투영 프로파일, Gap 및 특수 기호를 이용한 텍스트 영역의 어절 단위 분할

(Decomposition of a Text Block into Words Using Projection Profiles, Gaps and Special Symbols)

정 창 부 † 김 수 형 ††
(Chang Bu Jeong) (Soo Hyung Kim)

요 약 본 논문에서는 인쇄체 텍스트 영상에 대한 문자열 분리 방법과 어절 분리 방법을 제안한다. 문자열 분리 방법은 수평 투영 프로파일을 분석하고, 오분리된 문자열에 대하여 재귀적 투영 프로파일(Recursive Projection Profile) 분석을 수행한다. 어절 단위 분리는 문자열에 대한 연결요소 분석을 통하여 gap을 검출한 후, 계층적 군집화 기법에 의해 어절과 어절 사이에 존재하는 gap을 판별하여 어절 분리점을 결정한다. 또한 어절과 어절 사이에 존재하는 특수기호를 검출하여 어절 분리점을 추가하기 위해서, 연결요소의 종횡비와 골격선(skeleton)의 형태적 특징을 고려한다. 제안 방법의 성능 평가를 위하여 총 84개의 텍스트 영상에 대하여 실험하였고, 국내 상용 OCR 소프트웨어인 아르미와 성능 비교하였다. 최종 어절 분리에 대하여 제안 방법과 아르미가 각각 99.92%와 97.58%의 성능으로 측정됨으로써 제안 방법이 아르미에 비해 우수함을 보였다.

키워드 : 키워드 탐색, 어절 분리, Gap 클러스터링, 특수기호 검출

Abstract This paper proposes a method for line and word segmentation for machine-printed text blocks. To separate a text region into the unit of lines, it analyses the horizontal projection profile and performs a recursive projection profile cut method. In the word segmentation, between-word gaps are identified by a hierarchical clustering method after finding gaps in the text line by using a connected component analysis. In addition, a special symbol detection technique is applied to find two types of special symbols lying between words using their morphologic features. An experiment with 84 text regions from English and Korean documents shows that the proposed method achieves 99.92% accuracy of word segmentation, while a commercial OCR software named Armi 6.0 Pro^{TM} has 97.58% accuracy.

Key words : Keyword Spotting, Word segmentation, Gap clustering, Special symbol detection

1. 서 론

컴퓨터의 등장으로 인한 사무 자동화에도 불구하고 전 세계적으로 인쇄 문서의 발생량이 날로 증가하고 있다. 그러나 이러한 인쇄 문서상의 정보는 보관, 검색, 재생 및 수정에 대한 처리가 어려워 컴퓨터로의 입력에 대한 요구가 증대되고 있다. 대용량 문서의 컴퓨터 데이터베이스화에 따르는 기존의 고비용 수작업 입력의 대안으로 두 가지 방안이 연구되고 있다. 첫 번째는 광학 문자 인식(OCR: Optical Character Recognition)을 이용한 자동 변환(Image-to-Text Conversion) 방식이며, 두 번째는 문서 영상의 자동 색인(Indexing)을 통한 키워드 탐색(Keyword Spotting) 기법이다. 이들 두 가지는 서로 장단점이 있으며 각기 다른 응용을 목적으로 연구·개발되고 있다[1,2].

광학 문자 인식 및 키워드 탐색 방법의 공통적인 요소기술로는 문서영상의 전처리, 문서구조 분석, 텍스트 영역의 어절 및 문자단위 분할 등과 같은 전처리 단계가 있다. 이중에서도 텍스트 영역의 분할은 문서 인식 과정에서 인식률에 영향을 미치는 중요한 단계임에도 불구하고 양적, 질적으로 실제적인 연구가 부족한 실정이다. 더군다나 대부분의 문서 영상 분할은 OCR 패키

· 본 연구는 한국과학재단 지역대학우수과학자 지원연구(R-05-2003-000-10396-0) 지원으로 수행되었음
† 학생회원 : 전남대학교 전산학과
cbjeong@iip.chonnam.ac.kr
†† 종신회원 : 전남대학교 컴퓨터정보학부 부교수
shkim@chonnam.chonnam.ac.kr
논문접수 : 2003년 11월 24일
심사완료 : 2004년 6월 17일

지의 활용을 위한 문자 분리에만 집중되었고[3-8], 어절 분리에 대한 연구가 발표된 사례는 극히 드물다[9,10]. 문서 영상의 분할에 관한 연구는 크게 상향식(bottom-up)과 하향식(top-down) 접근 방법을 고려할 수 있다. 상향식 방법은 가장 기본이 되는 화소 단위나 작은 데이터에서 시작하여 유사한 특성을 갖는 부분을 점차 큰 단위로 병합해 나가는 방법으로, 대표적인 방법은 연결요소(CC: Connected Component) 분석 방법이 있다. 이 접근 방법은 많은 계산량에 의한 처리시간 비용이 크고, 초기의 잘못된 확장에 의한 오류를 유발할 가능성이 있다는 단점을 가지고 있다. 하향식 접근 방법은 상향식과는 반대 방향으로 분할을 진행하며, 런 길이 평활화(run-length smoothing) 방법과 투영 프로파일(projection profile) 방법이 대표적이다. 이 방법은 문서를 빠르게 처리할 수 있는 반면에, 사각형의 블록(block)들로 구성되지 않은 복잡한 구조에 대해서는 효율적이지 못한 단점을 안고 있다[11-14].

본 논문에서는 인쇄체 텍스트 영역 영상을 띄어쓰기 단위인 어절 영상의 집합으로 분리하는 방법을 제안한다. 즉, 임의의 문서 영상에 대해 전처리 및 문서구조 분석 단계를 적용하고, 이때 출력되는 각각의 텍스트 영역에 대한 영상을 제안된 시스템의 입력으로 간주한다. 제안된 시스템은 한글 또는 영문이 포함된 임의의 텍스트 영역을 문자열(text line) 단위로 분리하고, 각각의 문자열을 어절(word) 단위로 분리하여 어절 영상의 집합을 출력한다. 기존의 유사한 연구들이 상향식 또는 하향식 접근 방법으로 분류되는 반면, 제안 방법은 투영 프로파일 방법과 연결요소 분석 방법을 혼합하여 사용한다. 즉, 텍스트 영역에 대해 계층적인 방법으로 문자열을 분리하고, 각 문자열들을 띄어쓰기 단위인 어절들의 집합으로 분할한다. 문자열 분리에서는 수평방향 투영 프로파일(HPP: Horizontal Projection Profile)을 계산하여 분리 지점을 구하고, 재귀적 투영 프로파일 분석 방법을 추가하여 정확도를 개선한다. 어절 분리에서는 분리된 문자열에 대하여 연결요소 분석을 수행하고, 수직으로 병합된 연결요소의 최소 인접 사각형간 거리 값에 대해 계층적 군집화 기법을 사용하여 어절 분리 지점을 계산한다. 또한, 띄어 쓰지는 않았지만 어절의 분리를 가능하게 하는 특수기호들을 검출하여 어절 분리점을 수정함으로써 보다 정확한 어절 분리를 수행한다. 성능 평가를 위한 실험은 86개의 한글 및 영문 텍스트 영역에 대하여 어절 분리 정확도를 계산하고, 아르미와 실험 결과를 비교한다. 제안한 어절 분할 시스템은 2장에서 자세하게 기술한다. 또한 3장에서는 실험 결과 및 분석에 대해서, 4장에서는 결론에 대해 논한다.

2. 어절 분할 시스템

제안된 어절 분할 시스템은 인쇄체 텍스트 영상을 문자열 단위로 분리하고, 각각의 문자열을 띄어쓰기 단위인 어절들로 분할한다. 문자열 분리에서는 수평 투영 프로파일 분석 기법을 사용하여 일차적인 분리를 수행한다. 그리고 수평 투영 프로파일 분석에 의하여 문자열 분리가 성공적으로 이루어지지 않은 영역에 대하여 RPPC(Recursive Projection Profile Cut) 방법[15]을 적용, 이차적 문자열 분리를 수행한다. 어절 분리에서는 각각의 분리된 문자열 영상에 존재하는 gap들에 대하여 크기를 측정한 후, 어절과 어절사이에 존재하는 gap과 그렇지 않은 gap으로 군집화하고 어절 영상의 추가 분리를 가능하게 하는 특수기호를 검출한다. 최종 어절 분리 과정은 군집화 결과로 얻어진 어절 분리점과 특수기호의 정보를 이용하여 수행한다. 그림 1은 제안된 시스템의 구조를 보여준다. 입력 텍스트 영상은 300dpi의 이진 영상이며, 기울어짐 교정 등의 전처리가 수행되었다고 가정한다.

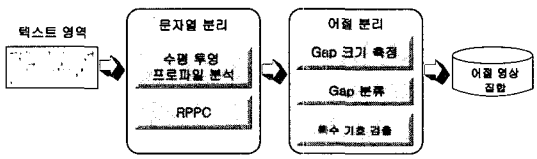


그림 1 어절 분할 시스템 구성도

2.1 문자열 분리

문자열 분리에서는 입력으로 주어지는 텍스트 영상을 문자열 단위로 분리한다. 제안 방법은 투영 프로파일(projection profile)을 이용한 방법으로써 두 단계로 구성되어있다. 첫 번째 단계에서는 텍스트 영상을 수평방향으로 투영시킨 후, 프로파일의 값이 0인 지점에서 문자열 분리를 수행한다. 문자들이 하나의 문자열에만 걸쳐 쓰여져 있는 일반적인 경우에는 이와 같은 수평 투영 프로파일 분석에 의하여 정확한 문자열 분리가 가능하다(그림 2). 그러나 그림 3(a)와 같이 영문 텍스트 영상에서 상위 문자열의 lower zone과 하위 문자열의 upper zone이 겹쳐 있는 경우, 수평 투영 프로파일이 겹치거나 서로 이웃하게 되어 문자열 분리점을 찾지 못한다. 또한 그림 3(b)처럼 하나 이상의 문자가 여러 개의 문자열에 걸쳐 쓰여져 있는 경우에는 단순한 수평방향 투영 프로파일 분석만으로 문자열 분리점을 결정하게 되면, 잘못된 문자열을 어절 분리 단계로 출력하게 된다. 이런 텍스트 영상에서는 문자열과 문자열의 사이에서 결정되는 분리점, 즉 수평방향 투영 프로파일 값이

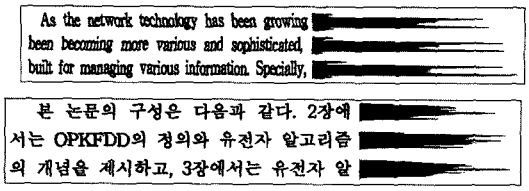
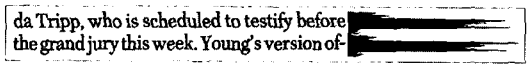
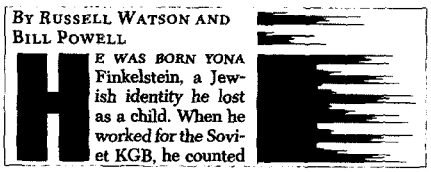


그림 2 영문과 한글 텍스트 영역에 대한 수평방향 투영 프로파일 결과



(a) RPPC-1 사례



(b) RPPC-2 사례

그림 3 수평방향 투영 프로파일만의 문자열 분리 실패 예

0인 위치를 구하지 못하여 문자열의 오분리가 발생한다. 그림 3(a)와 그림 3(b)와 같은 오류의 형태를 각각 RPPC-1과 RPPC-2로 정의한다(그림 3). 다음 단계에서 이런 오류에 대한 처리를 수행한다.

두 번째 단계에서는 첫 번째 단계의 결과로 얻어진 문자열들 중, 다른 문자열들에 비해서 높이가 상대적으로 큰 값으로 나타남으로써 두 개 이상의 문자열을 포함하고 있다고 추정되는 문자열 영상을 처리 대상으로 한다. 주어진 문자열 영상의 높이가 일차적으로 분리된 문자열의 높이에 대한 최빈값의 두 배보다 크면, 그 문자열 영상은 두 개 이상의 문자열 영역을 포함한다고 판단하여 재분리를 수행한다. 그런데, RPPC-1에 해당하는 문자열에서 서로 다른 문자열에 있는 소수의 문자들 상위부분과 하위부분이 수평 투영 시 겹쳐지는 정도가 검은 화소들의 런-길이(run-length)에 대한 최빈값의 두 배보다 작았지만, RPPC-2에서 두 문자열에 걸쳐있는 문자들의 수평 투영값은 그렇지 않음을 실험을 통하여 알 수 있었다. 그리하여 임계값 α를 검은 화소들의 런-길이에 대한 최빈값의 두 배라 정의하고 RPPC-1과 RPPC-2의 분류에 이용한다. RPPC에 의한 문자열 분할 알고리즘은 아래와 같다.

단계 1. 수평 투영 프로파일에서 임계값 α보다 작은 부분이 있으면 단계 5로 가고, 그렇지 않으면 단계 2로 간다. 이때 임계값 α는 문자열 영역에서 검은 화소들의 런-길이에 대한 중앙값에 대한 10%로 정의한다.

단계 2. 영역을 수직 방향으로 투영한다.

단계 3. 영역의 좌측으로부터 처음 수직방향 투영값(VPP: Vertical Projection Profile)이 0이 되는 부분에서 수직 방향으로 분리한다. 0의 투영값을 찾지 못하면 문자열 분리를 종료한다.

단계 4. 단계 3에서 분리된 영역 중 우측 영역에 대하여 수평방향으로 투영을 실시한다. 이때, α보다 작은 부분이 있으면 단계 5로 가고, 그렇지 않으면 단계 6으로 간다.

단계 5. α보다 작은 부분에서 문자열 분리를 수행하고 문자열 분리를 종료한다.

단계 6. 단계 3에서 분리된 우측영역에 대하여 단계 2부터 다시 수행한다.

위 알고리즘의 단계 1과 5는 그림 3(a)와 같은 RPPC-1의 처리를 위함이고, 단계 2~6은 그림 3(b)와 같은 RPPC-2의 경우를 처리하기 위함이다.

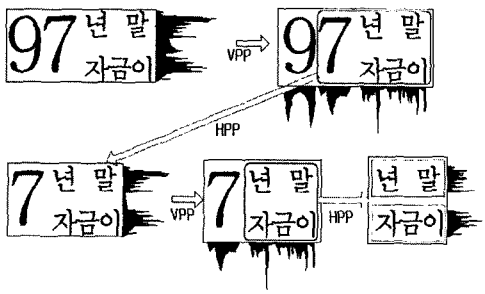


그림 4 RPPC-2의 처리 과정의 예

그림 4는 그림 3(b)와 같은 RPPC-2의 경우에 대한 처리 과정을 보여 주고 있다. 먼저 영상을 수직방향으로 투영하면 '9'와 '7' 사이에서 투영값이 0인 지점이 좌측을 기준으로 처음 관측된다. 그 다음 '9'와 '7' 사이에서 수직방향 분할을 한 후, 우측 영역에 대하여 수평방향 투영을 실행한다. 그러나 α보다 작은 수평방향의 투영값을 찾을 수 없기 때문에 우측 영역에 대하여 이전 처리를 반복하게 된다. 최종적으로 그림 4의 마지막 실행 결과처럼 두 개의 문자열로 분리가 성공적으로 수행된다.

2.2 어절 분리

이번 절에서는 문자열 분리 후 얻어진 문자열 영상을 어절 단위로 분리한다. 어절 분리를 위하여 제안 방법에서는 gap 크기 정보와 특수기호 정보를 사용한다. Gap은 그림 5에서 볼 수 있듯이 문자열을 수직 방향으로 투영한 후 얻어진 투영값이 0인 부분, 즉 흰-런(white-run)으로 정의된다. 이때 gap과 gap 사이에 존재하는 검은 화소의 집합을 어절 후보라 정의한다. 이렇게 얻어

진 gap들은 어절과 어절사이에 존재하는 IWG(Inter-Word Gap)과 IWG이 아닌 gap으로 분류한다. IWG이 아닌 gap을 편의상 ICG(Inter-Character Gap)으로 표기한다. 문자열로부터 얻어진 gap들을 IWG과 ICG으로 각각 분류하기 위하여 gap 크기 정보를 사용하는 최소 평균 거리법을 적용한다[16,17]. 최종적으로 어절 후보 사이에 존재하는 특수기호를 연결요소의 종횡비와 골격선(skeleton)의 형태적 특징 등을 이용하여 검출하고 어절 분리점, 즉 IWG를 개선한다.

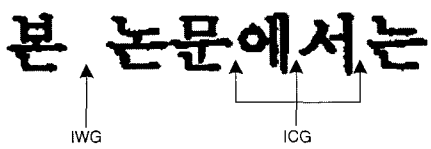


그림 5 수직방향 투영 결과 얻어진 gap

2.2.1 Gap 크기 측정

임의의 문자열 상에 존재하는 gap을 추출하기 위한 과정은 그림 6에 예시되어 있다. 먼저 주어진 문자열에 대하여 8-방향 연결요소(connected component)를 구한 후, 각 연결요소의 최소 외접 사각형(BB: bounding box)을 생성한다(그림 6(a)). 그리고 수직방향으로 겹치는 연결요소의 BB들을 병합(grouping)하여 어절 후보의 BB로 확장한다(그림 6(b)). 이때 gap은 이웃하는 두 어절 후보의 BB 사이에 존재하는 빈 공간으로 정의되며, 각 gap의 크기는 이웃하는 두 BB 사이의 수평 거리(BB 거리)로 측정한다.

BB의 크기가 임계치 이하인 어절 후보는 gap 추출 시 고려하지 않는다. 이는 최종 결과인 어절 영상에서 점(·)이나 쉼표(,), 따옴표(' 또는 ") 등과 같이 내용과 관계없는 요소들을 최대한 제외시키기 위함이다. 영문의 'i'나 'j', 한글의 'ㅎ' 등에서의 크기가 작은 연결요소는 BB의 병합과정에서 다른 연결요소와 병합되어 제외되지 않는다.

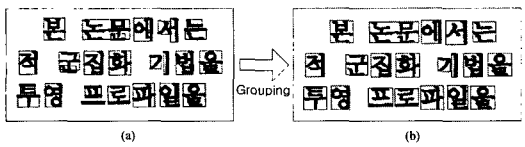


그림 6 어절 후보의 BB 생성

2.2.2 Gap 분류

Gap 분류 단계에서는 gap들을 두 클래스, IWG과 ICG로 군집화를 수행한다. 문자열로부터 얻어진 n개의 gap들을 $\{g_1, \cdots, g_n\}$이라 하고, 이들이 분류될 클래스의 집합을 {ICG, IWG}라 정의하자. 그리고 계층적 군집화를 통해 n개의 gap을 두 클래스 중 하나로 분류하면, 문자열을 구성하는 어절의 개수는 IWG으로 대응되는 gap의 개수 + 1이 된다. 그러나 문자열 영상이 하나의 어절로만 구성되어 있거나 ICG가 존재하지 않고 모든 gap들이 IWG으로 구성되어 있는 예외적인 경우에는 군집화 방법에 의한 gap 분류는 실패할 수밖에 없다. 이러한 예외적인 경우들을 위해 군집화 결과를 분석하여 수정하는 단계를 수행한다.

주어진 gap들을 IWG과 ICG으로 분류하기 위해 본 논문에서는 계층적 군집화 방법인 최소 평균 거리법(Average Linkage Method)[16]을 사용한다. 이 방법은 n개의 군집(cluster)이 두 개의 군집이 될 때까지 가까운 두 개의 군집들을 묶어 나가는 상향식(bottom-up) 군집화 방법이다. 우선 각각의 gap을 원소로 갖는 n개의 군집을 초기화한다. 그리고 군집들 간의 유사도를 계산한 후 가까운 두 군집을 병합한다. 이때 유사도는 두 군집의 중심점 사이의 맨하탄(Manhattan) 거리를 사용하여 계산된다. 이 과정을 두 개의 군집이 남을 때까지 반복한다. 군집화 결과 큰 평균을 갖는 군집에 속한 gap들이 IWG으로 분류된다. 그림 8은 그림 7과 같이 전 단계에서 측정된 gap의 크기를 이용하여 gap들을 ICG와 IWG로 군집화하는 과정을 보여준다. 초기 군집

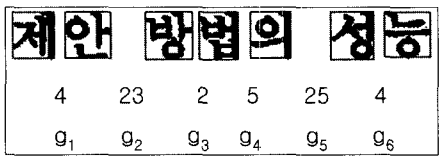


그림 7 Gap 크기 측정 결과

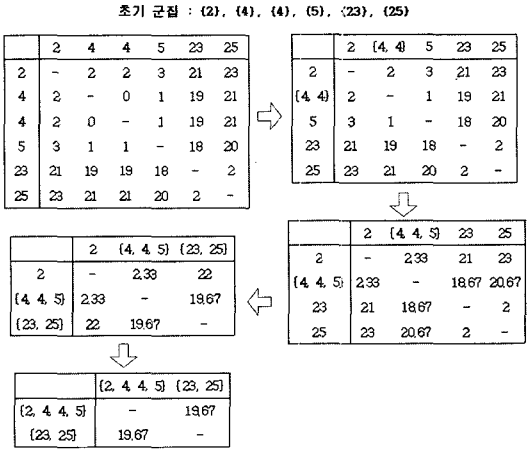


그림 8 최소 평균 거리법 적용 과정

은 6개이고, 그림 8에서의 음영이 들어간 부분처럼 각 군집간의 거리가 최소인 군집들을 합병한다.

제안 방법은 gap들을 두 클래스, IWG와 ICG 중 하나로 군집화하기 때문에 gap들이 어느 한 클래스의 패턴만으로 구성되어 있는 경우, 동일 클래스로 군집화가 될 gap들이 다른 클래스로 잘못 분류하게 된다(분류 오류-1). 또한 문자열 영상에서 이상적으로 큰 IWG이 존재하는 경우-그림 9와 같이 어절과 어절사이에 두 칸 이상의 공백(크기가 66인 gap)이 존재-와 gap 크기 측정 과정의 제외된 연결요소로 인하여 인접 IWG의 크기가 증가한 경우에 IWG으로 분류될 gap들을 ICG으로 잘못 분류될 수 있다(분류 오류-2). 이러한 문제점들을 해결하기 위하여 문자열에 적응적인 임계치(β)를 이용한 휴리스틱 규칙을 제안방법에 적용하였다. 수정된 gap 분류 알고리즘은 아래와 같다.

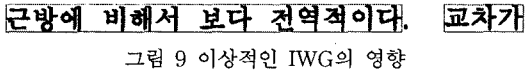


그림 9 이상적인 IWG의 영향

단계 1. gap의 개수가 1인 경우, gap의 크기가 이전 문자열에서의 최소 ICG 크기 값의 75%보다 크면 IWG로 분류하고, 그렇지 않으면 ICG로 분류하고 단계 7로 간다.

단계 2. 주어진 gap들을 ICG와 IWG 클래스로 분류하기 위하여 최소 평균 거리법의 계층적 군집화를 실행한다.

단계 3. 분류된 두 군집(ICG, IWG)간의 평균 차이가 임계값 β보다 작으면 다음 단계로 가고, 그렇지 않으면 단계 5로 간다. 이때 임계값 β는 현재 문자열의 BB들에 대한 최대 높이의 20%로 정의한다.

단계 4. 두 군집의 평균 합, 즉 전체 gap들의 평균이 β보다 크면 모든 gap들을 IWG로 분류하고, 그렇지 않으면 ICG로 분류하고 단계 7로 간다.

단계 5. ICG 클래스로 분류된 군집의 분산이 β보다 크면 ICG 클래스로 분류된 gap들을 대상으로 다시 군집화를 실행하고 다음 단계로 가고, 그렇지 않으면 단계 7로 간다.

단계 6. 단계 5에서 분류된 새로운 두 군집(ICG_NEW, IWG_NEW)간의 평균 차이가 $\beta\times2$보다 크면 새로운 IWG_NEW 군집의 gap들을 ICG 군집에서 IWG 군집으로 이동하고 단계 5부터 다시 실행한다. 만일 $\beta\times2$보다 작으면 단계 7로 간다.

단계 7. IWG 클래스로 분류된 gap 정보를 이용하여 문자열 영상을 어절 단위로 분리하고 알고리즘을 종료한다.

위 알고리즘의 단계 1, 3, 4는 분류 오류-1, 단계 5, 6은 분류 오류-2를 처리하기 위함이다. 실제 실험 영상에서 분류 오류-1과 분류 오류-2의 경우는 각각 15회와 6회가 나타났으며, 이들은 위 알고리즘에 의해 수정되어 정상적인 gap 분류를 수행하였다.

2.2.3 특수기호 검출

일반적으로 문자열을 띄어쓰기 단위인 어절 집합으로 분리하는데 IWG만을 이용하지만, 그림 10과 같이 어절 속에 위치하는 특수기호(~, -, (,), {, }, [,] 등)들을 통해서도 서로 독립적인 어절들로 분리할 수 있다. 그러나 이러한 방법은 어절 사이에 위치한 연결요소가 특수기호라고 인식한 후에 가능하다. 본 논문에서는 검출할 특수기호들을 어절 사이에 빈번히 나타나는 두 가지의 유형으로 구분하고, 이들에 대한 인식과정 없이 휴릭스틱 특징(BB의 종횡비, 골격선의 기하학적 구조 등)을 이용하여 위치를 파악하고 어절 분리점을 수정한다.

그림 10 특수기호로 인해 재분리가 가능한 문자열의 예

우선 첫 번째 유형의 특수기호는 그림 10의 ①처럼 수평방향으로 종횡비(aspect ratio)가 큰 연결요소로서 '~'와 '-' 등이 있는데, 이를 검출하기 위하여 병합된 연결요소의 높이($Height_{GroupedCC}$)와 폭($Width_{GroupedCC}$), 무게중심의 수직성분($VerticalCenter_{GroupedCC}$), 문자열의 높이($Height_{Line}$) 등을 고려한다. 다음과 같은 조건을 만족시키는 연결요소를 첫 번째 유형의 특수기호로 분류한다.

<조건 1> $Height_{GroupedCC} < Height_{Line}\times50\%$

<조건 2> $Width_{GroupedCC} > Height_{GroupedCC}\times2$

<조건 3> $Height_{Line}\times25\% < VerticalCenter_{GroupedCC} < Height_{Line}\times75\%$

여기서 <조건 1>과 <조건 2>는 연결요소의 수평성분에 대한 종횡비를 고려하고, <조건 3>은 밑줄(underline)과 구별하기 위함이다.

두 번째 유형의 특수기호는 그림 10의 ②와 ③과 같이 수직방향의 종횡비가 큰 연결요소로서 '{', '}', '[', ']', '(', ')' 등이 이에 해당된다. 이것들의 검출 과정은 연결요소의 종횡비를 고려한 후보군 추출과 추출된 후보들의 휴리스틱 특징을 이용하여 특수기호를 검출하는 두 단계로 구성된다. 첫 번째 단계에서는 수직방향에 대한 종횡비가 큰 특수기호 후보들이 검출되는데, 여기서 검출된 후보들은 연결요소의 높이와 폭을 고려한 아래의

두 조건을 만족시킨다.

<조건 1> $Width_{GroupedCC} < Height_{Line} \times 50\%$

<조건 2> $Height_{GroupedCC} > Width_{GroupedCC} \times 2$

두 번째 단계에서는 처음 단계의 결과로 얻어진 후보군에서 검출하고자 하는 특수기호를 분류하기 위하여 몇 가지 휴리스틱 특징을 이용한다. 이전 단계에서 검출된 후보군에는 특수기호뿐만 아니라 'l', 'i', 'I', 'j', 'J', 'h', 'H', 't', 'ㅣ', 'ㅏ', 'ㅑ', 'ㅓ', 'ㅕ', 'ㅐ' 등에 해당하는 연결요소들도 포함된다. 이런 후보군에서 특수기호를 선별하기 위하여 형태 분석을 수행한다. 그러나 형태 분석을 연결요소의 모든 검은 화소들을 대상으로 수행하면 복잡하고 데이터량이 증가하기 때문에, 연결요소의 세선화된 골격선에 대하여 수행한다. 이때 세선화 방법으로 골격선의 국부적인 잡영 제거 및 보다 세밀한 형태적 특성을 유지할 수 있는 알고리즘을 적용하였다[18]. 최종적으로 형태 분석으로 얻어진 특수기호와 비특수기호의 휴리스틱한 형태적 특징을 이용하여 특수기호를 선별한다. 여기서 휴리스틱한 형태적 특징에는 골격선의 끝점(end point) 개수와 검은 화소들간 연결성(수평방향, 역사선방향, 사선방향)의 구성과 횟수가 수직으로 대칭이 되는지의 여부 등이 있다. 두 번째 유형의 특수기호를 검출하는 두 번째 단계의 알고리즘은 다음과 같다.

단계 1. 후보군의 연결요소를 세선화하여 m-path 표현의 골격선을 구한다. 그리고 골격선이 아래의 두 조건을 만족하지 않으면 단계 10으로 간다.

<조건 1> 끝점(end-point)의 개수가 2개이다.

<조건 2> 골격선의 검은 화소 개수가 골격선 높이의 150% 값보다 작다.

단계 2. 위의 단계에서의 골격선에 대한 BB(Boundary Box)를 분석하고, 수직방향에 대하여 골격선의 BB를 상위와 하위 영역으로 이등분한다. 아래의 조건을 만족하지 않으면 단계 10으로 간다.

<조건 3> 상위 영역에서 검은 화소들이 존재하는 라인의 개수가 상위 영역의 라인 개수의 75% 이상이다.

단계 3. 전체 영역과 각각의 영역별로 검은 화소들의 수평 및 대각 연결 횟수를 분석한다. 분석된 결과는 아래의 표와 같이 정의한다.

단계 4. 연결요소의 기울임, 즉 이탤릭 여부를 검사한다. 아래의 조건을 만족하면 단계 7로, 만족하지 않으면 다음 단계를 수행한다.

<조건 4> $\frac{cntDiagonal2}{cntDiagonal1 + cntDiagonal2} > 75\%$ AND $cntDiagonal2 > \frac{Width_{SkeletonBB}}{2}$

($Width_{SkeletonBB}$: 골격선에 대한 BB의 폭)

단계 5. 아래의 조건들을 만족하면 '[' 와 같은 특수기호로 구분하고 단계 9로, 만족하지 않으면 다음 단계를 수행한다. 여기서 γ는 상위 영역과 하위 영역의 대칭 여부에 대한 임계치로써, $\gamma = \frac{2}{3}$ 로 정의한다.

<조건 5> $cntHorrizonTop + cntHorrizonBottom > Width_{GroupedCC}$

<조건 6> $cntDiagonal1 + cntDiagonal2 < Width_{GroupedCC}$

<조건 7> $\frac{\max(cntHorrizonTop, cntHorrizonBottom)}{cntHorrizonTop + cntHorrizonBottom} < \gamma$

단계 6. 아래의 조건을 만족하면 '(', '{' 와 같은 특수기호로 구분하고 단계 9로, 만족하지 않으면 단계 10으로 간다.

<조건 8> $cntDiagonal1 + cntDiagonal2 > Width_{GroupedCC} / \gamma$

<조건 9> $\frac{\max(cntDiagonal1, cntDiagonal2)}{cntDiagonal1 + cntDiagonal2} < \gamma$

단계 7. <조건 7>과 아래의 조건을 만족하면 기울임이 있는 '[', 즉 '[' 과 같은 특수기호로 구분하고 단계 9로, 만족하지 않으면 단계 10으로 간다.

<조건 10> $cntHorrizonTop + cntHorrizonBottom > Width_{GroupedCC} \times \gamma$

단계 8. <조건 9>와 아래의 두 조건을 만족하면 기울임이 있는 '('나 '{', 즉 '('나 '{'와 같은 특수기호로 구분하고 단계 9로, 만족하지 않으면 단계 10으로 간다.

<조건 11> $\max(cntDiagonal2Top, cntDiagonal2Bottom) > \min(cntDiagonal2Top, cntDiagonal2Bottom) \times 2$

<조건 12> $\max(cntDiagonal2Top, cntDiagonal2Bottom) = \frac{Width_{GroupedCC}}{2}$

단계 9. 검출된 특수기호의 BB 정보를 이용하여 IWG 정보를 수정한다.

단계 10. 알고리즘을 종료한다.

위의 알고리즘에서 <조건 1-3>은 특수기호와 비특수기호간의 형태 특징들 중, 뚜렷한 차이를 가져오는 특징들을 이용하여 비특수기호의 일부를 후보군에서 제외시킨다. <조건 4>는 연결요소의 기울임 여부를 판단하여 기울임이 있는 특수기호와 기울임이 없는 특수기호에

방향 / 영역	수평방향	역사선방향(↘)	사선방향(↙)
전체영역	*cntHorrizon*	*cntDiagonal1*	*cntDiagonal2*
상위영역	*cntHorrizonTop*	*cntDiagonal1Top*	*cntDiagonal2Top*
하위영역	*cntHorrizonBottom*	*cntDiagonal1Bottom*	*cntDiagonal2Bottom*

대하여 서로 다른 조건으로 처리하기 위함이고, <조건 5-9>는 기울임이 없는 특수기호, <조건 10-12>는 기울임이 있는 특수기호를 선별하는데 이용된다. 여기서 γ는 골격선의 어떤 특징에 대한 상위 영역의 값과 하위 영역의 값이 대칭으로 구성되었는가를 알아보기 위한 임계치이다. <조건 7과 9>에서 γ에 비교되는 값이 γ보다 더 크다는 것은 어느 한 영역의 값이 다른 영역의 것보다 월등히 커서 대칭 구조로 볼 수 없음을 의미한다. 본 논문에서 제안하는 알고리즘들에서 사용되는 임계치 또는 파라미터(parameter)는 본 논문의 실험 영상뿐만 아니라, 잡지와 다른 논문지의 텍스트 영상에 대한 실험을 통하여 튜닝(tuning)되었다.

3. 실험 결과 및 분석

본 논문에서 제안한 문자열 분리 및 어절 분리 방법을 평가하기 위하여 총 84개의 텍스트 영역을 사용하였다. 텍스트 영상은 '정보처리논문지'에 게재된 논문의 일부로서, 각각 1편의 한글과 영문 논문에서 텍스트 영역에 해당되는 부분만을 수작업으로 추출하여 만들어졌다(표 1). 또한 텍스트 영상은 SHARP JX-330P 스캐너에 의해 300 dpi 해상도를 갖는 이진 영상으로 스캐닝되었다. 본 논문에서 제안한 방법의 성능을 입증하기 위하여 국내의 유명한 OCR 상용제품인 '아르미 6.0 프로'의 성능과 비교하였다. 이때 아르미의 인식 옵션으로 숫자와 영어, 한글, 한자를 인식문자로 설정하였고, 이탤릭 구별이 가능하게 하였으나 후처리에 해당하는 단어 검사는 옵션에서 제외하였다.

표 1 실험에 사용된 텍스트 영상

	한글	영문	계
영상개수	36	48	84

3.1 문자열 분리 성능 분석

문자열 분리에서 제안 방법과 아르미는 총 904개의 문자열을 모두 성공적으로 분리하였다(표 2). 이는 실험 대상이 되는 텍스트 영상의 내용이 논문의 양식을 따르면서 일정한 줄간격으로 문자열들을 구성하고 있기 때문이다.

표 2 문자열 분리 실험 결과

	제안 방법		아르미	
	한글	영문	한글	영문
총 문자열 수	524	380	524	380
분리 성공 횟수	524	380	524	380
성공률(%)	100	100	100	100
	100(904/904)		100(904/904)	

3.2 어절 분리 성능 분석

어절 분리의 성능은 gap 정보 획득 후 군집화를 통해 분류된 IWG로 어절 분리를 실행하는 1차 과정과 특수기호 검출로 수정된 IWG로 어절 분리를 실행하는 2차 과정으로 나누어 분석하였다.

3.2.1 특수기호 검출을 고려하지 않은 어절 분리

1차 과정의 성능은 어절 후보사이에 존재하는 gap들을 ICG와 IWG로 얼마나 정확하게 분류하는 가를 가지고 측정하였다. 그래서 오류는 IWG를 ICG로 분류하는 것과 반대로 ICG을 IWG로 분류하는 두 가지 유형이 있다. 전자의 오류 유형은 두 어절을 하나의 어절로 결정하여 분리를 실행하지 않는 결과를 가져온다. 그림 11처럼 영문으로 구성된 문자열에서는 IWG와 ICG의 차이가 작아서 일부 IWG가 ICG로 분류되어 어절 분리가 실패하는 경우가 발생할 수 있는데, 이러한 오류가 아르미에서는 11번, 제안 방법에서는 단 1번만 보였다. 후자의 오류 유형은 하나의 어절을 두 개의 어절로 과분리하는 결과를 가져온다. 또한 아르미에서는 그림 12와 같이 하나의 어절인 '표시하였다'가 '표시'와 '하였다'의 두 어절로 과분리하는 경우가 10번 발생하였지만 제안 방법은 1번 경우에만 과분리가 발생하였다.

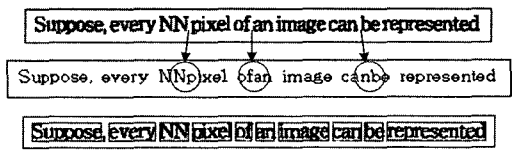


그림 11 제안방법과 아르미와의 어절 분리 실험 결과 1
(상: 원영상, 중: 아르미 결과, 하: 제안방법 결과)

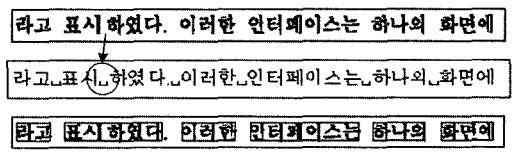


그림 12 제안방법과 아르미와의 어절 분리 실험 결과 2
(상: 원영상, 중: 아르미 결과, 하: 제안방법 결과)

어절 분리 오류는 이론상으로 앞서 언급한 두 가지 유형이 있지만, 실험에서 아르미는 몇 개의 문자열 영상을 텍스트로 변환하지 못하고 문자열 영상을 대신 보여주었다. 그림 13에서 우하단의 아르미 결과를 참조하면, 크고 진한 글씨로 구성된 문자열과 그렇지 않은 문자열들이 존재함을 알 수 있다. 여기서 크기가 작은 부분은 정상적으로 변환된 텍스트이고, 화살표가 지시하는 크고 진한 부분은 문자열 영상으로 남아있다. 왜 이러한 오류가 생기는지 정상적으로 텍스트 변환되는 다른 문자열

영상과 비교해보았지만 오류의 원인을 발견할 수 없었다. 아르미에서는 이와 같은 오류가 한글·영문 영상에서 각각 4개씩, 8개의 문자열 영상에서 나타났으며, 이들 문자열 영상에 존재하는 어절 영상(140여개)들의 분리는 모두 실패하였다. 모든 오류를 고려한 어절 분리 실험은 표 3의 결과에서처럼 제안 방법이 2.3%의 차이로 아르미보다 우수함을 보였다.

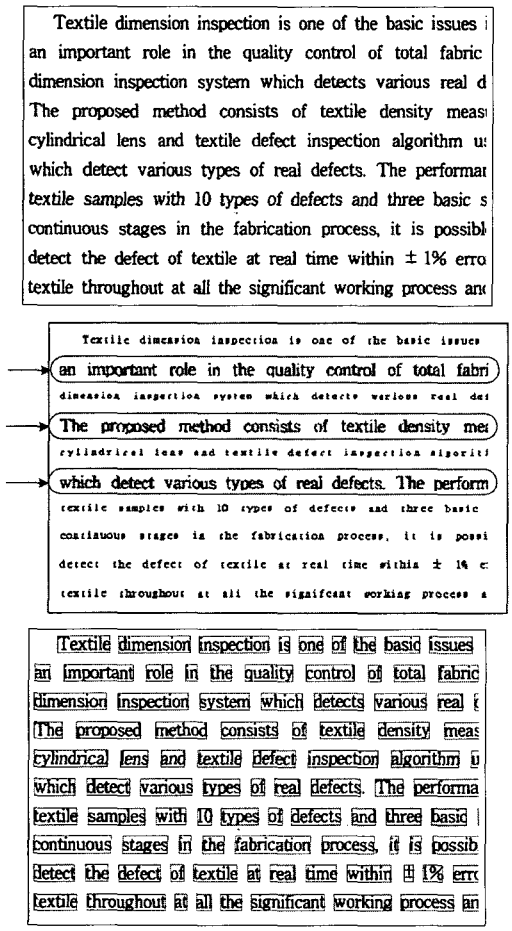


그림 13 제안방법과 아르미와의 어절 분리 실험 결과 3 (상: 원영상, 좌하: 아르미 결과, 우하: 제안방법 결과)

표 3 Gap 분류에 의한 어절 분리 실험 결과(특수기호 고려하지 않음)

	제안 방법		아르미	
	한글	영문	한글	영문
총 어절 수	3910	3153	3910	3153
분리 성공 횟수	3909	3152	3826	3071
성공률(%)	99.97	99.97	97.85	97.40
	99.97(7062/7064)		97.65(6897/7064)	

3.2.2 특수기호 검출을 고려한 어절 분리

최종적인 어절 분리를 위해서는 특수기호 검출이 선행되어야 한다. 그림 14와 같이 검출된 특수기호의 정보를 이용하여 군집화된 IWG의 결과만으로 분리할 수 없었던 어절들을 분리할 수 있었다. 첫 번째 유형의 특수기호 검출에서는 아르미가 제안 방법보다 우수하였는데, 이는 수직방향 투영으로 특수기호 후보군을 결정하는 제안 방법이 특수기호 '-'가 인접한 연결요소들과의 접촉때문에 검출되는데 실패하였기 때문이다. 그리고 두 번째 유형의 특수기호 검출에서 아르미는 그림 15와 그림 16처럼 '['이나 기울임이 있는 ')' 등의 특수기호를 알파벳으로 변환하여 실패하였지만, 제안 방법에서는 그림 16처럼 특수기호가 인접한 연결요소와 수직으로 겹치게 되어 후보군에서 제외되는 경우에서만 특수기호 검출이 실패하였다. 특수기호 검출의 성능은 제안 방법이 4%의 차이로 아르미보다 우수함을 보였다(표 4).

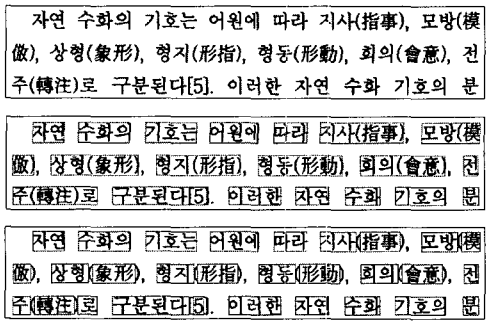


그림 14 제안방법과 아르미와의 특수기호 검출 실험 결과 1 (상: 원영상, 중·하: 특수기호를 고려하지 않은 경우와 고려한 경우)

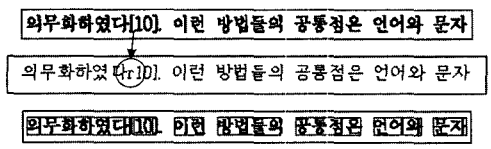


그림 15 제안방법과 아르미와의 특수기호 검출 실험 결과 2 (상: 원영상, 중: 아르미 결과, 하: 제안방법 결과)

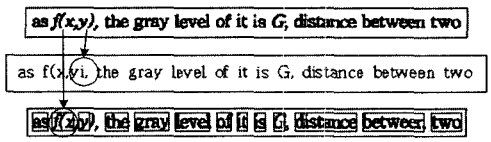


그림 16 제안방법과 아르미와의 특수기호 검출 실험 결과 3 (상: 원영상, 중: 아르미 결과, 하: 제안방법 결과)

표 4 특수기호 검출 실험 결과

	제안 방법				아르미			
	한글		영문		한글		영문	
	-, ~	(, [, {	-, ~	(, [, {	-, ~	(, [, {	-, ~	(, [, {
총 개수	37	144	48	134	37	144	48	134
검출 성공 횟수	37	144	46	133	33	140	48	124
성공률(%)	100.00	100.00	95.83	99.25	89.19	97.22	100.00	92.54
	100		97.54		93.20		96.27	
	98.77(360/363)				94.73(345/363)			

표 5 최종 어절 분리 실험 결과

		제안 방법		아르미	
		한글	영문	한글	영문
총 어절 수		3,997	3,199	3,997	3,199
Gap 분류만의 어절 분리	개수	3,909	3,152	3,826	3,071
	성공률(%)	97.80	98.53	95.72	96.00
특수기호 검출 수행 후 어절 분리	개수	3,996	3,195	3,910	3,114
	성공률(%)	99.97	99.87	97.82	97.34
총 성공률		99.92(7,189/7,196)		97.58(7,024/7,196)	

표 6 아르미와의 각 단계별 성능 비교 (성능 %)

	제안 방법	아르미
문자열 분리	100	100
Gap 분리	99.97	97.65
특수기호 검출	98.77	94.73
최종 어절 분리	99.92	97.58

표 5는 텍스트 영상에 대한 문자열 분리와 gap 군집화, 특수기호 검출 등 모든 처리를 수행한 어절 분리 실험의 결과로써, 제안 방법이 아르미보다 약 2.3%의 차이로 우수함을 보인다. 표 6은 제안 방법과 아르미를 텍스트 영상에 대한 어절 분리 시스템의 각 단계별 성능으로 비교하고 있다.

4. 결 론

본 논문에서는 인쇄체 텍스트 영상으로부터 어절 영상을 추출하는 시스템을 제안하였다. 텍스트 영상을 문자열 단위로 분할하기 위해 수평·수직 방향의 투영 프로파일 분석을 재귀적으로 적용하는 방법을 채택하였으며, 각각의 문자열을 띄어쓰기 단위인 어절의 집합으로 분리하기 위해 문자열 상에 존재하는 gap과 특수기호를 추출하는 방법을 제안하였다. 추출된 gap 중 어절과 어절사이에 존재하는 gap을 찾기 위해 계층적 군집화 방법을 사용하였으며, 특수기호 검출을 위해 세선화을 실행하고 골격선의 형태적 특징을 이용하였다. 84개의 한글 및 영문 텍스트 영상에 대하여 실험한 결과, 100%의 문자열 분리와 99.92의 어절 분리 성공률을 관측하였다. 그리고 제안 방법의 우수한 성능을 입증하기 위하여 아르미의 실험 결과와 비교한 결과, 제안 방법이 약 2.4%의 차이로 우수하였다.

본 논문에서 제안한 어절 분할 시스템은 문서 영상을 텍스트 문서로 변환하는 문자인식 시스템(OCR)의 성능 개선을 위해 활용될 수 있으며, 아울러 전자도서관 등에서와 같이 문서 영상의 자동 색인(automatic indexing) 및 주제어 기반 검색을 위한 시스템의 핵심 요소 기술로 활용될 수 있다.

참 고 문 헌

[1] AIM'96 Conference Handbooks, *Association for imaging and information methodologies*, 1996.

[2] J. L. George, "Digitization: a literature review and summary of technical processes," Information Services Group, Oct. 1994.

[3] 장명욱, 천대녕, 양현승, "연결화소를 이용한 문서 영상의 분할 및 인식", 한국정보과학회 논문지, Vol 20, No.12, pp.1741-1750, 1993.

[4] 김두식, 이성환, "한·영 혼용 문서의 디지털 라이브러리 구출을 위한 효과적인 문서 기울기 교정 및 문자 분할 방법", 한국정보과학회 봄 학술발표논문집, Vol. 23, No.1, pp.293-296, 1996.

[5] 배진학, 박세현, 김항준, "영·숫자 한글 문서에서 문자 분리 및 인식", 정보과학회 논문지, 제23권 제9호, pp.941-949, 1996.

[6] 김두식, 이성환, "한글과 영·숫자가 혼용된 문서를 위한 효과적인 문자 분할 방법", 제 8회 영상 처리 및 이해에 관한 워크샵 발표논문집, pp.19-26, 1996.

[7] 임장준, "인쇄된 한영 혼용 문서 인식을 위한 문자 분할 방법과 문자의 한글과 영어의 구별", 포항공과대학교 대학원 석사학위논문, 1998.

[8] 최정호, 김태균, 남궁재찬, 신문 자동인식 시스템의 개

발, 연구보고서, 1991.
[9] 정규식, 권희웅, "내용기반의 인쇄체 영문 문서 영상 검색을 위한 특징기반 단어 검색", 정보과학회 논문지 (B), 제26권, 제10호, pp.1204-1218, 1999.
[10] 조현목, 이경무, 최영우, "Projection Profile을 이용한 새로운 자동 문서영상의 영역분리 및 분류 알고리즘", 제9회 영상처리 및 이해에 관한 워크샵, pp.136-140, 1997.
[11] S. N. Srihari, S. Lam, V. Govindaraju, R. Srihari and J. J. Hull, "Document understanding: research directions," CEDAR-TR-92-1, May 1992.
[12] 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.
[13] F. R. Jenkins, T. A. Nartker and S. V. Rice, "Result of the fifth annual test of OCR technology by UNLV's Information Science Research Institute," *Inform Magazine*, pp.20-25, Sep. 1996.
[14] K. Marukawa, T. Hu, H. Fujisawa and Y. Shima, "Document retrieval tolerating character recognition errors-estimation and application," *Pattern Recognition*, Vol.30, No.8, pp.1361-1371, 1997.
[15] 류대석, 강선미, 이성환, "매개변수에 무관한 새로운 문서 구조 분석 방법", 한국정보과학회 가을 학술발표 논문집, Vol.26, No.2, pp.482-484, 1999.
[16] E. Gose, R. Johnsonbaugh and S. Jost, *Pattern recognition and image analysis*, Prentice Hall, 1996.
[17] Soo H. Kim, S. Jeong, G.S. Lee, and C.Y. Suen, "Gap Metrics for Handwritten Korean Word Segmentation," IEE Electronics Letters, Vol. 37, No. 14, pp. 892-893, July 2001.
[18] Lei Huang, Genxun Wan, Chanping Liu, "An Improved Parallel Thinning Algorithm," Proc. 7th International Conference on Document Analysis and Recognition, pp.780-783, 2003.

정 창 부
1999년 호남대학교 컴퓨터공학과 졸업(학사). 2001년 전남대학교 전산통계학과 졸업(이학석사). 2003년~현재 전남대학교 전산학과 박사과정. 관심분야는 문서영상전처리, 영상처리

김 수 형
1986년 서울대학교 컴퓨터공학과 졸업(학사). 1988년 한국과학기술원 전산학과 졸업(공학석사). 1993년 한국과학기술원 전산학과 졸업(공학박사). 1993년~1996년 삼성전자 멀티미디어연구소 선임연구원. 1997년~현재 전남대학교 컴퓨터정보학부 부교수. 관심분야는 패턴인식, 문서영상전처리, 웨이블릿변환, 유비쿼터스컴퓨팅