 등으로 진행되었다. 본 절에서는 테이블 영상의 선 성분 및 셀 추출 관점에서 관련 연구를 기술하고, 추가로 텍스트 영역의 단어 분리에 대한 연구를 설명한다.

Watanabe 등[9]은 셀의 좌상단 모서리를 사용하여 테이블을 표현하였다. 우선 수직선과 수평선 성분을 추출하기 위하여 두 개의 필터를 이용하고, 좌상단 모서리를 추출하기 위하여 또 다른 두 개의 필터를 사용한다. 이 방법은 좌상단 모서리만을 교차점으로 이용하기 때문에 잘못된 교차점 정보에 대한 해결책이 없으며 선 성분의 변형이 있으면 필터의 효과가 반감될 수 있다.

Kim 등[10]에서는 8 방향 체인 코드를 이용한 방법과 히스토그램을 이용한 방법, 런길이를 이용한 방법 등의 세 가지 방법을 적용하여 문서 영상의 테이블을 벡터화하였다. 8 방향 체인 코드를 이용한 방법은 세선화 과정이 필요하므로 처리 시간이 많이 소요되는 단점이 있고, 히스토그램을 이용한 방법은 문자와 테이블의 선 성분을 흑화소의 누적 정도로 테이블을 구성하는 것으로써 복잡한 구조의 테이블 영상에서는 정확률이 떨어진다. 마지막으로 런길이를 이용한 방법은 흑화소의 연속된 수를 이용하여 문자와 선 성분을 구별하는 것인데 선 성분에 굴곡이 생기면 구별이 힘들어진다.

Lee 등[13]은 선 성분을 추출하기 위하여 Opening 연산과 조건부 Closing 연산을 수행하였다. 그리고 교차점을 찾기 위하여 [9]에서 제안한 방법을 사용하였는데, 이미 모폴로지 연산으로 인해 선 성분에 대해서만 교차점을 찾으므로 [9]의 실험보다는 좋은 결과를 보였다.

Taylor 등[8]에서는 9×9의 필터들을 이용하여 일차적으로 모서리에 위치하는 네 종류의 기본 교차점(ㄱ, ㄱ, ㄴ, ㄴ)을 찾고, 나머지 부류의 확장 교차점(ㅏ, ㅜ, ㅓ, ㅗ, ㅗ)을 찾기 위하여 이미 찾아진 기본 교차점을 적당한 방법으로 조합하였다. 그러나 Arias 등[11]은 기본 교차점과 확장 교차점의 관계를 계층 구조로 정의하여 확장 교차점의 추출에서 소요되는 계산량을 감소시켰다.

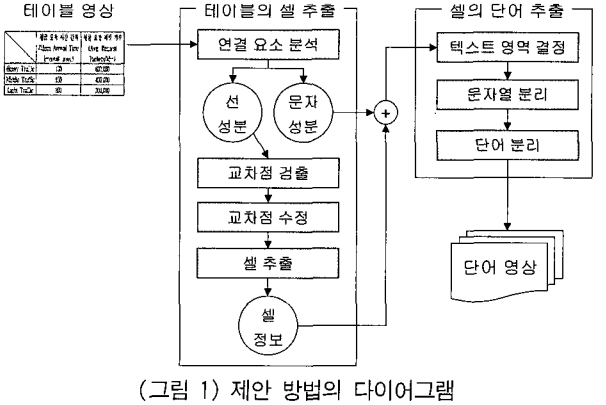
앞서 설명한 방법들은 추출된 교차점에 대한 신뢰도가 떨어진다. 즉, 잘못된 교차점 정보에 대한 해결책이 없으므로, Neves 등[14]에서는 [11]에서 사용한 필터 대신 모폴로지 연산을 수행하여 9개의 교차점 정보(위치와 형태)를 구하고,

0번의 가상 교차점을 추가하였다. 그리고 이웃하는 교차점과의 관계를 분석하여 잘못된 교차점을 찾아내어 수정함으로써 셀 추출의 정확률을 향상시켰다.

위의 연구들이 테이블 영상에서 셀을 추출하는 문제를 다루었다면, [7]은 텍스트 영상을 단어단위 영상으로 분리하는 방법을 제안하였다. 제안된 방법은 한글 또는 영문이 포함된 임의의 텍스트 영역을 문자열단위(text lines)로 분리하고, 각각의 문자열을 단어단위(words)로 분리한다. 기존의 유사한 연구들이 상향식 또는 하향식 접근 방법으로 분류되는 반면, 이 논문은 투영 프로파일 방법과 연결 요소 분석 방법을 혼합하여 사용한다. 즉, 텍스트 영역에 대해 투영 프로파일을 분석하여 문자열단위로 분리하고, 각 문자열들은 연결 요소 분석을 이용하여 띄어쓰기 단위인 단어단위로 분리한다. 문자열 분리에서는 수평방향 투영 프로파일(HPP: Horizontal Projection Profile)을 계산하여 분리 지점을 구하고, 재귀적 투영 프로파일 분석 방법을 추가하여 정확도를 개선한다. 단어 분리에서는 분리된 문자열에 대하여 연결 요소 분석을 수행하고, 수직으로 병합된 연결 요소의 최소 인접 사각형간 거리 값에 대해 계층적 군집화 기법을 사용하여 단어 분리 지점을 계산한다. 또한, 띄어 쓰지는 않았지만 추가적인 단어 분리를 가능하게 하는 특수 기호들을 검출하여 더욱 정확한 단어 분리를 수행한다.

3. 제안 방법

테이블 영상에서의 단어 분리는 테이블의 물리적 구조를 파악하여 셀의 위치를 구하고, 추출된 셀의 내부에 존재하는 문자열을 단어단위로 분리하는 두 단계의 과정으로 수행된다. 첫 번째 단계는 테이블 영상으로부터 마스크 연산을 이용하여 셀을 구성하는 교차점의 정보(형태, 위치)를 추출하고, 이웃하는 교차점들을 분석하여 교차점의 정보를 수정하고 최종적으로 셀의 위치를 결정한다. 두 번째 단계에서는 추출된 셀 정보를 이용하여 셀 내부에 있는 문자열을 추출하고 갭 군집화와 특수 기호 검출을 이용하여 단어 분리를 실행한다. (그림 1)은 제안 방법의 단계별 수행 과정을 도식화한 것이다.



(그림 1) 제안 방법의 다이어그램

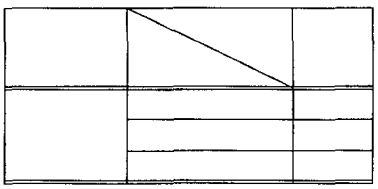
3.1 테이블의 셀 추출

3.1.1 교차점 검출

테이블 영상에는 일반적으로 (그림 2)의 (a)와 같이 테이블 프레임에 해당하는 선 성분(수직선, 수평선, 대각선 등)과 문자 성분(문자, 숫자, 특수 기호 등)이 같이 존재한다. 따라서 선 성분으로 구성된 테이블의 셀을 추출하는 과정에서는 불필요한 문자 성분을 제외시켜야 하기 때문에, 연결 요소 분석을 통하여 테이블의 프레임에 해당하는 연결 요소만 고려한다. 즉, 8 방향 연결 요소 분석에서 폭(또는 높이)이 가장 큰 연결 요소를 테이블의 프레임으로 결정한다(그림 2의 (b)). 여기서 구해지는 문자 성분과 선 성분에 대한 연결 요소 분석 결과는 셀 추출 후의 단어 분리에서도 이용되는데, (그림 2)의 (a)처럼 문자 성분과 선 성분(주로 대각선)이 혼재한 셀(1행 2열)에서 단어를 분리할 때에 선 성분을 제외하는 효과를 얻을 수 있다.

메시지크기	메시지 발생률 / 방안	0.01
16플릿	Probe 기법	0
	Lopez(Th=16)	0
	Lopez(Th=32)	0

(a) 원본 테이블 영상

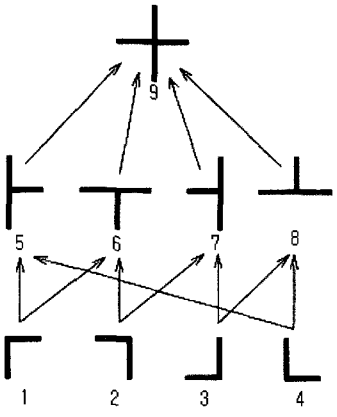


(b) 테이블의 프레임에 해당하는 연결 요소

(그림 2) 연결 요소 분석을 이용하여 테이블의 프레임에 해당하는 연결 요소 결정

셀을 구성하는 교차점 검출은 (그림 2)의 (b)처럼 문자 성분이 제외된 선 성분을 대상으로 교차점의 위치와 유형을 결정한다. 교차점은 독립적인 9가지 유형으로 분류할 수 있지만, [11]에서 제안된 (그림 3)과 같은 계층적인 방법을 이용하여 교차점의 유형을 결정한다면 계산량을 감소시킬 수 있다. 우선 교차점 검출은 (그림 3)의 하위 계층에 속하는 1~4번 교차점 유형에 해당하는 4개의 마스크 연산자를 이용한다. [11]에서는 9×9 형태의 마스크 연산자를 이용하지만, 테이블의 선 성분이 가는 경우에는 비틀림과 같은 변형이 조금만 있어도 교차점 검출이 어려워질 수 있기 때문에 본 논문에서는 7×7 형태의 마스크 연산자를 이용한다. 마스크 연산자의 축소로 교차점이 아닌 점이 교차점으로 검출될 가능성이 높아지지만 이미 문자 성분이 제외되었기 때문에 우려하지 않아도 된다.

우선 테이블 프레임의 검은 화소에 대하여 4개의 마스크 연산자를 검사하고 만족하는 마스크 연산자의 번호를 교차점 유형으로 결정한다. 만약 두 개의 마스크 연산자를 만족하면 두 개의 교차점 유형이 결합된 상위 계층의 교차점 유

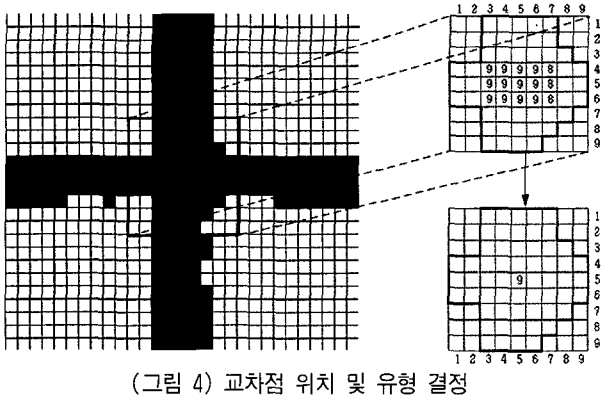


(그림 3) 교차점 유형에 대한 계층도

형으로 결정하고, 세 개 이상의 마스크 연산자를 만족하면 최상위인 9번 유형의 교차점으로 결정한다. 예를 들어 7번 유형의 교차점인 경우에는 하위 계층의 2번과 3번 유형의 마스크 연산자를 만족하므로 그들이 결합한 상위계층의 7번 유형으로 결정될 것이다.

테이블의 수직선과 수평선이 교차 또는 접촉할 때 하나의 교차점이 발생하는데, 위의 마스크 연산자를 이용하면 테이블의 선의 굵기에 따라 여러 개의 교차점 성분이 검출될 수 있다. 그러므로 여러 개의 교차점 성분을 분석하여 한 개의 교차점 성분으로 집약해야 한다. (그림 4)처럼 다수의 교차점들이 무리로 검출되었을 때, 최종 교차점의 위치는 그 무리의 중앙으로, 유형은 다수결의 원칙으로 결정되는데 동수일 경우에는 높은 값의 유형으로 결정한다.

정상적인 셀은 4개의 교차점 성분, 즉 위치와 유형이 적절하게 조합되어야 하기 때문에, 검출된 교차점들의 수직 위치를 고려한 연결리스트를 구성한다. (그림 5)는 교차점의 유형과 위치(수평, 수직)로 구성된 연결리스트를 표 형식으로 나타낸 것으로써, (1, 1) 요소인 "1-(3, 5)"는 영상의 (3, 5)에 위치한 점이 유형 1의 교차점임을 의미한다. 일반적인 테이블의 경우에는 교차점의 수직 위치만 고려하여 구성된 연결리스트를 이용하여 셀을 추출할 수 있지만, (그림 2)처럼 병합된 셀들이 있는 테이블의 경우에는 연결리스트의 수정이 불가피하다. 그러므로 연결리스트에서 상하로 이웃하



(그림 4) 교차점 위치 및 유형 결정

	1	2	3	4
1	1-(3, 5)	6-(203, 5)	6-(473, 4)	2-(604, 4)
2	5-(3, 139)	9-(203, 138)	9-(474, 138)	7-(604, 138)
3	5-(3, 143)	9-(203, 143)	9-(474, 143)	7-(604, 143)
4	5-(203, 190)	9-(474, 190)	7-(604, 190)	
5	5-(203, 243)	9-(474, 242)	7-(604, 242)	
6	5-(3, 296)	9-(203, 296)	9-(474, 295)	7-(604, 295)
7	4-(2, 300)	8-(202, 300)	8-(473, 300)	3-(604, 300)

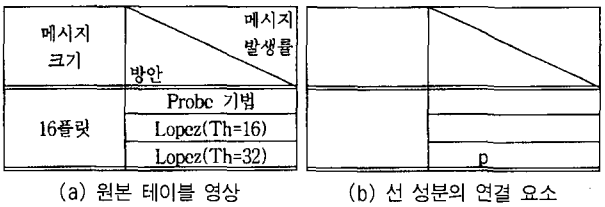
(그림 5) 교차점들의 수직 위치를 고려한 연결리스트

	1	2	3	4
1	1-(3, 5)	6-(203, 5)	6-(473, 4)	2-(604, 4)
2	5-(3,139)	9-(203, 138)	9-(474, 138)	7-(604, 138)
3	5-(3,143)	9-(203, 143)	9-(474, 143)	7-(604, 143)
4	0-(3,190)	5-(203, 190)	9-(474, 190)	7-(604, 190)
5	0-(3,243)	5-(203, 243)	9-(474, 242)	7-(604, 242)
6	5-(3,296)	9-(203, 296)	9-(474, 295)	7-(604, 295)
7	4-(2,300)	5-(202, 300)	5-(473, 300)	3-(604, 300)

(그림 6) 가상의 교차점이 삽입되어 수정된 연결리스트

는 교차점들의 수평 위치의 차이가 임계치 이상이면 병합된 셀이 있는 것으로 간주하고 가상의 교차점(0번 유형)을 삽입한다. 수평으로 병합된 셀이 있는 경우도 이와 같은 방법으로 처리가 가능하다. (그림 6)은 가상의 교차점을 삽입하여 (그림 5)의 연결리스트를 수정한 것이다.

선 성분이 명확한 테이블일 경우에는 지금까지 처리만으로도 교차점 검출을 올바르게 수행할 수 있다. 그러나 문서 영상의 획득 과정이나 기울어짐 교정 등에서 문서 영상의 변질 등이 발생하여 테이블의 선 성분에 왜곡이 생기고 교차점 검출이 실패하거나 교차점이 잘못 검출(거짓 교차점)될 수 있기 때문에, 정확한 셀 추출을 위해서는 이웃하는 교차점들과 관련성을 분석하여 올바른 교차점으로 수정해야 한다. 교차점 수정은 [14]에서 제안한 방법을 응용한 것으로써, 교차점 유형에 따른 이웃 교차점으로의 부적합 리스트를 사전 정보로 이용하여 잘못 검출된 교차점을 찾아내고, 적절한 교차점으로 수정하는 것이다. (그림 7)과 같이 문자



(a) 원본 테이블 영상 (b) 선 성분의 연결 요소

(그림 7) 문자 성분과 선 성분의 접촉으로 거짓 교차점이 발생하는 예제 영상

	1	2	3	4
1	1-(3, 5)	6-(203, 5)	0-(273, 4)	2-(473, 4)
2	5-(3, 139)	9-(203, 138)	0-(274, 138)	7-(474, 138)
3	5-(3, 143)	9-(203, 143)	0-(274, 143)	7-(474, 143)
4	0-(3, 190)	5-(203, 190)	0-(274, 190)	7-(474, 190)
5	0-(3, 243)	5-(203, 243)	0-(274, 242)	7-(474, 242)
6	5-(3, 296)	9-(203, 296)	8-(274, 295)	7-(474, 295)
7	4-(2, 300)	8-(202, 300)	0-(273, 300)	3-(473, 300)

(a) 거짓 교차점이 발생한 연결리스트

	1	2	3	4
1	1-(3, 5)	6-(203, 5)	0-(273, 4)	2-(473, 4)
2	5-(3, 139)	9-(203, 138)	0-(274, 138)	7-(474, 138)
3	5-(3, 143)	9-(203, 143)	0-(274, 143)	7-(474, 143)
4	0-(3, 190)	5-(203, 190)	0-(274, 190)	7-(474, 190)
5	0-(3, 243)	5-(203, 243)	0-(274, 242)	7-(474, 242)
6	5-(3, 296)	9-(203, 296)	0-(274, 295)	7-(474, 295)
7	4-(2, 300)	8-(202, 300)	0-(273, 300)	3-(473, 300)

(b) 거짓 교차점이 수정된 연결리스트

(그림 8) 거짓 교차점이 발생한 연결리스트에 대한 교차점 수정 결과

메시지크기	메시지 발생률 / 방안	0.01
16플릿	Probe 기법	0
	Lopez(Th=16)	0
	Lopez(Th=32)	0

(a) 이중선으로 불필요한 셀이 생기는 영상

a.	(a1) 인간은 누구나 언어를 사용할 줄 안다. (a2) 언어 사용이야말로 인간과 다른 동물을 구별할 수 있게 하는 가장 큰 특징이다.	*CDS(a1,a2)* = 0.6837
b.	(b1) 인간은 누구나 언어를 사용할 줄 안다. (b2) 언어는 인간 사회에서 어떤 개념을 특정한 소리를 사용하여 지시하자는 약속이다.	*CDS(b1,b2)* = 0.5418

(b) 완전 개방형 테이블

CN	*RD*	*TT*	*SD*
c1	10:00:09	2	10:00:07
c3	10:00:12	2	10:00:10
c4	10:00:14	2	10:00:12
c6	10:00:20	2	10:00:18
c8	10:00:19	2	10:00:17

(c) 좌우 개방형 테이블

(그림 9) 테이블 영상에서의 셀 추출 결과

성분과 선 성분의 접촉이 발생하면, 접촉한 문자 성분은 선 성분으로 포함되고 교차점 검출에도 영향을 주어서 (그림 8)의 (a)처럼 거짓 교차점이 검출된다. 이러한 거짓 교차점은 교차점 수정 작업을 통하여 (그림 8)의 (b)처럼 가상 교차점으로 변경된다.

그러나 [14]에서는 테이블에서의 교차점 위치는 고려하지 않고 이웃하는 교차점 유형간의 관계만 분석하기 때문에 불필요한 계산량이 줄이기 위하여 다음과 같은 교차점의 위치를 고려한 제약조건을 추가한다. 테이블의 모서리에 해당하는 4개의 교차점은 각각 1~4번 유형의 교차점이어야 하며, 가장자리에 위치한 교차점들은 상하좌우의 위치에 따라 각각 5~8번 유형의 교차점과 0번의 가상 교차점이 가능하다. 또한 그 밖의 위치, 즉 내부에 있는 교차점의 후보 리스트에서 1~4번 유형의 교차점은 부적합으로 제약을 둔다. 그러나 (그림 9)의 (b)나 (c)처럼 외곽에 존재하는 선 성분이 일부 또는 전부 없는 개방형 테이블의 경우는 [14]의 교차점 수정으로도 처리가 불가능하다. 이런 경우에는 테이블의 모서리에 해당하는 교차점들이 앞서 분석한 선 성분의 연결요소에 대한 최외곽 사각형(bounding box)의 꼭지점과 유사한 위치에 있는가를 검사하여, 유사한 위치에 있지 않으면 1~4번 유형의 교차점을 추가한다.

최종적으로 교차점 수정을 마친 연결리스트로부터 유형 0번의 가상 교차점을 제외한 인접한 동일 라인에 있는 2개의 교차점과 동일 컬럼에 있는 2개의 교차점, 즉 4개의 교차점 조합으로 하나의 셀을 결정한다. 그러나 (그림 9)의 (a)처럼 이중선 등으로 생기는 셀은 높이나 폭이 임계치 이하인지 확인하여 제거를 해야 한다. (그림 9)는 다양한 테이블 영상에서 셀을 추출한 결과이다.

3.2 셀에서의 단어 분리

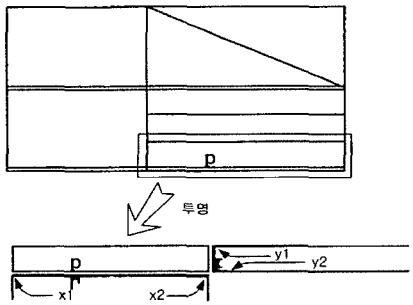
셀에서의 단어 분리는 우선 추출된 셀 정보를 이용하여 셀 내부에 존재하는 문자 성분의 범위, 즉 텍스트 영역을 구한다. 텍스트 영역은 문자열 단위로 분할되고, 분할된 문자열은 갭의 군집화를 이용하여 단어로 분리되며 특수 기호 검출을 통하여 추가 단어 분리가 수행된다.

3.2.1 텍스트 영역 결정

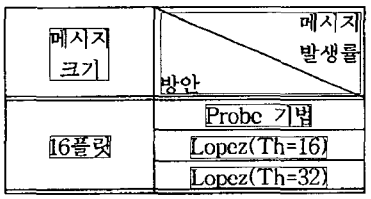
셀 내부에 있는 문자 성분의 분석은 일반적으로 셀의 범위에 있는 검은 화소의 분석(투영 프로파일, 연결 요소 분석 등)을 통해 가능하다. 그러나 이미 테이블의 선 성분을 분석할 때 제외되었던 문자 성분들의 정보(연결 요소)를 이용하면 보다 빠르고 쉽게 처리할 수 있다. 즉, 문자 성분의 연결요소가 해당 셀의 내부에 있는지를 검사함으로써 셀 내부의 텍스트 영역을 결정할 수 있고, 셀 내부에 존재할 수 있는 선 성분(대각선, 선의 번짐 등)의 영향을 감소 또는 제거도 가능하다. 그러나 3.1절에서 설명하였듯이 문자 성분과 선 성분의 접촉이 발생하면 문자 성분은 선 성분으로 포함되기 때문에, 추출된 셀 정보를 이용하여 이러한 문자 성분을 선 성분과 분리하고 문자 성분으로 추가해야 한다.

우선 교차점의 수정 단계에서 검출된 거짓 교차점이 선 성분에 위치하면 이러한 접촉이 발생한 것으로 결정한다.

접촉이 발생한 셀은 거짓 교차점의 위치와 유형을 분석하여 알아내고, 테이블 프레임으로부터 해당 셀의 투영 프로파일을 분석하여 셀 내부 영역을 파악한다(그림 10). 즉, 각 투영 프로파일의 시점과 종점에서부터 중심방향으로 이웃하는 두 값을 비교해 나갈 때, 현재의 값이 이전 값의 10%보다 적으면 셀 내부 영역의 정보(x1, x2, y1, y2)로 결정한다. 이렇게 구해진 셀 내부 영역에 존재하는 테이블 프레임의 연결 요소를 텍스트 영역의 문자 성분에 추가한다. (그림 10)은 셀 내부 영역을 결정하는 방법을 설명하는데, 최종적으로 셀 내부 영역에 있는 'p'에 대한 연결 요소가 문자 성분으로 추가된다. 선 성분과의 접촉으로 누락된 문자 성분까지 추가되어 셀 내부의 문자 성분이 결정되면 이들을 포함하는 최외곽 사각형을 해당 셀의 텍스트 영역으로 설정한다. (그림 11)은 선 성분에 접촉한 문자 성분이 올바르게 텍스트 영역으로 포함됨을 보여준다.



(그림 10) 셀 내부 영역 결정하는 방법



(그림 11) 선 성분에 접촉한 문자 성분이 있는 테이블에서 텍스트 영역 검출

3.2.2 문자열 분리 및 단어 분리

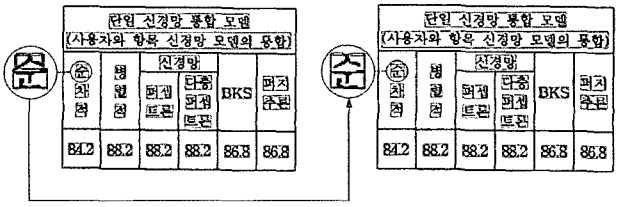
텍스트 영역에 대한 단어 분리는 [7]의 단어 분리 알고리즘을 이용하지만, 단어들이 일반 문서의 텍스트 영역이 아닌 테이블의 셀 내부에 있기 때문에 부가적인 제약 조건이 필요하다. 예를 들어, 셀에 한 문자만 존재하고 문자에 대한 수평 히스토그램에서 갭이 발생하게 되면 그 문자는 두 개의 문자열로 분리될 것이다. 또한 한 단어의 문자들이 셀이라는 환경 때문에 여러 단어로 분리될 수도 있다. 그러므로 제안한 방법은 [7]의 단어 분리 알고리즘이 테이블의 셀에 있는 문자 성분을 올바르게 처리할 수 있도록 수정하여 응용하였다.

텍스트 영역에서의 문자열 분리는 연결 요소의 최외곽 사각형을 수평으로 투영하여 히스토그램을 구하고, 히스토그램에서 투영값이 0인 부분을 문자열간의 갭으로 결정한다. 그러나 셀의 문자열이 한 글자로만 이루어진 경우에는 문자열 내의 투영값이 0인 지점이 분석되어 하나의 문자열이 두 개 이상의 문자열로 분리될 수 있다. 이렇게 과다분리된 문자열을 병합하는 후처리는 다른 문자열 갭의 정보를 이용한 아래의 조건에 의하여 실행된다. 아래의 조건에서 N은 해당 셀의 문자열 개수, $LineGap(i)$은 i 번째 문자열 갭의 크기를 의미하고 $AverageOfLineGaps$는 테이블에 존재하는 문자열 갭의 평균값을 나타낸다.

$$N \geq 3\text{일때}, LineGap(i) < \frac{\sum_{i=1}^{N} LineGap(i)}{2\times(N-1)}, (i \neq j)$$

$$N < 3\text{일때}, LineGap(i) < \frac{AverageOfLineGaps}{2}$$

(그림 12)의 (a)에서는 '순'이라는 한 글자로 구성된 문자열에서 투영값이 0인 곳의 발생으로 문자열이 과다분리 되었지만, (b)에서는 동일 셀의 다른 문자열 간격을 참조하는 위의 첫 번째 조건을 만족하므로 과다분리된 두 개의 문자열을 병합하였다.



(a) 문자열의 과다분리가 발생하는 영상 (b) 수정된 문자열

(그림 12) 문자열의 과다분리가 발생하여 수정되는 예제

분리된 문자열은 갭의 군집화와 특수 기호 검출을 통하여 단어단위로 최종 분리한다. 그러나 문자열 분리 과정에서의 예외 사항처럼 하나의 단어를 구성하는 문자들이 셀이라는 공간에서 단어사이의 간격만큼 상당한 간격으로 위치할 수 있다. [7]의 단어 분리 방법은 이런 문자들을 독립된 단어로 분리하므로 분리된 단어를 한 단어로 병합하는 후처리가 필요하다. 이런 후처리는 동일 문자열에 있는 모든 단어들이 한 문자로만 구성될 가능성이 거의 없다는 것을 이용한다. 즉, 단어 분리에서 추출된 단어영상의 폭과 높이를 분석하여 문자의 수를 추정하고, 분리된 단어들의 추정된 문자수가 모두 1이면 분리된 단어들을 한 단어를 구성하는 문자들로 간주하고 하나의 단어로 통합한다.

(그림 13)의 (a)에서 왼쪽에 위치한 셀들의 문자열은 의미상 모두 한 단어로 구성되어 있다. 이런 문자열들에 대하여 갭의 군집화를 이용한 단어 분리를 수행하면 다른 문자열들은 한 단어로 분리하지만, "이미징"의 단어는 세 문자로 분리하게 된다. 왜냐하면 다른 문자열에서는 갭이 하나씩만 구해지기 때문에 재 군집화로 인하여 단어 간의 갭으로 결정되었지만, "이미징" 문자열에서의 문자간의 갭들은 크기가 비교적 차이가 나서 두 군집으로 분류된다. 결과적으로 크

기가 큰 문자간의 갭은 단어간의 갭으로 오분류된다. 그러나 (그림 13)의 (b)에서는 분리된 단어("이", "미", "지")에서 추정된 문자수가 모두 1이므로, 한 단어가 오분리된 것으로 간주하고 하나의 단어로 통합한 결과이다.

급 지	0.3
노 출	0.5
이 미 징	1.8
인 화	0.7
출 지	0.3
세 정	0.2

(a) 단어의 문자들이 분리

급 지	0.3
노 출	0.5
이 미 징	1.8
인 화	0.7
출 지	0.3
세 정	0.2

(b) 후처리에 의해 수정

(그림 13) 단어 분리의 후처리 예

4. 실험 및 결과

4.1 실험 데이터

제안 방법의 성능 평가를 위하여 총 100개의 테이블 영상을 사용하였다. 테이블 영상은 정보과학회에서 제공하는 원문 검색 서비스를 이용하여 다운받은 논문 영상을 [7]의 시스템으로 전처리하고 테이블 영역만을 저장한 것으로써, 300dpi의 이진 영상이며 크기는 849×117 픽셀부터 1500×1770 픽셀까지 다양하다. <표 1>은 테이블의 셀 구성이나 테두리의 선 형태에 따라 분류하여 실험 데이터를 분석한 것이고, (그림 14)는 테이블의 분류별로 예제 영상을 보여주고 있다. 실험은 Pentium-4 2.0 GHz PC 상에서 수행되었다.

<표 1> 실험 데이터의 구성

분 류	개 수
일반적인 셀 구성의 테이블	60
병합된 셀이 있는 테이블	20
하나의 셀로 구성된 테이블	10
테두리 선이 일부 또는 전부 없는 테이블	10
합 계	100

	평균 도착 시간 간격 (Mean Arrival Time Interval: μsec)	평균 요청 패킷 개수 (Avg. Request Packets/Min)
Heavy Traffic	100	600,000
Middle Traffic	150	400,000
Light Traffic	200	300,000

(a) 일반적인 셀 구성의 테이블

Category	Function Name	Attribute Name	Bank A		Bank B		Bank C	
			Uses	Type	Uses	Type	Uses	Type
여신	계좌 계설	상담원정보	○	Text	○	Text		
		현금	○	Num			○	Num
		수수료	○	Num	○	Num		
		인출한도	○	Num			○	Num
		패스워드	○	Text, Num	○	Num	○	Text, Num
		예금과목	○	Text	○	Text	○	Text
		계좌번호	○	Num	○	Text	○	Num

(b) 병합된 셀이 있는 테이블

1. if User_Model is dislike and Movie_Model is dislike then dislike (CF:0.89)
2. if User_Model is like and Movie_Model is dislike then like (CF:0.88)
3. if User_Model is dislike and Movie_Model is like then like (CF:0.75)
4. if User_Model is like and Movie_Model is like then like (CF:0.93)
5. if User_Model is middle and Movie_Model is like then like (CF:0.67)
6. if User_Model is middle and Movie_Model is dislike then dislike (CF:0.67)
7. if User_Model is middle and Movie_Model is middle then like (CF:0.63)
8. if User_Model is dislike and Movie_Model is middle then dislike (CF:0.6)
9. if User_Model is like and Movie_Model is middle then like (CF:0.91)

(c) 하나의 셀로 구성된 테이블

CN	*RD*	*TT*	*SD*
c1	10:00:09	2	10:00:07
c3	10:00:12	2	10:00:10
c4	10:00:14	2	10:00:12
c6	10:00:20	2	10:00:18
c8	10:00:19	2	10:00:17

(d) 테두리 선이 일부 또는 전체가 없는 테이블

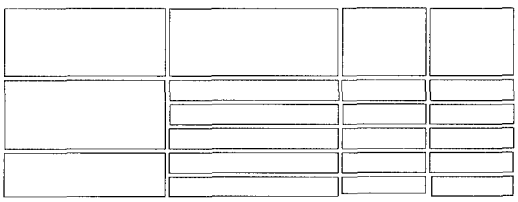
(그림 14) 실험 데이터의 분류별 예제 영상

4.2 성능 평가

테이블 영상에서의 셀 추출 방법과 전체 단어 분리 방법에 대한 성능은 정확도와 수행 시간의 측정으로 평가하였다. 제안된 셀 추출 방법은 실험 데이터인 100개의 테이블 영상에 있는 2,313개의 셀을 모두 성공적으로 추출하였다. 3절에서의 (그림 9)와 같이 여러 형태로 구성된 테이블의 셀들을 정상적으로 추출하였고, 또한 (그림 15)처럼 문자 성분과 선 성분이 접촉하여 거짓 교차점이 검출되는 테이블에서도 교차점 수정을 통하여 올바른 셀을 추출하였다.

트래픽 형태	메시지 발생률 / 메시지크기	0.01	0.1
균등분포	16 플릿	2.202	2.076
	32 플릿	2.391	2.146
	64 플릿	2.564	2.432
perfect-shuffle	16 플릿	1.947	1.95
	32 플릿	1.973	2.181

(a) 원본 테이블 영상



(b) 검출된 셀 영역 표시

(그림 15) 문자 성분과 수직선 성분이 접촉한 테이블에서의 셀 추출 결과

<표 2>는 추출된 셀 정보를 이용하여 단어 분리를 수행한 결과로써, 갭 정보만을 이용하여 단어 분리를 수행한 결과와 단어 사이에 있는 특수 기호를 검출하여 추가적인 단

어 분리까지 수행한 성능을 보여준다. 100개의 실험 영상에 존재하는 단어의 개수가 4,547개였는데, 제안 방법은 갭의 군집화를 이용하여 94.59%의 단어를 성공적으로 분리하였으며 단어 사이에 존재하는 특수 기호까지 검출하면 4.57%의 단어를 추가로 분리할 수 있다. 즉, 총 99.16%의 단어를 성공적으로 분리하였다.

〈표 2〉 단어 분리 결과

		갭 정보에 기반한 단어 분리		특수 기호 검출 후 단어 분리	
영상 개수	단어 개수	성공	실패	성공	실패
100	4547	4301(94.59%)	246(5.41%)	4509(99.16%)	38(0.84%)

(그림 16)은 (그림 15)의 테이블 영상에서 단어 분리를 수행한 결과이다. (그림 16)에서 (a)는 갭의 군집화만을 이용하여 단어를 분리한 것으로써 "perfect-shuffle"의 영상처럼 단어 사이에 존재하는 특수 기호 '-' 때문에 단어 분리를 실패했지만, (b)는 그와 같은 특수 기호를 검출하여 "perfect"와 "shuffle"의 단어로 분리되었다. 또한 (그림 16)에서는 선 성분(수직선)과 문자 성분(숫자)의 접촉이 발생하였지만 단어 분리가 성공적으로 수행되었다. 100개의 실험 영상 중, 이런 접촉은 5개의 영상에서 총 7번의 접촉이 발생하였으며 모두 성공적으로 처리되었다.

트래픽 형태	메시지 발생률 / 메시지크기	0.01	0.1
균등분포	16 플릿	2.202	2.076
	32 플릿	2.391	2.146
	64 플릿	2.564	2.432
perfect-shuffle	16 플릿	1.947	1.95
	32 플릿	1.973	2.181

(a) 갭 군집화만을 이용한 단어 분리

트래픽 형태	메시지 발생률 / 메시지크기	0.01	0.1
균등분포	16 플릿	2.202	2.076
	32 플릿	2.391	2.146
	64 플릿	2.564	2.432
perfect-shuffle	16 플릿	1.947	1.95
	32 플릿	1.973	2.181

(b) 특수 기호 검출을 통한 후처리까지 수행

(그림 16) 단어 분리를 수행한 결과

평균적으로 615,000개의 픽셀로 구성된 테이블 영상에 대하여 제안 방법은 셀 추출에 1.399초, 단어 분리에 0.071초로 총 1.47초의 시간이 소요되었다. 셀 추출 단계의 수행 시간이 단어 분리 단계의 수행 시간보다 상당히 오래 걸린 것은 단어 분리에서 필요한 연결 요소 분석을 셀 추출 단계에서 미리 수행하기 때문이다. 그러므로 관련 연구의 셀 추출 방법과 단순하게 수행 시간을 비교하는 것은 제안 방법의 성능을 평가하는 방법으로 적합하지 않다.

<표 3>은 [7]에서 제안된 시스템과 본 논문의 제안 방법을 결합하여 실험한 결과로써, 50개의 한글 영상에 대해서 [7]의 시스템은 90.84%의 성능을 보인 반면에, 본 논문의 제안 방법을 추가하면 93.50%의 향상된 성능을 얻었다. 성능 향상의 주요 원인은 [7]의 시스템에서 처리하지 못한 테이블 영역의 단어를 본 논문의 제안 방법으로 추출이 가능하기 때문이다. 또한 실험 대상인 50개의 한글 영상 중에서 테이블 영역이 포함된 14개 영상들에 대해서만 실험을 하였을 때, [7]의 시스템은 테이블 영역에 존재하는 다수의 단어를 분리하지 못하여 80.50%의 성능을 보였지만 본 논문에서 제안한 방법을 추가 실행하면 성능을 93.62%로 향상시킬 수 있었다.

〈표 3〉 [7]의 시스템에 제안 방법을 추가하여 실험한 결과

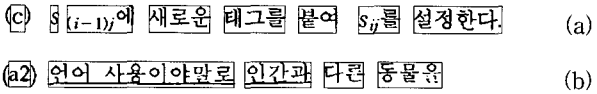
영상번호	단어 수				추출된 단어 수				정확도(%)	
	텍스트 영역	비텍스트 영역		전체 영역	텍스트 영역	비텍스트 영역		전체 영역	추가 전 시스템	추가 후 시스템
		테이블	그림			테이블	그림			
1	474	0	0	474	472	0	0	472	99.58	99.58
2	707	0	0	707	707	0	0	707	100.00	100.00
3	581	0	35	616	581	0	0	581	100.00	94.32
4	570	0	27	597	569	0	0	569	99.82	95.31
…	…									
35	374	208	0	582	353	208	0	561	60.65	96.39
36	546	74	0	620	537	71	0	608	86.61	98.06
…	…									
46	432	43	19	494	420	43	0	463	85.02	93.72
47	400	169	0	569	387	169	0	556	68.01	97.72
48	491	0	10	501	486	0	0	486	98.98	97.01
49	477	0	15	492	470	0	0	470	98.53	95.53
50	402	53	45	500	378	53	0	431	75.60	86.20
									90.84	93.50

4.3 오류 분석

제안 방법의 실패는 갭 군집화를 이용한 단어 분리 과정에서 실패하는 경우와 단어를 구성하는 한 문자가 특수 기호로 잘못 검출되어서 한 단어가 분리되는 경우, 단어 사이에 있는 특수 기호를 검출하지 못하여 분리하지 못하는 경우 등 세 가지 유형으로 분석이 가능하다.

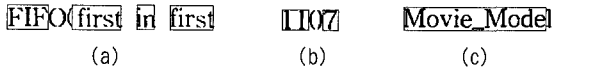
총 4,547개의 단어 중 38개의 단어를 분리 실패하였는데, 그 중 15개의 단어가 첫 번째 유형의 원인으로 실패하였다. 첫 번째 유형의 원인은 재 군집화의 실패와 밑줄의 폰트 속성 때문에 여러 개의 단어가 하나로 처리되기 때문이다. (그림 17)의 (a)와 같은 오류는 재 군집화의 실패로 인한 오류로서, "$s_{(i-1)j}$"에서의 갭이 다른 문자들 간의 갭보다 상대적으로 커서 단어 간의 갭으로 처리되었다. 즉, 개별 문자가 아닌 단어 단위의 출력을 위해 추가의 단어분리를 고려하는 과정에서 발생된 오류로서, 본 논문의 주 관심사는 아니며 참고문헌 [7]의 알고리즘을 개선함으로써 해결이 가능할 것이다. 한편 (그림 17)의 (b)와 같은 오류를 해결하기 위해서는 단어 영상에서 밑줄(underline)의 존재유무를 판단한다. 밑줄의 유무를 판단하기 위해서는 모폴로지 연산이나 허프

변환 등의 방법을 고려할 수 있지만, 분리된 모든 단어 영상에 대한 처리 시간의 증가에 비해 성능 개선의 효과가 미비하기 때문에 이런 종류의 오류는 허용하였다.

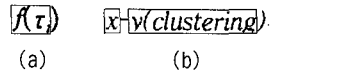


(그림 17) 단어 분리의 첫 번째 오류 유형의 예제 영상

두 번째 유형의 오류는 특수 기호와 유사한 문자 영상이 훼손되어 특수 기호로 오검출되는 것으로써 9개의 단어가 분리에 실패하였다. (그림 18)에서 (a)와 (b)는 영문자 'O'나 숫자 '0'의 영상이 중간 연결 요소가 끊어짐으로 인해 특수 기호 '('나 ')'로, (c)는 영문자 '1'의 획 일부가 손실되어 특수 기호 '|'처럼 오검출되는 예를 보여준다. 마지막으로 세 번째 유형의 오류는 특수 기호의 영상이 훼손되거나 이탤릭체 등의 폰트 속성 때문에 특수 기호의 형태적 특징이 변하여 특수 기호로 검출하지 못하는 것이다. 이런 원인으로 단어 사이에 있는 7개의 특수 기호를 검출하지 못하였고, 14개의 단어를 분리하는데 실패하였다. (그림 19)에서 (a)는 특수 기호 '('가 이웃하는 영문자 'f'와 접촉하여 실패하였고, (b)는 특수 기호의 폰트 속성이 이탤릭체여서 검출에 실패하였다.



(그림 18) 단어 분리의 두 번째 오류 유형의 예제 영상



(그림 19) 단어 분리의 세 번째 오류 유형의 예제 영상

위의 그림과 같이 두 번째와 세 번째 유형의 오류는 해당 문자 영상이 영상 획득 과정 등에서 훼손되어 일반 문자가 특수 기호의 형태적 특징을 가지거나 특수 기호가 형태적 특징을 잃어버리는 경우이다. 이러한 오류는 원본 영상의 훼손에 어떻게 대처해야 하는 가에 관한 문제로서 영상처리 과정에서의 개선 등과 같은 다른 측면의 해결방법이 시도되어야 할 것이다.

5. 결 론

본 논문에서는 테이블 영상에서 셀을 추출하고 추출된 셀 내부의 문자 성분을 단어단위로 분리하는 방법을 제안하였다. 테이블 영상으로부터 셀을 추출하기 위하여 연결 요소를 분석하여 테이블 프레임을 찾고, 테이블 프레임에 대해서만 교차점 검출을 수행하였다. 이렇게 검출된 교차점은 이웃하는 교차점과의 관계를 분석함으로써 올바른 교차점인가를 확인하고 수정하여 셀을 추출한다. 추출된 셀 내부에 있는 텍스트 영역은 이미 분석했던 문자 성분에 대한 연결 요소 정보를 재사용하여 결정하고, 결정된 텍스트 영역은 문자열로 분리된다. 마지막으로 분리된 문자열에 대하여 갭 군집화와 특수 기호 검출을 이용하여 단어 분리를 수행한다. 원문 검색 서비스에서 300dpi의 해상도로 제공되는 논문 영상으로부터 추출한 총 100개의 테이블 영상에 대해 실험한 결과, 99.16%의 단어 분리 성공률을 얻을 수 있었다.

본 논문에서 제안한 테이블 영상의 단어단위 영상 분할 방법은 전자 도서관 등에서 사용되는 문서 영상의 자동 색인 및 주제어 기반 검색을 위한 시스템의 핵심 요소 기술로 활용될 수 있다.

참 고 문 헌

[1] D. Doermann, "The Retrieval of Document Images: A Brief Survey," *Computer Vision and Image Understanding*, Vol.70, No.3, pp.287-298, 1998.

[2] S. Marinai, E. Marino, G. Soda, "Indexing and Retrieval of Words in Old Documents," *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Vol.1, pp.223-227, Aug., 2003.

[3] I. S. Oh, Y. S. Choi, J. H. Yang, S. H. Kim, "A Keyword Spotting System of Korean Document Images," *Lecture Notes in Computer Science 2555*(Proc. 5th International Conference on Asian Digital Libraries, Singapore), pp.530, Dec., 2002.

[4] S. Marinai, E. Marino, F. Cesarini, G. Soda, "A General System for the Retrieval of Document Images from Digital Libraries," *Proceedings of the First International Workshop on Document Image Analysis for Libraries*, pp.150-173, Jan., 2004.

[5] Y. Lu, L. Zhang, C. L. Tan, "Retrieving Imaged Documents in Digital Libraries Based on Word Image Coding," *Proceedings of the First International Workshop on Document Image Analysis for Libraries*, pp.174-187, Jan., 2004.

[6] Y. Lu, L. Zhang, C. L. Tan, "A Search Engine for Imaged Documents in PDF Files," *Proceedings of the 27th annual International Conference on Research and development in Information Retrieval*, pp.536-537, July, 2004.

[7] C. B. Jeong, S. H. Kim, "A Document Image Preprocessing System for Keyword Spotting," *Lecture Notes in Computer Science 3334*(Proc. International Conference on Asian Digital Libraries 2004, Shanghai, China), pp.440-443, Dec., 2004.

[8] S. Taylor, R. Fritzson, J. Pastor, "Extraction of Data from Pre-printed Forms," *Machine Vision and Applications*, Vol.5, No.3, pp.211-222, 1992.

[9] T. Watanabe, Q. Luo, N. Sugie, "Layout Recognition of Multi-Kinds of Table-Form Documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.17,

pp.432-445, April, 1995.

[10] W. S. Kim, J. B. Shim, Y. B. Park, K. A. Moon, S. Y. Ji, "Research on the Table Vectorization in the Document Image," *Journal of Korea Information Processing Society*, Vol.3, No.5, pp.1147-1159, Aug., 1996(text in Korean).

[11] J. F. Arias, R. Kasturi, "Efficient Extraction of Primitives from Line Drawings Composed of Horizontal and Vertical Lines," *Machine Vision and Applications archive*, Vol.10, pp.214-221, Dec., 1997.

[12] L. Y. Tseng, R. C. Chen, "Recognition and Data Extraction of Form Documents based on Three Types of Line Segments," *Pattern Recognition*, Vol.31, No.10, pp.1525-1540, 1998.

[13] S. H. Lee, K. M. Lee, "Table Extraction and Analysis Algorithm from Document Images," *Hongik Journal of Science and Technology*, Vol.2, pp.129-138, Dec., 1998.

[14] L. A. P. Neves, J. Facon, "Methodology of Automatic Extraction of Table-Form Cells," *XIII Brazilian Symposium on Computer Graphics and Image Processing* (SIBGRAPI' 00), pp.15-21, Oct., 2000.

[15] K. C. Fan, M. L. Chang, "Form Document Identification using Line Structure based Features," *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Vol.1, pp.704-709, Sept., 2001.

[16] D. Xi, S. W. Lee, "Reference Line Extraction from Form Documents with Complicated Backgrounds," *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Vol.2, pp.1080-1084, Aug., 2003.

[17] J. H. Shamilian, H. S. Baird, T. L. Wood, "A Retargetable Table Reader," *Proceedings of the 4th International Conference on Document Analysis and Recognition*, pp.158-163, Aug., 1997.

[18] D. Lopresti, G. Nagy, "A Tabular Survey of Automated Table Processing," *Lecture Notes In Computer Science*, Vol.1941, pp.93-120, 1999.

정 창 부

e-mail : cbjeong@honam.ac.kr

1999년 호남대학교 컴퓨터공학과(학사)

2001년 전남대학교 전산통계학과(이학석사)

2005년 전남대학교 전산학과 박사수료

2005년~현재 호남대학교 정보통신대학 인터넷소프트웨어학과 전임강사

관심분야 : 문서영상전처리, 패턴인식

김 수 형

e-mail : shkim@chonnam.chonnam.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1993년 한국과학기술원 전산학과(공학박사)

1993년~1996년 삼성전자 멀티미디어연구소 선임연구원

1997년~현재 전남대학교 전자컴퓨터정보통신공학부 부교수

관심분야 : 패턴인식, 문서영상전처리, 웨이블렛변환, 유비쿼터스 컴퓨팅