論文 97-22-3-2

연결요소를 이용한 한·영 혼용문서의 구조분석 및 낱자분리

學生會員 김 민 기* 終身會員 권 영 빈* 正會員 한 상 용*

Bilingual Document Analysis and Character Segmentation using Connected Components

Min-Ki Kim*, Young-Bin Kwon*, Sang-Yong Han* *Regular Members*

※이 논문은 한국과학재단(93-0100-02-01-3)과 96년도 중앙대학교의 연구비 지원에 의한 결과임.

요 약

본 논문은 연결요소 분석에 의한 상향식 방법을 이용하여 문서의 구조를 분석하고, 연결요소의 배치형태를 이용하여 한·영 혼용문서에서의 낱자분리 방법을 제안하였다. 기존의 연구가 문서를 문자블럭과 그래픽으로 나누는데 비하여 본 논문에서는 문서를 크게 네가지의 요소 즉, 문자, 도표, 그래픽, 구분선으로 분류하였다. 문자영역은 문자블럭, 문자열, 단어, 낱자의 단계를 거치며 구조를 점차적으로 세분하였고, 문자를 제외한 나머지는 그림영역으로 나누어 구분선, 도표, 그래픽으로 분류하였다. 그리고 일반적인 한·영 혼용의 문서에 대하여 한글의 낱자분리를 정확히 수행하기 위하여 각 단어를 한글단어와 영어단어로 그 형태를 결정한 후, 한글단어에 대하여 한글의 특성에 맞는 낱자합병과 접촉문자분리를 수행하는 방법을 제안하였다. 제안된 방법에 따라 여러종류에 문서에 대하여 구조분석과 낱자분리를 실험한 결과 제안된 방법이 매우 효과적임을 알 수 있었다.

ABSTRACT

In this paper, we described a bottom-up document structure analysis method in bilingual Korean-English document. We proposed a character segmentation method based on the layout information of connected component of each character. In many researches, a document has been analyzed into text blocks and graphics. We analyzed a document into four parts: text, table, graphic, and separator. A text is recursively subdivided into text blocks, text

*중앙대학교 컴퓨터공학과
論文番號:96145-0511
接受日字:1996年 5月 11日

lines, words, and characters. To extract the character in bilingual text, we proposed a new method of word separation of Korean or English. Futhermore, we used a character merging and segmentation method in accordance with the properties of Hangul on the Korean word blocks. Experimental results on the various documents show that the proposed method is very effectively operated on the document structure analysis and the character segmentation.

Ⅰ. 서 론

문자인식에 대한 연구는 지난 40여년간 꾸준히 수행되어 상당한 수준에 도달하였으며[1], 특히 인쇄체 문자에 대해서는 만족할 만한 인식결과를 보이고 있다[2, 3]. 이러한 연구결과에도 불구하고 상용화가 잘 이루어지지 않는 이유는 오프라인 문자인식 기술이 그 자체만으로는 다양한 문서에 적용하기가 어렵기 때문이다.

기존의 오프라인 문자인식에 대한 연구들은 낱자로 분리된 문자를 인식대상으로 하였거나[2, 3] 매우 제한된 형태의 문서[4]를 인식 대상으로 하였기 때문에 실생활속의 여러 가지 인쇄물을 자동으로 인식할 수 없었다. 또한 문자인식 연구자들은 제각기 자국에서 사용되는 문자 만을 주대상으로 연구를 수행하였기 때문에 하나의 문서에 여러나라의 문자들이 나타날 경우에는 인식에 많은 문제를 나타내고 있다.

이러한 문제를 해결하기 위해서는 임의의 형태를 갖는 문서로부터 문자, 도표, 그래픽 등을 구분할 수 있는 문서구조 분석과 문자영역에 대한 낱자분리가 필수적이다[5]. 그리고 문서구조 분석 과정에서 추출한 정보를 이용하여 문자집합을 대분류할 수 있다면 각 문자집합에 대한 기존의 인식기를 수정하지 않고서도 다양한 문자집합이 혼용된 문서를 효과적으로 인식할 수 있다[6]. 물론 문자집합에 대한 대분류 없이 인식하는 방법도 가능하나, 이 경우에는 기존 인식기의 수정이 불가피하여 확장성이 크게 떨어지는 단점이 있다.

문서구조 분석 방법에는 문서영상을 구성하는 기본요소를 찾아내어 이를 기반으로 문서 전체의 구조를 분석해 가는 상향식 방법[7]과 문서 전체를 작은 영역으로 분할해 가는 하향식 방법[8]이 있다. 또한 이 두 가지를 병행하여 분석하는 방법이 사용되기도 한다. 상향식 방법의 대표적인 예로서는 연결요소분석(connected component analysis) 방법이 있으며, 하향식 방법의 대표적인 예로서는 런길이 평활화(run length smoothing) 방법이 있다. 런길이 평활화 방법은 구현이 간단하고 처리속도가 빠르지만 복잡한 문서에서 다양한 구조를 세밀하게 분석하는 능력이 떨어진다. 연결요소분석 방법은 구현이 복잡하고 처리속도가 상대적으로 느리지만 세밀하게 구조를 분석할 수 있는 장점을 가지고 있다.

Fletcher[9]는 하프변환(Hough transform)을 응용하여 입력영상에서 추출한 연결요소를 논리적인 문자열로 그룹핑하였으며, 이를 이용하여 그래픽과 문자를 구분하는 방법을 제안하였다. Wahl[10]은 런길이 평활화 알고리즘으로 문자, 선, 그래픽, 하프톤(half tone)영상을 분리하는 방법을 제안하였다. 이 방법은 처리속도가 빠르며, 비교적 간단한 조건과 방법으로 문자블럭과 그래픽 블럭을 쉽게 구분할 수 있다. 하지만 기울어짐에 크게 제약을 받으며, 런길이에 대한 임계값이 고정되어 있어 다양한 문서에 대한 적응성이 떨어진다. Hirayama[11]는 런길이 평활화 방법과 연결요소분석 방법을 혼용한 방법을 제안하였다. 먼저 런길이 평활화 방법으로 문자열을 연결하고, 그 특성을 이용하여 문자그룹을 구성한 후에 문자그룹의 가장자리 정보로부터 단을 구성하여 모든 그룹을 다시 적절한 블럭으로 나누고 있다. 나뉘어진 각 블럭은 투영윤곽(projection profile) 방법에 의하여 문자와 그림으로 구분하고 있다.

문서구조분석 방법은 문서영상의 전체적인 구조에 분석의 촛점을 맞추는데 비하여, 낱자분리는 문서영상 중에서 문자영역만을 대상으로 하고있다. 그러므로 문서내부의 문자영역을 분석하여 개개의 낱자영상을 추출하는 것을 낱자분리라 한다. 낱자분리 방법은 크게 내부분할과 외부분할의 두 가지로 나눌 수 있다. 내부분할은 낱자분리 과정에 문자인식이 병행되는 방법으로써, 접촉된 문자는 분리시키고 형태에

따라 분리된 한글은 결합시킴으로써 낱자를 구성하도록 하기 위하여 문자인식을 수행하면서 적절한 분할장소 또는 합병될 문자를 찾아가는 방법이다[4, 8, 12, 13]. 이에 반해 외부분할은 인식과는 무관하게 문자의 형태만으로 낱자분리를 수행하는 방법이다[14, 15, 16]. 일반적으로 외부분할은 문자의 형태 정보만으로 낱자를 구분하기 때문에 형태적 모호성에 따른 낱자분할의 한계성이 있으나 분리과정이 빠르고 인식기에 독립적인 장점이 있다.

본 논문은 연결요소 분석에 의한 상향식 방법을 이용하여 문서의 구조를 분석하고, 연결요소의 배치형태를 이용하여 한·영 혼용문서에서 한글과 영문자를 구분한 후 낱자 분리를 수행하는 방법을 제안한다. 논문의 구성을 살펴보면 다음과 같다. 2장에서는 제안하는 문서분석 방법을 설명하고, 3장에서는 한·영 혼용문서가 내포하고 있는 낱자분리의 문제점을 살펴본 후 제안하는 낱자분리 방법을 상세히 기술한다. 4장에서는 구현된 시스템의 실험결과를 보이고, 마지막으로 5장에서 결론 및 앞으로의 과제에 대하여 설명하기로 한다.

Ⅱ. 문서구조분석의 단계

기존의 문서구조분석 방법의 대부분은 문서를 단순히 문자영역과 그림영역으로만 구분하고 있으므로 공문이나 일반 인쇄물 등에서 흔히 볼 수 있는 도표는 단순히 그림영역으로 분류되거나 또는 문자만을 분리하는 수준이었다[9, 17]. 장명욱[4]은 연결요소의 크기와 수정된 런길이 평활화 알고리즘에 의해서 문서를 문자부분과 영상부분으로 분할하고, 문자부분의 연결화소를 왼쪽에서 오른쪽으로 탐색하여 자소인식 결과에 따라 연결화소를 분할 및 결합시키면서 단어내의 열결화소들을 글자별로 분할하였다. 최봉희[15]는 입력된 문서영상에서 문자부분을 사용자가 지정하도록 하여 문자부분만을 대상으로 가로 및 세로방향에 대한 흑화소의 누적분포를 이용하여 문자열을 추출하고, 문자열을 다시 흑화소에 의한 누적분포로 문자분할을 수행하였다. 이동준[12], 이승형[16] 등은 표, 그림 등을 고려하지 않고 문자만을 대상으로 문자분할을 수행하였다.

본 논문에서는 기존의 문서구조분석 방법보다 세밀하게 문서를 분석하므로써 도표의 분리까지도 수행할 수 있는 방법을 제안하기로 한다. 구분 대상으로 정한 요소들은 그림 1의 트리(tree)와 같은 형태를 갖도록 하였다.

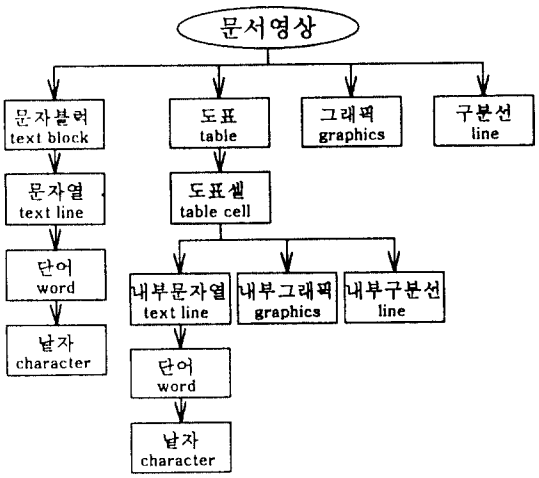


그림 1. 문서영상요소의 구분 트리
Fig. 1 Tree of document elements

기존의 연구가 문서를 문자블럭과 그래픽으로 나누는데 비하여 제안하는 시스템에서는 그림 1에서 볼 수 있듯이 문서를 크게 네가지의 요소 즉, 문자, 도표, 그래픽, 구분선으로 분류한다. 문자영역은 문자블럭, 문자열, 단어, 낱자로 그 구조가 점차적으로 세분화되어 저장되고, 문자를 제외한 나머지는 그림영역으로 구분선, 도표, 그래픽으로 세분된다. 이 때 구분선은 제외되도록 하였다. 구분선은 단구분선과 줄구분선을 말하며, 이는 따로 분리되어 문서내의 단을 구성하는데 사용된다. 도표는 하나의 도표내부에 도표를 구성하는 셀(cell)을 하나 이상 가지고 있어야 하며, 각 도표셀은 그 내부에 문자, 그래픽, 구분선 등의 모든 요소를 포함할 수 있고 또한, 다른 도표를 포함할 수도 있으나 이에 대한 포함관계는 분석대상에서 제외시켰다. 이와같은 방법을 이용하게 되면 국내의 신문구조분석에도 유용하게 접근이 가능한 장점이 있다.

2.1 연결요소 그룹핑

2.1.1 연결요소 추출

윤곽선 추출방법[18]을 사용하여 입력영상으로부터 연결요소를 추출한다. 이 중에서 외부윤곽선(external contour)이 주로 분석의 대상이 되며, 내부윤곽선(internal contour)은 외부윤곽선의 그룹핑과 도표의 구성에 사용된다. 연결요소의 크기는 최소외접사각형의 폭과 너비의 합으로 나타낸다. 본 연구에서는 다양한 문서에 대한 적응성을 위하여 분석의 기본이 되는 연결요소의 평균크기(AVG_CM)를 아래와 같이 구하였다.

$$AVG\_CM = \sum_{i=1}^{N} (CM(i).width + CM(i).depth)/N$$

CM(i).width, CM(i).depth는 i번째 연결요소에 대한 최소외접사각형의 너비와 높이를 나타내고, N은 연결요소의 총 갯수이다. 연결요소는 크기순으로 정렬이 되며, 외부윤곽선 연결요소와 내부윤곽선 연결요소는 따로 분리되어 저장된다.

2.1.2 구분선 추출

그림 1의 문서영상 요소의 구분트리가 분리하고자 하는 대상 중에서 가장 처음으로 분리하는 것은 구분선이다. 모든 외부윤곽선 연결요소 중 가로 대 세로 또는 세로 대 가로의 비가 10이상이고 일정크기 이상이면 구분선으로 분리되며, 문자의 단구분이나 줄구분을 위하여 사용된다. 구분선의 존재를 판단하는 구체적인 조건은 아래와 같다. 여기서 CM은 하나의 연결요소를 나타낸다.

<u>가로구분선으로 판단되는 조건</u>:

$$(CM.width > CM.depth * 10 \text{ and } CM.depth < AVG\_CM * 0.5 \text{ and } CM.width > AVG\_CM * 3) \text{ or } (CM.width > CM.depth * 15 \text{ and } CM.depth > AVG\_CM)$$

<u>세로구분선으로 판단되는 조건</u>:

$$(CM.depth > CM.width * 10 \text{ and } CM.width < AVG\_CM * 0.5 \text{ and } CM.depth > AVG\_CM * 3) \text{ or } (CM.depth > CM.width * 15 \text{ and } CM.width > AVG\_CM)$$

2.1.3 연결요소의 크기별 그룹핑

구분선이 추출된 후 나머지 연결요소는 크기별로 그룹핑되는데, 이는 비슷한 크기의 연결요소를 서로 그룹지어 같이 처리하기 위해서이다. 즉, 이 단계에서는 비슷한 크기를 갖는 문자, 또는 그래픽 요소를 하나의 그룹으로 묶어주는 역할을 수행한다. 그림 2는 크기에 의한 연결요소의 그룹핑 방법을 설명하고 있다.

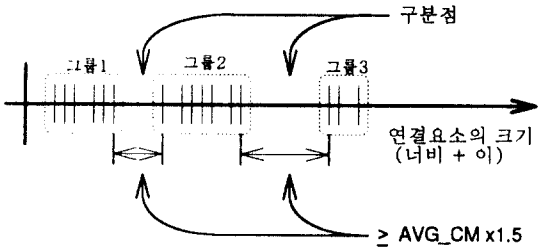


그림 2. 크기에 의한 연결요소의 그룹핑
Fig 2. Grouping of connected components by size

그림 2에서 볼수 있듯이 크기로 정렬된 연결요소의 대열에서 크기에 의한 그룹의 구분점은 서로 인접한 두 연결요소 간의 크기 차이가 AVG_CM의 1.5배 이상인 곳이 되며, 이러한 크기의 구분점에 의해 연결요소는 몇 개의 그룹으로 분리된다.

2.2 문자열, 도표, 그래픽 추출

2.2.1 문자열 구성

실제적인 문서구조 분석을 위한 첫 단계로 그룹핑된 연결요소로부터 문자열을 구성한다. 먼저 그룹내의 연결요소를 최소 x좌표값 순으로 정렬을 수행한 후, 연결요소의 y좌표값이 중첩되는 조건을 만족하는 연결요소들을 같은 문자열로 구성한다. 이때 두 연결요소간의 x좌표값이 임계치보다 작은 경우에 한하여 같은 문자열로 확장해 가기 때문에 단순한 수평투영에 비하여 문서의 기울어짐에 강한 특성을 갖게 된다. 그러나 두 연결요소 사이에 구분선이 존재하는

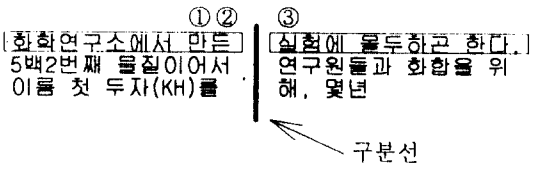


그림 3. 연결요소의 문자열 구성
Fig 3. Composing text lines from connected components

경우에는 서로 다른 문자열이 된다. 그림 3의 연결요소 ①과 ②는 구분선에 의해서 나뉘지 않으므로 같은 문자열로 구성될 수 있지만, ②와 ③은 구분선에 의해서 나뉘어지므로 같은 문자열로 구성될 수 없다.

2.2.2 도표 추출

앞에서 구성된 문자열을 제외한 나머지 요소로부터 도표와 그래픽 요소를 추출한다. 일반적으로 도표는 사각형의 형태를 갖고 있으므로 사각형 모양을 한 요소를 골라내어 도표인지 확인을 해야한다. 연결요소의 윤곽선에 위치한 점들 중에서 $x+y$, $-x+y$, $x-y$, $-x-y$ 각각의 값이 최소인 점을 찾아, 그 점들이 연결요소의 최소외접사각형의 각 꼭지점과 이루는 거리를 계산하여 사각형인지를 판단한다[그림 4]. 이 방법은 문서가 10° 이내로 기울어져 있는 경우에 적용할 수 있다. 사각형 조건을 만족할 경우 그 요소의 내부에 포함되는 모든 내부윤곽선 연결요소를 찾는다. 이때 내부윤곽선 연결요소가 여러개 존재하면서 서로 분리되어 있으면 도표로 확정한다. 도표추출 알고리즘을 슈도코드로 정리하면 다음과 같다.

[도표 추출 알고리즘]

Step 1 : 연결요소를 크기 순으로 정렬한다.

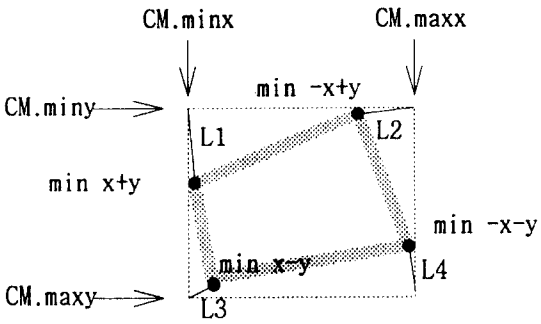
Step 2 : 연결요소 중 사각형이 될 수 있는 것을 추출한다.

Step 3 : 사각형 연결요소의 내부 연결요소를 찾는다.

Step 4 : 사각형 연결요소와 내부 연결요소의 위치관계를 확인한다.

Step 5 : 이웃한 사각형 연결요소를 병합하여 도표를 구성한다.

도표로 확정이 되었으면 도표의 내부요소를 결정하게 된다. 먼저 최외각에 존재하게 되는 외부윤곽선 연결요소인 도표 전체에 포함되는 그래픽, 구분선, 문자열과 나머지 요소를 모두 골라낸다. 여기에서 골라진 문자열등의 요소는 도표를 구성하는 요소들로서, 이들은 다시 도표 내부 셀과의 포함관계에 의해서 다시 구분된다. 이 때 그래픽 요소를 하나라도 가지게 되는 셀은 그래픽셀이 되어 내부의 모든 요소를 그래픽으로 처리하게 되며, 이는 그래픽 요소를 전혀 갖지 않는 텍스트셀과 구분된다.



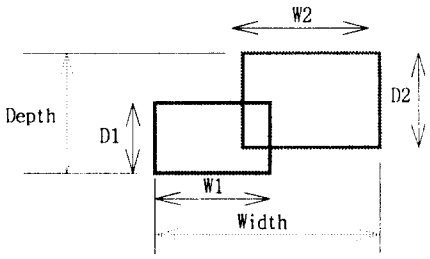
if (L1+L2+L3+L4 < TH) then 사각형

그림 4. 연결요소의 사각형 판단조건

Fig 4. A constraint of Rectangular shape of connected components

2.2.3 그래픽 추출

구분선과 도표가 추출된 그림영역의 나머지는 실제의 그래픽 요소가 된다. 따라서 그래픽 요소를 완성하기 위하여 그림 5와 같이 먼저 그래픽 요소 중 서로 겹치는 것을 합병하여 크기를 키워나간다. 더 이상 합병되는 그래픽 요소가 없으면 합병된 그래픽 요소에 겹쳐지는 도표, 문자열과 나머지 요소를 그래픽의 내부요소로 간주하여 합병한다.



If (D1+D2 < Depth*1.2 && W1+W2 < Width*1.2) then 합병

그림 5. 그래픽 요소의 합병

Fig 5. Merging of graphic elements

Ⅲ. 한·영 혼용문서의 낱자분리 방법의 제안

3.1 한·영 혼용문서에서의 낱자분리 문제점

한글문서나 영어문서에서 문서구조분석 방법은 크게 다르지 않다. 하지만 가장 커다란 차이를 보이는

부분은 낱자분리 부분이다. 왜냐하면 그림 6에서 볼 수 있듯이 한글의 구조는 영어의 구조와 근본적으로 다르기 때문이다. 영어는 'i', 'j'를 제외하고는 낱자가 모두 하나의 연결요소로 구성되어 있는 반면, 한글은 여러개의 연결요소가 조합되어 낱자 한 자를 구성하게 된다.

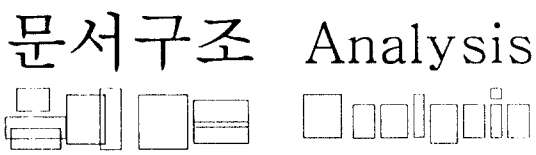


그림 6. 한글과 영어의 낱자형태
Fig 6. Character shapes of Hangul and English

이러한 특성때문에 영어문서의 낱자분리시 처리해야 할 것은 접촉문자(touching character)에 국한되지만, 한글문서는 접촉문자와 더불어 자소로 분리되어 추출된 연결요소의 합병도 고려해야 한다[12, 15, 16, 19]. 하지만 이러한 자소의 합병은 결코 쉬운 문제가 아니다. 자소조합 형태에 따라 한글의 유형은 그림 7과 같이 분류할 수 있다. 그림 7에서 볼 수 있듯이 3, 4번 유형은 수직방향으로 투영하여 겹치는 연결요소를 합병하면 완전한 낱자로 구성되며, 2, 6번 유형은 글꼴에 따라서 낱자로 구성될 수 있다. 하지만 1, 5번 유형은 수직방향의 투영으로 합병이 불가능하다. 1, 5번 유형에 나타나는 연결요소를 단순히 합병할 수 없는 이유는 한 문서에 한글, 영어, 숫자 등이 동시에 나타나기 때문이다.

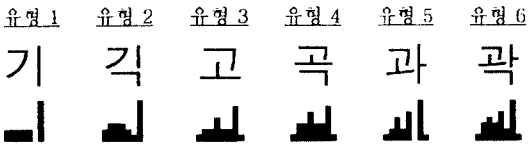


그림 7. 한글의 유형과 수직투영 형태
Fig 7. Six types of Hangul and corresponding shapes of vertical projection

그림 8은 한글 '이', 숫자의 '01', 영문자의 'ol'이 모두 유사함을 보이고 있다. 한글 '이'의 합병은 문자에 대한 정보없이는 해결될 수 없으므로 궁극적으로는 문자인식과 더불어 해결해야 할 과제이다. 하지만 본 논문에서는 단어에 대한 정보를 이용하여 이 문제를 효과적으로 해결할 수 있는 방법을 고려하였다.

그림 8. 유사한 연결요소 형태
Fig 8. Some similar shapes of connected components

한·영 혼용문서에서는 이러한 자소합병의 문제만이 아니라 접촉문자의 분리에서도 발생하게 된다. 그림 9와 같이 여러개의 낱자가 서로 접촉되어 이웃한 문자와 일부 혹은 전체가 접촉된 경우 접촉 문자의 갯수를 추정하는 것이 쉽지 않다. 또한 한글 낱자 중 일부 자소가 이웃한 자소와 접촉되어 있는 경우에는 단순한 분할로서 이를 완전한 낱자로 구분하는 것은 매우 어려운 문제이다.

그림 9. 접촉된 자소의 형태
Fig 9. Touched shapes of characters

지금까지 살펴본 바와 같이 한글은 그 형태상의 특성때문에 영어에서 발생할 수 없는 문제점을 갖고 있으므로, 한·영 혼용문서에서 낱자를 분리하는 것은 한글과 영어의 특성을 모두 고려하여 수행해야 한다. 따라서 본 논문에서는 연결요소를 분석하여 단어를 구성한 뒤, 각각의 단어에 대하여 한글단어와 영어단어로 나누고, 한글단어에 대해서 한글의 특성에 맞는 자소합병과 접촉문자의 분리 방법을 제안하므로써 한글과 영어가 혼용된 문서에 대한 효율적인 분석 결과를 얻을 수 있도록 하였다.

3.2 한·영 혼용문서에서의 낱자분리

3.2.1 낱자구성과 단어구성

문서구조분석에서 완성된 문자열내의 연결요소로부터 낱자와 단어를 구성하게 된다. 이를 위하여 문자열 내의 연결요소는 최소 x좌표값으로 정렬된 후, 인접한 두 요소의 위치관계와 거리에 따라 일정 조건을 만족하면 낱자로 합병하게 된다. 낱자가 구성된 후, 새로이 구성된 낱자들 간의 간격에 의해서 다시 단어로 구분된다.

그림 10-①과 같이 x축으로 겹쳐진 부분(MERGED_X)이 낱자후보의 너비, 합병될 연결요소의 너비에 따라 겹치는 조건을 만족하면 낱자로 구성되며, 또한 ②와 같이 낱자간 간격(GAP)이 문자열내 평균간격(AVG_GAP)의 1.7배 보다 크면 다른 단어로 분리된다.

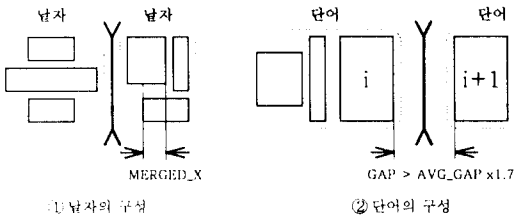


그림 10. 낱자와 단어의 구성
Fig 10. Merging to character and word

3.2.2 단어의 한·영 구분

문자열내의 연결요소는 일단 낱자와 단어로 구분되어 졌지만, 이 낱자구분은 완벽한 것이 아니다. 그 이유는 3.1절에서 설명한 바와 같이 한글과 영어는 그 구조때문에 낱자분리에 차이를 보이기 때문이다. 그림 11은 입력문서에 대하여 낱자와 단어구분을 수행한 결과이다. 이 그림에서 볼 수 있듯이 '서울'의 '서'는 'ㅅ'과 'ㅓ'로 따로 구분되어 있고, 'Seoul이'의 '이'도 'ㅇ'과 'ㅣ'로 따로 분리되어 있다. 하지만 사람이 인식하듯이 쉽게 'ㅇ'과 'ㅣ'를 '이'로 합병할 수 없는데, 그 이유는 그림 8의 문제점과 같이 문맥의 이해와 어울어지지 못하면 구분할 수 없기 때문이다.

이 문제를 해결하기 위한 한·영 단어구분 방법을 제안하기로 한다. 한·영 단어구분을 위해서 낱자의 모양을 구분해본 결과 그림 12와 같이 16개의 서로 다른 특성을 갖는 형태로 정의하게 되면 모든 모양을 포함하게 되는 것을 발견하였다.

그림 11. 초기의 낱자구분 결과
Fig 11. The result of initial character segmentation

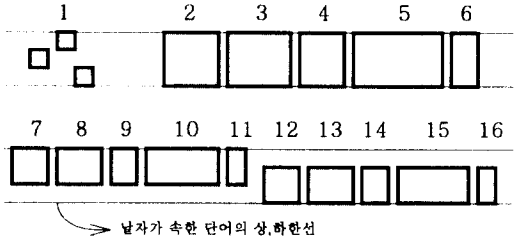


그림 12. 낱자의 형태
Fig 12. The shapes of characters

이에따라 다음과 같은 낱자 형태정의 알고리즘을 구성할 수 있다. 알고리즘에서 CH.width, CH.depth는 낱자의 폭과 높이를 나타내고, CH.short는 폭과 높이 중에서 짧은 부분, CH.long은 긴부분을 나타낸다.

[낱자 형태정의 알고리즘]

크기 정의:

if CH.depth > TL.depth0.7 **then** size = 1
else if CH.depth > TL.depth0.33 **then** size = 2
else size = 3

위치 정의:

if TL.miny-CH.midy > TL.depth0.4 **then** pos = 1
else if TL.miny-CH.midy > TL.depth0.8 **then** pos = 2
else pos = 3

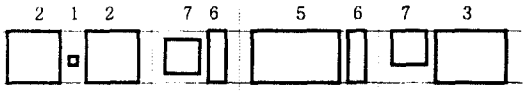
장평정의:

if CH.short/CH.long > 0.8 **then** ec = 1
else if CH.width/CH.long > 0.5 **then** ec = 2
else if CH.depth/CH.width > 0.5 **then** ec = 3
else if CH.width > CH.depth **then** ec = 4

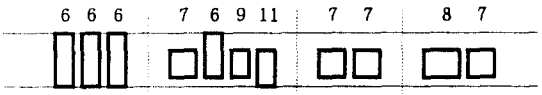
```
    else if CH.depth > CH.width then ec = 5

낱자 형태결정:
    if size == 3 then TYPE = 1
    else if size == 1 then TYPE = ec + 1
    else if pos < 3 then TYPE = ec + 6
    else TYPE = ec + 11
```

수직투영에 의해서 일차로 구성된 각 낱자는 크기, 위치, 장평이 조사되고, 이 세가지 조건에 의해서 형태가 결정된다. 그림 13은 한글단어와 영어단어가 갖는 낱자의 배치형태를 나타내고 있으며 단어를 한글과 영어로 분류하는 알고리즘은 다음과 같다.

```
[단어를 한글과 영어로 분류하는 알고리즘]

Type(p):앞 글자의 낱자유형, Type(c):현재 글자의 낱자유형,
Type(n):다음글자의 낱자유형
초기값:Type(p) = 1;count = 0;

do {
   if (Type(p) == 1 and Type(c) >= 11)
      then 영어단어로 판단;
   if (Type(p) != 1 and Type(c) > 11)
      then 영어단어로 판단;
   if ((Type(p) == 6 or Type(p) == 11)
      and (Type(c) == 6 or Type(c) == 11))
      then {count ++; if (count >= 2)
      then 영어단어로 판단}
   if ((Type(c) < 11 and 앞낱자와의 높이차 > 임계치)
      then 영어단어로 판단
   if ((Type(p) > 6 and Type(p) < 10) and (Type(c) > 6
      and Type(c) < 10) and Type(n) != 6)
      then 영어단어로 판단
} while (c != 단어내의 마지막 낱자)
한글단어로 판단;
```



① 한글단어의 낱자배치 모양



② 영어단어의 낱자배치 모양

그림 13. 한·영 낱자의 배치모양
Fig 13. Character arrangements of Hangul and English

3.2.3 단어형태에 의한 낱자수정

단어의 형태가 '한글'이라고 결정되면 한글단어에 대하여는 낱자수정 과정을 거치게 된다. 낱자수정 과정에는 잘못 합병된 낱자를 다시 합병하는 과정과 한글의 접촉문자를 분리하는 두 가지 과정이 있다. 만약 단어의 형태가 '영어'라고 결정되면 낱자수정 과정을 거치지 않고 바로 접촉문자 분리 과정을 수행한다. 영어의 경우는 대부분 외각 연결요소가 하나의 낱자를 구성하고, 'i', 'j'와 같이 분할된 연결요소도 연결요소의 x좌표값이 중첩된다. 그러므로 영어단어의 경우는 별도의 병합과정 없이 접촉된 문자의 갯수를 추정하여 분할한다.

그림 7에서 분류한 한글낱자의 유형 중 1, 5번 유형은 초기 낱자분리 방법으로는 자소가 하나의 낱자로 구성될 수 없다. 따라서 한글단어라고 판단된 단어에 대하여 1, 5번 유형의 분리되어 있는 자소를 합병하여 완전한 낱자로 재구성한다. 그림 14에는 합병대상이 되는 요소를 표시해 놓았는데 ①, ③번은 그림 7의 1, 5 유형과 같이 원래의 낱자 한 자 내에서 분리되어 존재하는 자소를 합병하는 것이고, ②, ④번은 그림 9-②와 같이 떨어진 낱자내의 자소가 앞뒤의 자소와

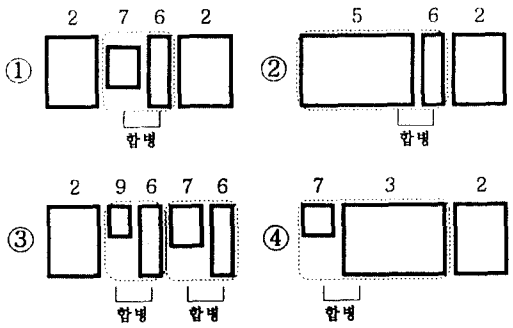


그림 14. 한글단어의 자소합병
Fig 14. Merging characters to word in Hangul

접촉하게 되어 커다란 연결요소로 존재하게 된 형태이다. 이런 형태는 일단 합병을 시킨 뒤 이어질 접촉문자의 분리과정에서 처리하게 된다.

접촉문자는 대부분 그림 12의 낱자유형 중 3번과 5번의 모양을 지니게 된다. 자소합병 후에는 이러한 접촉문자를 분리해 주어 낱자수정 과정을 마치게 된다. 아래의 수식은 한글단어의 낱자유형 중 3, 5번 형태의 낱자가 몇 개의 낱자접촉에 의해 생긴 것인지, 접촉문자의 낱자 갯수를 예측한다. TL.depth는 낱자가 속한 문자열의 높이를 나타낸다.

$$\text{접촉문자의 낱자갯수} = \lfloor \frac{CH.width + TL.depth * 0.33}{TL.depth * 0.8} \rfloor$$

여기서 보면 낱자의 높이에 대한 너비의 비율을 80%로 정하여 접촉문자의 낱자 갯수를 예측하였다. 이렇게 조사된 낱자의 갯수가 2~6개인 경우에만 접촉문자로 결정하여 균일한 크기로 분리하고, 그 외의 경우는 접촉 가능성이 희박하므로 처리하지 않는다.

3.2.4 문자블럭의 구성

낱자수정까지 끝난 후, 제안하는 문서분석의 마지막 단계로 문자열을 합병하여 문자블럭을 구성한다. 문서내의 모든 문자열은 최소의 y좌표값 순으로 정렬된 후, 그 순서대로 수직투영의 겹침에 의해서 문자블럭으로 구성된다. 그림 15에서 볼 수 있듯이 새로 구성되는 문자블럭(TB)과 문자열(TL)의 간격(GAP)이 문자블럭의 최상위 문자열(Top TL) 높이의 2배 이하이고 TB와 TL이 x축으로 겹친 MERGED_X가 Top TL 또는 TL의 3분의 1 이상이면 TL은 TB와 같은 문자블럭으로 합병된다.

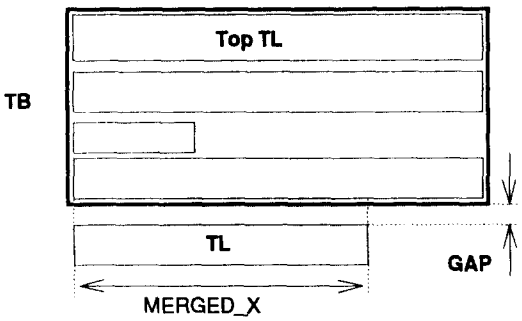


그림 15. 문자블럭의 구성
Fig 15. Merging to character block

Ⅳ. 실험결과 및 분석

제안한 시스템은 SPARC station IPX에서 C언어로 구현하였으며, 총 56페이지에 달하는 여러 종류의 문서에 대하여 분석실험을 하였다. 표 1은 실험에 사용된 문서에 대한 자료이다

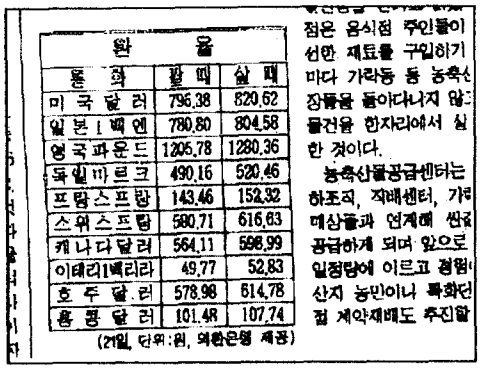
표 1. 실험대상 문서
Table 1. Documents used in expements

	화일이름	문 서 종 류	문서크기
여러가지 문서	PAPER	2단으로 된 영어논문 일부	1189×796
	NEW1	신문일부	962×733
	NEW2	신문일부	1334×858
	TABLE1	도표	1624×745
	TABLE2	도표	900×733
	MEMOPAD	메모지	1241×475
낱자분리	B414~B430	책 한쪽씩 17쪽 (패턴인식론, 오영환)	1089×1682
구조분석	V1~V24	책 한쪽씩 24쪽 (Computer Vision, Ballard)	1158×1792
	D1~D9	책 한쪽씩 9쪽 (Digital Image Processing Methods, E.R. Dougherty)	1076×1613

위의 실험대상 문서 중 윗쪽 6개의 문서는 다양한 문서에 대한 분석실험을 위한 것이며, B414~B430는 한·영 혼용문서의 낱자분리 실험을 하기 위한 문서이다. 그리고 아래의 V1~V24와 D1~D9는 문자영역과 그림영역의 분리실험을 위한 데이타이다.

다음의 그림 16은 표 1의 문서 데이타 중 'NEWS1'에 대하여 분석한 결과를 나타낸다. 다음의 그림에서 굵은 선의 사각형으로 된 부분은 도표를 나타내며, 가는 선의 사각형은 각각의 분리된 요소를 나타낸다. 또한 낱자에서 빗금으로 표시된 것은 접촉문자인 것을 분리해낸 결과이다.

표 2는 문서(B414~B430) 17페이지에 대한 낱자분리 실험결과이고, 표 3은 표 2의 TEXT에 해당하는 부분 중 프로그램 코드가 나타난 부분을 제외한 순수한 문자부에서 한글과 영어가 차지하는 빈도와 각 단어의 분리율을 나타낸 것이다.



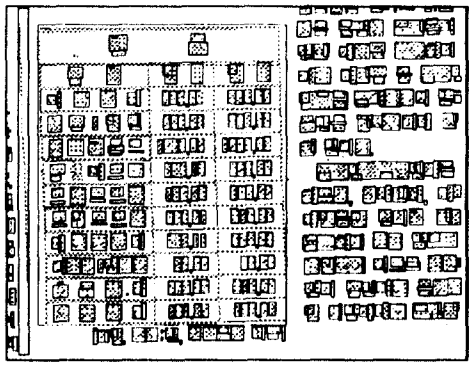
환 율		
통 화	팔 때	살 때
미 국 달 러	796.38	820.62
일 본 1 백 엔	780.80	804.58
영 국 파 운 드	1205.78	1280.36
독 일 마 르 크	490.16	520.46
프 랑 스 프 랑	143.46	152.32
스 위 스 프 랑	580.71	616.63
캐 나 다 달 러	564.11	598.99
이 태 리 1 백 리 라	49.77	52.83
호 주 달 러	578.98	614.78
홍 콩 달 러	101.48	107.74

(21일, 단위:원, 외환은행 제공)

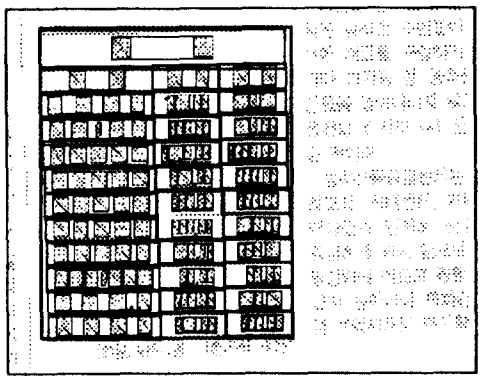
점은 음식점 주인들이 선한 재료를 구입하기 마다 가락동 등 농축산 장들을 돌아다니지 않 물건을 한자리에서 실 한 것이다.

농축산물공급센터는 하조직, 직배센터, 가 매상들과 연계해 싼값 공급하게 되며 앞으로 일정량에 이르고 경험 산지 농민이나 특화단 접 계약재배도 추진할

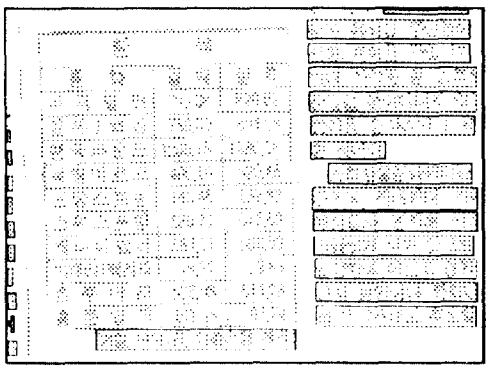
(a) 영상데이타



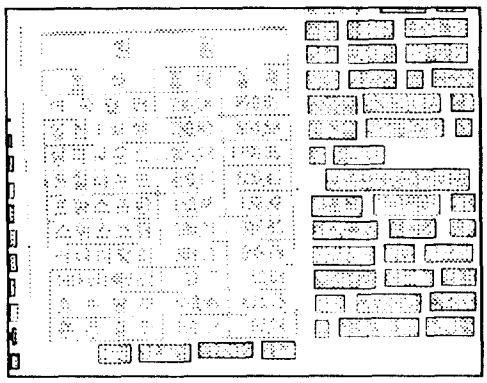
(b) 연결요소 추출결과



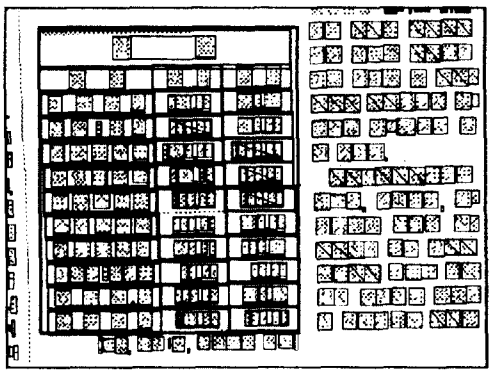
(c) 도표추출 결과



(d) 문자열 추출결과



(e) 단어추출 결과



(f) 낱자추출 결과

그림 16. NEWS1에 대한 문서분석 및 낱자분리 실험결과
Fig 16. Experimental results on 'NEW1' document in structure analysis and character segmentation

표 2. 문서의 낱자분리 실험결과
Table 2. Experimental results on character segmentation

	TEXT	TABLE	전체
문자블럭수	83		83
문자열수	448	128	576
단어수	3,007	271	3,278
합병낱자수	1,812	136	1,948
분리된낱자	21	4	25
낱자수	9,748	636	10,384
총 낱자수	9,748	636	10,384
낱자오류	94	0	94
분리율	99%	100%	99.1%

표 3에서 한 단어란 공백이 없이 연결된 문자열을 의미하는 것으로 한 단어내에 한글과 영어가 모두 나타나는 경우에는 낱자가 많은 쪽으로 단어의 유형을 결정하여 단어수를 산정하였다. 표 3에서 알 수 있듯이 한글이 영어로 오분류되는 경우는 거의 없었으나, 상당수의 영어가 한글로 오분류 되었다. 영어가 한글로 오분류된 경우는 대문자만으로 쓰여진 약어로 이것이 전체오류의 85%를 차지하고 있다. 그러나 대문자만으로 구성된 영어단어가 한글로 오분류될 경우에는 영어낱자가 한글과 같은 형태를 갖기 때문에 대문자 'I', 'J'를 제외하고는 대부분 정확한 낱자분리로 이어진다.

표 3. 문서의 단어분리 실험결과
Table 3. Experimental results on word segmentation

	한 글	영 어	전 체
단어수	2429	83	2512
정분류	한글→한글(2428)	영어→영어(56)	2484
오분류	한글→영어(1)	영어→한글(27)	28
단어분리율	99.96%	67.47%	98.89%

표 4는 문서(V1~V24, D1~D9) 33페이지에 대한 문자영역과 그림영역의 추출실험을 수행한 결과이다. 분석속도는 페이지당 평균 10여초가 소요되었다.

제안한 시스템은 문서의 기울어짐을 어느 정도 흡수할 수 있으나, 문서의 기울어짐에 영향을 받는 부분은 한·영 단어로의 형태결정과 한글단어에 대한 낱자구분을 수행하는 곳이다. 즉, 기울어진 문서에 대하여 각 낱자의 형태파악의 잘못으로 단어의 형태결정에 오류가 생겨, 결국 낱자분리까지 그 오류가 이어지게 되었다. 또한 한글의 낱자모양에 대해 세로크기에 대한 가로크기의 비를 80%로 정하였기 때문에, 이로 인해 낱자를 분리하는 경우에 오류가 발생할 수 있으며, 글꼴에 따라서는 그림 10에서 정의한 낱자형태와 일치하지 않을 수도 있으므로 역시 오류가 발생할 수 있다. 하지만 일반적 형태의 글꼴을 가진 문서가 10° 이하로 기울어진 상태로 입력되면 만족할 만한 분석결과를 얻을 수 있으며, 비교적 복잡한 문서에 대해서도 그림과 문자는 물론 도표까지 구분할 수가 있다. 또한, 본 논문에서 제안한 한·영 단어로의 형태구분에 의한 단어의 낱자분리는 문자인식이 없이도 한·영 혼용문서의 낱자분리에 좋은 결과를 보여주었다.

표 4. 문서의 영역추출 실험결과
Table 4. Experimental results on region analysis

	문자영역	그림영역	전 체
영역 수	299	167	466
바르게 추출된 영역수	286	166	452
추출률	95.6%	99.4%	97.0%

V. 결 론

문서의 형태로 저장되어 있는 많은 양의 정보는 주로 광파일의 형태로 저장되어 정보검색이 매우 어렵게 되어있다. 이러한 정보를 전산화하고 정보검색을 효율적으로 수행하기 위해서는 문자인식 기술 이전에 문서분석 기술이 반드시 선행되어야 한다. 단순하게 낱자만을 인식할 수 있는 문자인식기에 문서분석 기술을 도입하면 다양한 문서로부터 낱자를 추출하여 인식기로 처리할 수 있을 뿐만 아니라, 문서내의 여러 가지 요소를 세밀하게 분석하여 체계적으로 데이타 베이스화 할 수 있다.

본 논문에서는 한·영 혼용의 다양한 문서를 문서구

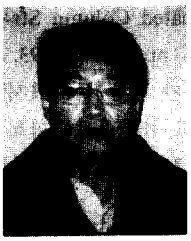
조 분석에서 널리 사용되고 있는 연결요소분석 방법에 의하여 문자, 그래픽, 도표, 등으로 분석하기 위한 방법을 제안하였다. 또한, 일반적인 한·영 혼용의 문서에 대하여 한글의 낱자분리를 정확히 수행하기 위하여 각 단어를 한글단어와 영어단어로 그 형태를 결정한 후, 한글단어에 대하여 한글의 특성에 맞는 낱자합병과 접촉문자분리를 수행하는 방법을 제안하였다. 이 방법을 일반적인 책 17페이지에 대하여 실험한 결과 낱자분리에 99.1%의 정확성을 보였으며, 구조분석을 실험한 결과 33페이지의 문서에서 95.6%의 문자영역 추출률과 99.4%의 그림영역 추출률을 얻었다. 따라서, 본 논문에서 제안한 문서분석 방법으로 일반적인 문서를 대상으로 한·영 혼용문서에서의 구조분석 뿐만 아니라 한글 낱자분리에도 좋은 성능을 기대할 수 있다. 그러나 한글 횡모음 'ㅣ'와 유사한 형태를 갖는 숫자 '1'과 특수문자 '(', '[' 등이 한글 단어에 붙어서 나타날 경우 종종 낱자 분할에서 오류가 발생한다. 따라서 향후 한글이 숫자나 특수문자 등과 결합되어 나타날때 이를 효과적으로 분할하는 방법에 대한 연구가 보완되어야 할 것이다.

참 고 문 헌

1. S. Mori, C. Y. Suen, and K. Yamamoto, Historical Review of OCR Research and Development, Proc. of the IEEE, Vol. 80, No. 7, pp. 1029-1058, 1992.
2. Tin Kam Ho, H. S. Baird, Perfect Metrics, Proc. of the ICDAR93, pp. 593-597, 1993.
3. S. Kahan, T. Pavlidis, and H. S. Baird, On the Recognition of Printed Characters of Any Font and Size, IEEE tran. on PAMI, Vol. 9, No. 2, pp. 274-288, 1987.
4. 장명욱, 천대녕, 양현승, "연결화소를 이용한 문서영상의 분할 및 인식", 정보과학회논문지, 제20권, 제12호, pp. 1741-1751, 1993.
5. Y.Y. Tang, C.Y. Suen, "Document Structures: A Survey", Proc. of the ICDAR93, pp. 99-102, 1993.
6. Hong Guo, et al., Realization of A High-Performance Bilingual Chinese-English OCR System, Proc. of the ICDAR95, pp. 978-981, 1995.
7. A. Dengel, "Initial Learning of Document Structure", Proc. of the ICDAR93, pp. 86-90, 9. 1993.
8. D. Wang and S. N. Srihari, "Classification of Newspaper Image Block Using Texture Analysis", CVGIP, Vol. 47, pp. 327-352, 1989.
9. L.A. Fletcher and R. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, pp. 910-918, 1988.
10. F.M. Wahl, K.Y.Wong and R.G.Gasey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents", CVGIP, pp. 375-390, 1982.
11. Y. Hirayama, "A Block Segmentation Method for Document Image with Complicated Column Structures", Proc. of the ICDAR93, pp. 91-94, 1993.
12. 이동준, 이성환, "한글 및 영숫자 혼용 문서에서의 문자분할 및 인식", 1994년도 한국정보과학회 가을 학술발표논문집, Vol. 21, No. 2, pp. 403-406, 1994.
13. J. Schurmann et al., "Document Analysis-From Pixels to Contents", Proc. of the IEEE, Vol. 80, No. 7, pp. 1101-1119, 1992.
14. H.Fujisawa, Y.Nakano and K.Kurino, "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis", Proc. of the IEEE, Vol. 80, No. 7, pp. 1079-1092, 1992.
15. 최봉희, 이인동, 김태균, "문자영역 추출과정에서의 오분리의 교정", 정보과학회논문지, 제21권, 제1호, pp. 86-93, 1994.
16. 이승형, 전종익, 조용주, 남궁재찬, "신문 자동인식 시스템을 위한 문자의 분류에 관한 연구", 1989년도 한글 및 한국어정보처리 학술발표논문집, pp. 209-215, 1989.
17. S. Tsujimoto and H. Asada, "Major Components of A Complete Text Reading System", Proc. of IEEE. Vol. 80, No. 7, pp. 1133-1149, 1992.
18. 박문규, 권영빈, "다양한 결합문자를 갖는 계층지도의 인식", 제2회 문자인식 워크샵, pp. 244-256, 1994.
19. 김의정, 김태균, "인쇄체 문서인식을 위한 문자

추출에 관한 연구", 제2회 문자인식 워크샵, pp. 171-179, 1994.

김 민 기(Min-Ki Kim) 학생회원
1966년 11월 7일생
1989년 2월 : 중앙대학교 전자계산학과 졸업(이학사)
1992년 7월 : 중앙대학교 대학원 전자계산학과 졸업(공학석사)
1994년 9월~현재 : 중앙대학교 대학원 컴퓨터공학과 박사과정 재학
※주관심분야 : 문자인식, 패턴인식, 영상처리, 유전자 알고리즘



권 영 빈(Young-Bin Kwon) 종신회원
1955년 10월 24일생
1978년 2월 : 아주대학교 전자공학과 졸업(공학사, 전 교수석)
1981년 2월 : 한국과학기술원 졸업(공학석사)
1986년 1월 : 프랑스 파리 ENST 졸업(공학박사)
1986년 3월~현재 : 중앙대학교 컴퓨터공학과 교수
※주관심분야 : 문자인식, 패턴인식, 컴퓨터비젼, 인공지능, 멀티미디어, HCI

한 상 용(Sang Yong Han) 정회원
1952년 10월 2일생
1975년 2월 : 서울대학교 공과대학 졸업(공학사)
1984년 6월 : University of Minnesota 컴퓨터공학과 졸업(공학박사)
1995년 3월~현재 : 중앙대학교 컴퓨터공학과 부교수
1977년 1월~1978년 12월 : KIST 시스템공학센터 연구원
1984년 6월~1995년 12월 : IBM DSD 및 Watson 연구소 연구원
1985년 9월~1988년 2월 : Rensellaer Polytechnique Institute 겸임교수
1996년 1월~1996년 8월 : AMD Visiting Scholar
※주관심분야 : VLSI 및 CAD, Virtual Prototyping, 알고리즘