論 文

문서의 영역분리와 레이아웃 정보의 추출

正會員 趙 鎔 周* 正會員 南宮 在 贊*

The Block Segmentation and Extraction of Layout Information in Document

Yong Joo Cho* Jae Chan Namkung* *Regular Members*

要 約

본 논문은 이미 출판된 문서를 대상으로 문서의 영역을 분리하고, 문서의 구성요소와 레이아웃 정보를 추출하는데 새로운 알고리즘을 제안한다.

먼저 300 dpi로 입력된 문서에서 문서를 이루는 각 요소를 영역화 하기 위하여 레이 블링과 블럭화 작업을 행한다. 둘째로 블럭화된 문서의 각 요소를 대상으로 부분영역으로 분리를 수행한다. 세째로 추출된 부분영역에서 그림영역을 추출하고 문자영역에 대해서는 문자열 추출 및 개별 문자 추출을 한다. 마지막으로 이렇게 추출된 정보로 문서의 레이아웃 인식을 위한 정보를 추출하였다.

실험은 어느정도의 형식을 가진 학회 논문지를 대상으로 하였으며, 문자와 그림 영역의 분류 및 문자열 추출에 대해서 98.5%의 성공율을 얻고, 레이아웃 인식을 위한 정보의 추출에서도 98%의 성과를 보였다.

ABSTRACT

In this paper, we suggest a new algorithm applied to the segmentation of published documents to obtain constituent and layout information of document.

Firstly, we begin the process of blocking and labeling on a 300dpi scanned document. Secondly, we classify the blocked document by individual sub-regions. Thirdly, we group sub-regions into graphic areas and text areas. Finaly, we extract information for layout recognition by using the data.

From an experiment on papers of an academic society, we obtain the above 98% of region classification rate and extraction rate of information for the layout recognition.

I. 서 론

최근에 들어와 정보 사회의 급진전은 다양한 분야에서 매우 빠르게 성장하고 있다. 이러한 분야로서는

*光云大學校 電子計算機工學科
Dept. of Computer Engineering, Kwangwoon University
論文番號 : 92-113(接受1992. 6. 27)

사무자동화가 주류를 이루고 있으며 주로 문서를 생성하는 측면에서 많은 발전이 있었다. 여기에서 문서(Document)가 차지하는 비중은 DTP(Desk Top Publishing) 시스템과 워드프로세서(Wordprocessor)의 성장과 함께 모든 사무실 또는 인간이 생활하는 모든 환경에서 필수요소가 되어버렸으며 폭주하는 문서를 보관 또는 재입력하기 위해서 인간은 많은 노력을 하고 있다.

본 논문의 목적은 이미 출판 또는 출간된 문서를 대상으로 문서에서 문자와 그림영역을 추출한뒤, 추출된 블럭정보를 이용하여 문서의 레이아웃 정보를 추출하는 것이다.

문서인식에 관한 연구는 문서의 레이아웃 정보 추출, 문자의 구별, 문자인식 그리고 인식된 정보의 표현으로 나누어 질 수 있는데, 본 논문은 레이아웃 정보의 추출 부분에 해당된다. 그림 1에는 문서인식의 전체적인 개략도를 보인다. 문서의 인식에 관한 연구는 방대한 데이타량과 빠른 속도를 요구하기 때문에 연구 및 개발에 있어서 많은 문제점이 있었으나, 지금에 와서는 하드웨어 및 소프트웨어 기술의 발달에 따라 실용화 되어지고 있다. 대표적인 문서인식 시스템을 예로들면 미국 뉴욕 주립대학교의 Wang과 Srihari[1]는 신문을 중심으로 RLSA(Run Length Smoothing Algorithm)을 적용하여 문서를 영역화하는 연구를 하였다. 처리에 이용된 컴퓨터는 SUN 3/60이었고 결과는 매우 만족할만한 수준이었다. 이 팀은 이후에도 많은 연구를 하였으며 문서인식의 선두주자로 인식되어진다. 미국에서는 주로 기업에의하여 많은 상품이 발표[2]-[10]되고 있으며, 영문자가 단순하다는 장점아래 CALERA사의 Wordscan나 Logitech사의 CatchWord등은 이미 시스템으로 출하되어 사용자들의 구미를 만족시켜 주고 있다. 현재 이러한 시스템은 MS-WINDOW 버젼으로 발표되고 있으며, 다양한 파일 포맷을 임포트(import)하고 있다. 일본에서는 교토대학의 Sakai[11]교수를 시발로 NTT[12], Fujitsu등에 의하여 문서인식 시스템이 개발되었고, 특히 Fujitsu의 ATLAS-JK은 문서인식 기능에 기계번역시스템까지 구비하여 판매되고 있는 실정에 있다. 우리나라에서는 그동안 연구적인 측면에서 많은 연구가 있었고, 특히 주식회사 코테크의 리덱스 시스템은 SUN SPARC 2 시스템에서 문서와 문자인식 시스템을 개발하여 시판하고 있다. 그러나 가격이 매우 고가여서 사용자가 이용하기에는 매우 불리한 여건에 있다.

본 연구에서는 MS-WINDOW 상에서 실행되는 문서인식 시스템을 개발하기 위하여 문서인식의 전처리 단계인 문서의 레이아웃 인식을 한글과 한국의 문서환경에 알맞게 알고리즘을 개발 및 수정하여 적용하였다.

레이아웃 인식에서는 추출되어진 문자가 정확하게 문자인식 단계에 전달되어야 하므로 레이아웃 인식이 실패한 경우 문자인식까지는 이루어질 수 없게 된다. 경우에 따라서는 사용자가 문자영역만 지정하여 수행할 수도 있으나, 본 연구에서는 문서전체를 대상으로 실험한다. 본 논문에서 사용 및 개발한 알고리즘을 간단히 살펴보면, 레이블링 알고리즘은 한글의 특성에 2×3 마스크를 한글의 특성에 맞게 적용하였으며, 레이블링 이후의 처리에서는 1차 레이블링에서 추출된 블럭에 의하여 모든 레이아웃의 정보를 추출하였다. 이후의 처리는 문서의 구조분석에 의한 분석된 데이타에 의하여 처리를 하였고, 블럭화는 그동안 화상공학 연구실에서 연구한 연구결과[13][14]를 이용하여 처리하였다.

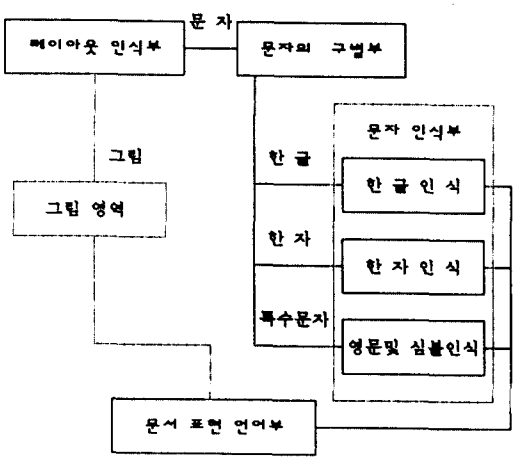

그림 1. 문서 인식의 개략도
Fig. 1. Block diagram of document recognition

Ⅱ. 문서의 구조 분석

본 장에서는 문서의 레이아웃인식을 위한 정보를

추출하기 위하여, 문서의 구조적인 분석을 ISO(International Organization for Standardization)에서 정의한 규격과 본 논문에서 사용한 학회의 논문지를 대상으로 구조를 비교 분석한다.

2.1 ISO에 의한 문서 구조

본 절에서는 본 논문의 대상으로 쓰인 이미지 정보가 속하는 문서에 대하여 조사한다. 국제 표준에 의하면 문서는 논리적 구조(Logical Structure)와 배치구조(Layout Structure)로 구성된다. 논리적인 구조는 문서의 전체적인 구성요소 즉, 장, 제목, 개요, 항, 내용등을 나타내고, 배치구조는 페이지(page), 프레임(frame), 블럭(block)등을 나타낸다. 그림 2와 그림 3에 문서의 논리 구조와 배치구조를 나타낸다.

이상에서 정의한 문서는 ODA(Open Document Architecture)에서 정의한 문서 형식이며, 이러한 문서의 구조는 레이아웃 정보의 추출과 인식 결과에 이용된다. 또한 논리적인 측면에서 인식된 결과는 사용자가 원하는 문서의 교환 형식으로 표현되어질 수 있다.

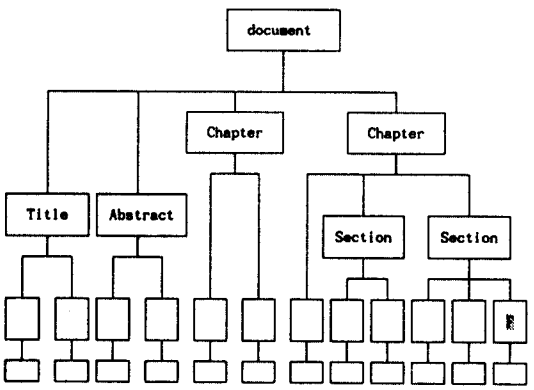

그림 2. 문서의 논리 구조
Fig. 2. Logical structure of document

2.2 실험 데이타

본 논문에 사용한 데이타는 어느정도 일정한 문서의 규격을 가진 전문 논문지(통신학회, 전자공학회, 정보과학회)를 대상으로 실험하였고 그림 4에는 입력된 데이타의 예를 보여주고 있다.

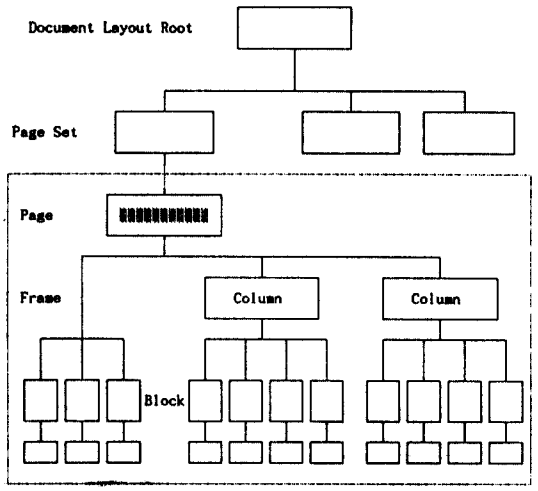

그림 3. 문서의 배치구조
Fig. 3. Layout structure of document

그림 4에서 보는 바와 같이 문서는 두개의 컬럼으로 구성되어 있으며 헤더와 풋터정보를 포함하고 있다. 헤더는 페이지 번호화 학회지명으로 이루어져 있고, 본문 안의 내용은 문자와 그림영역이 혼합되어 존재하고 있다. 본문은 문자를 이루는 부분을 제외하고는 모두 이미지로 간주하였다. 문서에서 문자열을 이루는 부분은 다음과 같은 성질을 가지고 있다.

1) 문자영역은 문자열로 이루어져 있고, 문자열 간격은 대체로 일정하다.
2) 문자열은 문자열 사이의 간격에 의해서 구분된다.
3) 단어는 문자로 구성되고 거의 일정한 간격으로 나란하게 있다.
4) 문자와 문자의 간격은 단어와 단어에 비하여 작다.

또한 문자영역과 그림영역은 다음과 같은 특징이 있다.

1) 문자블럭은 그림영역의 블럭보다 현저히 작다.
2) 그림영역의 가로, 세로 비율은 일정하지 않다.
3) 그림영역의 아래에는 그림을 설명하는 문자열이 있다.

이와 같은 성질을 이용하여 본 논문에서는 문자와 그림영역을 분리 및 추출하고, 레이아웃 인식을 위한 정보를 추출한다.

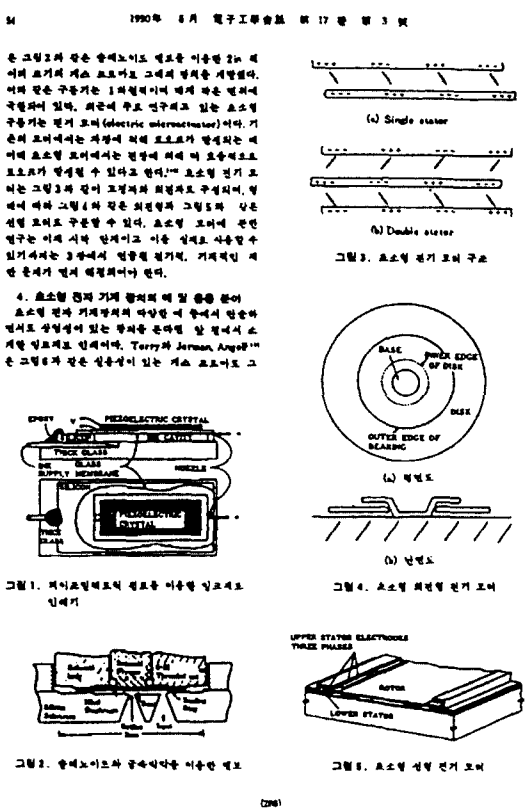

그림 4. 입력 데이타
Fig. 4. Input data

Ⅲ. 문서의 세그멘테이션(Segmentation)

일반적으로 문서를 이루는 구성요소는 크게 문자영역과 그림영역으로 분류할 수 있다. 이 중 문자영역은 각 문자들의 계층적인 구조(Hierachical structure)에 의해서 문자, 문자열, 문단, 문자블럭등으로 분류하고, 그림(Graphic)영역은 라스터(Raster) 그래픽과 지오메트릭(Geometric)그래픽으로 분류한다.

본 논문에서는 문서를 이루는 구성요소 중에서 문자영역 부분과 그림영역 부분을 분리하여 추출하고, 그 중 문자영역에 대해서는 각 문자들의 배열 상태를 조사하여 문서에서 문자영역을 계층적으로 추출하였다. A4 대상의 문서를 300 DPI의 해상도로 입력받은 데이타의 양은 대략 800만 화소의 점들로 구성되고, 약 1 Mbyte의 메모리 용량을 차지한다. 본 논문에서는 이러한 데이타를 처리하는데 있어서 전체 이진 영상에 대해서 1차 레이블링으로 한번 조사한 후, 이 레이블링 처리에서 블럭표현의 정보를 얻어 레이블링 처리 후의 모든 처리과정에서 블럭표현의 정보만으로 처리함으로써, 이진 영상을 계속적으로 조사해야 하는 번거로움을 없앴다.

3.1 레이블링(Labeling)

정보가 있는 부분은 흑, 정보가 없는 부분은 백으로 표현되는 전체 이진 영상의 문서 영역에서 정보가 있는 부분을 추출하기 위하여 전체 문서 영상에 대해 레이블링 처리를 한다. 레이블링이란, 이미지내에 연결된 각각의 영역에 고유의 레이블을 부여하는 처리이다. 레이블링 처리가 끝난 후, 각 레이블된 데이타에서의 위치정보를 추출하기 위하여 X축 시작점과 길이, Y축 시작점과 높이로 표현되는 블럭의 좌표를 추출하였다.

3.1.1 레이블링 알고리즘

일반적으로 영상에서의 레이블링 처리는 1차 레이블링 처리와, 2차 레이블링 처리를하여 연결된 영역을 추출하나, 본 논문에서는 1차 레이블링 처리후에 레이블된 데이타의 블럭 정보를 추출하여, 레이블링 처리후의 모든 처리과정에 1차 레이블링에서 추출된 블럭정보를 이용하여 처리함으로써 2차 레이블링 처리까지 해야하는 번거로움을 없앴다.

1차 레이블링 처리에서는 모든 이미지 영역에 각각 고유의 레이블을 부여하기 위하여 그림 5와 같은 2×3 마스크(mask)를 이용하여 처리하였으며 처리과정은 다음과 같다.

알고리즘 : 레이블링(labling)

단계 1 : 전체 이진 영상에서 흑화소의 영역을 찾기 위하여 이미지를 좌→우, 위→아래로 조사한다.

단계 2 : 현재 관찰된 화소가 백화소이면 레이블 0을 부여하고, 흑화소이면 단계 3으로 간다.

단계 3 : 그림 5의 마스크(mask)를 사용하여 주위 4방향의 레이블된 상태를 조사한다. 모두 레이블 0이면 새로운 레이블을 부여하고, 그렇지 않으면 단계 4로간다.

단계 4 : 그림 5의 마스크 중 우선 순위가 높은 레이블

을 현재의 주목화소에 부여한다.

단계 5 : 단계 1에서 단계 4를 2진 영상 전체에 대하여 수행한다.

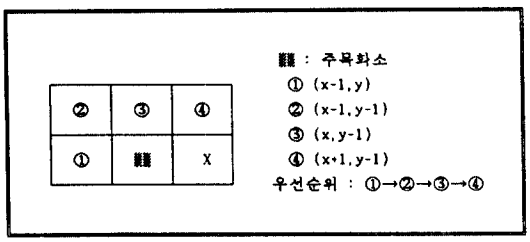

그림 5. 레이블링 마스크
Fig. 5. Labeling mask

3.1.2 영역의 블럭화

1차 레이블링 처리 후의 각각의 레이블들은 정보의 한 형태(문자, 그림, 분리선등)를 포함한다. 본 논문에서는 전체 이진영상에서 정보가 있는 부분에 대한 블럭표현을 얻기 위하여, 레이블링 처리와 동시에 블럭정보를 추출한다. 이 블럭 표현은 각각의 레이블된 집합을 포함하는 가장 작은 크기의 직사각형을 나타낸다. 전체 이진 영상을 조사하는 과정에서 레이블이 처음 발견된 위치를 Xc, Yc라 하면, 이 레이블 블럭 표현은 Xstart = Xc, Ystart = Yc, Xend = Xc + △X, Yend = Yc + △Y(△X : 블럭의 길이, △Y : 블럭의 높이)로 표현된다.

계속적으로 전체 이진 영상을 좌→우, 위→아래로 조사하면서 그림 5의 마스크로 4방향의 레이블을 조사하면서 같은 연결성의 레이블의 블럭표현과 현재 관찰중인 점과의 비교로 새로운 블럭표현을 한다. 이는 다음과 같다.

$$Xstart = \min(Xmin, Xc)$$
$$Ystart = \min(Ymin, Yc)$$
$$\triangle X = \max(Xmax, Xc)$$
$$\triangle Y = \max(Ymax, Yc)$$

비교 대상 블럭 : (Xmin, Ymin, Xmax, Ymax)

이와같은 작업을 2진영상 전체에 대해서 수행하여 이미지에 대한 블럭 표현을 얻는다. 그림 7은 전체 문서에 대하여 블럭정보를 추출한 결과이다.

3.2 문서의 구조정보 추출

문서를 구성하는 구조적인 정보에서 문서는 몇개의 행을 구성할 수 있고, 각 행별로 문자열이 구성된다.

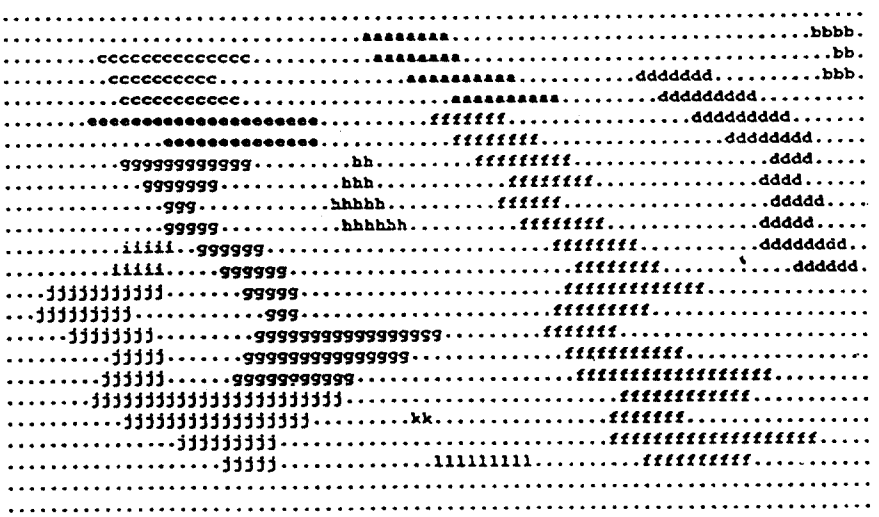

그림 6. 레이블링 처리
Fig. 6. Labeling process

본 논문에서는 전체 입력문서에서 블럭단위로 행을 추출하고 각 행별로 문자영역과 그림영역을 추출하기 위하여 위의 레이블링 처리 후 추출한 블럭정보를 이용, 열을 구분하여 문서를 몇개의 열과 행으로 구분되는 부분영역으로 나누어 처리하였다. 이에 따라 본 논문에서는, 먼저 전체 문서에 대한 열을 추출한 다음 각 열별로 문서에서의 행을 추출하였다. 그리고 이 행별로 추출한 영역에서 다시 문서의 열을 추출하여 문자 영역과 그림영역을 분리하고난 후에, 각 단원별로 블럭표현 데이타를 처리함으로써 후에 문서의 레이아웃의 구조를 추출하는데 이용하였다.

3.2.1 문서의 열(row)정보 추출

본 항에서는 블럭화된 데이타에서 문서의 열정보를 추출하기 위한 알고리즘에 관하여 논한다.

문서의 레이아웃을 구성하는 요소중에 헤더(header), 풋터(footer), 본문, 문자영역, 그림영역, 페이지 번호 등을 분리하기 위하여, 본 논문에서는 앞 절의 레이블링 처리 후에 추출한 문서의 블럭표현 정보를 이용하여 전체문서에 대해서 문서의 열을 추출하였다. 문서의 열정보 추출은 먼저 전체 블럭정보에서 각 블럭의 위치값과 블럭의 높이(△y)를 조사하여 가로축 라인으로 빈 공간이 있는 위치를 추출하였다.

본 논문에서는 이와같이 추출한 문서 전체의 열정보 중에서 문서의 구조적인 열정보를 추출한다. 문서의 구조적인 열 정보란 전체 문서에서 그림이나, 분리선, 또는 문자 영역에 의해 구분되는 열의 정보를 가리킨다. 본 논문에서는 위에서 얻은 각 열 정보 데이타에 대하여, 각 열 사이의 간격을 구한 다음, 아래의 공식으로 이들 높이에 대한 표준편차를 구하여, 표준편차 이상의 거리를 가지는 열을 문서에서의 구조적인 열 정보로 추출하였다.

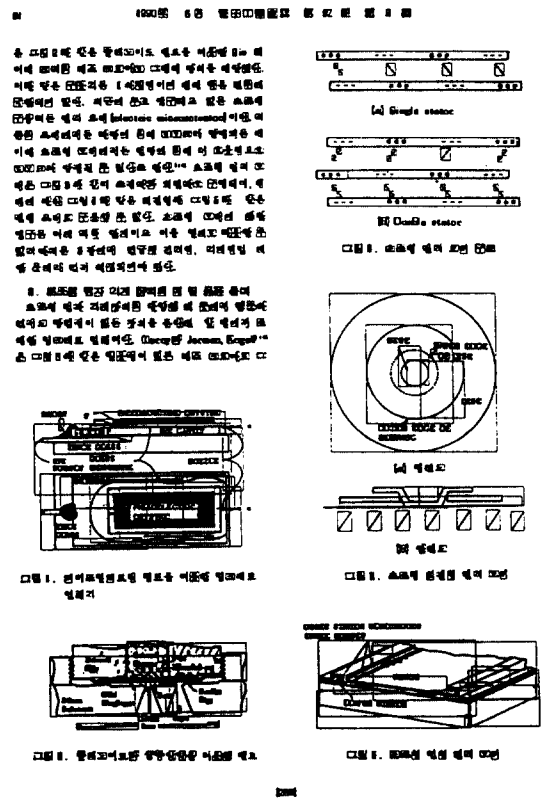

그림 7. 블럭화된 문서 데이타 예
Fig. 7. Blocked document data example

각 열 사이의 간격 : $Y1, Y2, Y3, \ldots, Yn$

각 열 사이 간격의 평균 :

$$Y = \frac{1}{n} \sum_{j=1}^{n} Yj$$

표준편차 :

$$\sigma(Y1, Y2, \ldots, Yn) = \frac{1}{n-1} \sum_{j=1}^{n} (Yj - Y)^2$$

그림 8에 위 식을 적용하여 추출한 문서에서의 구조적인 열 정보를 나타내었다.

3.2.2 문서의 행(column) 정보 추출

앞서의 문서에서의 구조적인 열정보를 이용하여, 각 열로 구분되는 영역에 대해서 행(column) 정보를 추출하였다.

행 정보의 추출은 앞에서 열 정보를 추출한 것과 같은 방식을 사용하였으며, 각 구조적인 열 정보로 표현되는 영역에 포함되는 블럭정보를 이용하여 세로로 빈 공간의 행에 대한 표준편차를 구하여 표준편차 이상의 빈 공간의 여백을 가지는 행을 부분 영역에 대한 행으로 추출하였다.

보통의 문서는 몇개의 행으로 이루어질 수 있다. 그리고 이 컬럼들 사이의 간격은 문자와 문자 사이의 간격과 단어와 단어 사이의 간격보다 훨씬 크며, 실험적인 결과로서 100 화소 이상으로 나타났다. 본 논문에서는 이들 추출된 행 중 간격이 100 화소 이상인 것을 추출하였다. 행 정보 추출의 결과를 그림 9에 나타내고, 부분 영역 추출에 대한 흐름도를 그림 10에 나타내었다.

3.3 문자영역 추출

본 논문에서는, 행과 열이 혼합된 임의의 문서에서 앞에서의 전체 문서에 대한 열 정보와 각 열에 대한 행 정보를 이용하여 문서를 부분적으로 나누어 처리하였다. 문서는 그림 9와 같이 부분적으로 나누어진 영역의 순서대로 처리하여 문자 영역과 그림 영역의 분리, 문자영역 및 문자열의 추출, 문자 블럭들의 정렬 및 혼합등을 수행하게 된다. 이렇게 문서를 부분

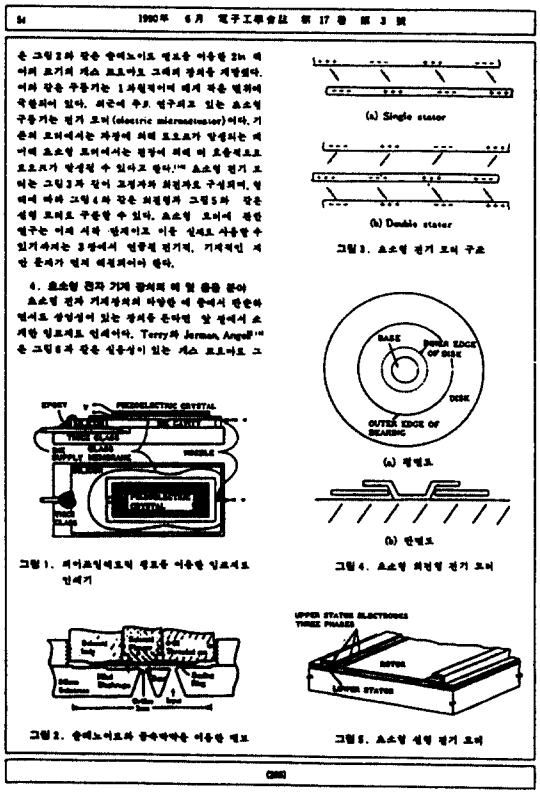

그림 8. 구조적인 열정보 추출
Fig. 8. Extraction of Structrual row information

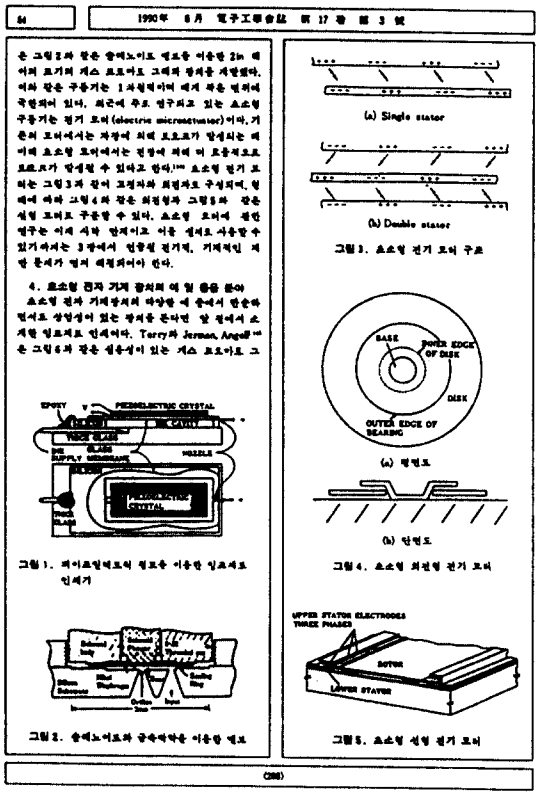

그림 9. 구조적인 행정보 추출
Fig. 9. Extraction of structural column information

적으로 나누어 처리함으로써 사람이 읽는 순서에 의하여 문자 블럭들을 추출하였다. 이러한 부분 영역들의 정보는 뒤에 레이아웃의 정보 추출에 대한 특징 요소로 이용되고, 부분영역들의 위치와 부분 영역내의 각 문자영역과 문자열 그리고 그림영역을 추출하였다.

3.3.1 블럭의 정렬

1차 레이블링 처리후에 추출한 블럭정보는 추출하는 데에서 문서의 행과 열을 고려하지 않고, 문서를 조사하는 순서대로 블럭의 순서가 정해져 있다. 이러한 블럭들을 문서의 구조에 맞게 부분영역별로 정렬(sorting) 하였다. 부분영역별 블럭들의 정렬에는 각 블럭들의 Y축 상위 정보를 사용, 퀵정렬(quick sorting)을 이용하여 정렬하였다. 이들 정렬된 블럭들의 정보는 이 후 문자영역과 그림영역의 분리 및 문자열의 추출, 그리고 개별문자의 추출에 사용된다.

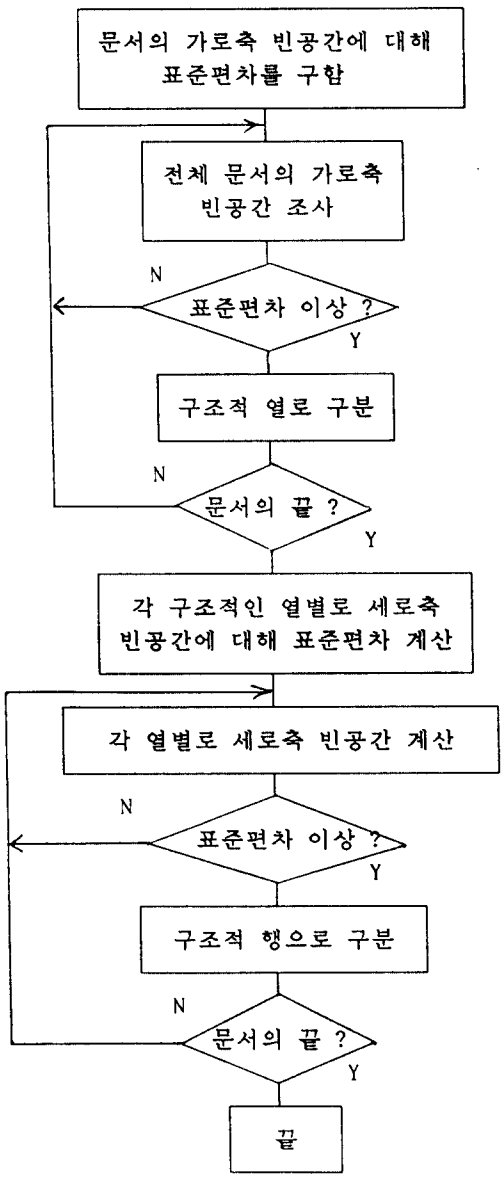

그림 10. 부분영역 추출에 대한 흐름도
Fig. 10. Flowchart of sub-region extract

3.3.2 문자열 추출

부분영역에서 문자영역은 보통 수개의 문자열과 그림영역으로 이루어져 있다. 본 논문에서는 이들 문자열을 추출하기 위하여 부분영역별 가로축 빈공간을 추출하여 열정보로 이용하였다. 이들 각 부분영역별 열 정보는 문자영역과 그림영역을 분리하여 주고, 문자 영역에 대하여서는 각 문자열을 분리하여준다. 먼저 각 부분영역에 포함되는 블럭들을 위→아래로 정렬하였다. 이들 정렬된 데이타에서 각 블럭정보의 위치 정보와 아래측 정보를 조사하여 각 문자열들 중 빈 공간을 추출하였으며, 이 빈 공간을 문자열과 문자열 사이의 간격으로 추출하였다. 각 블럭들을 조사하면서 x축 최대 최소, y축 최대 최소를 비교하면서, 열에 대한 블럭 표현을 얻었다.

3.3.3 문자 영역과 그림영역의 분리

일반적으로 문서를 이루는 구성요소는 텍스트, 그래픽, 라스터그래픽으로 이루어진다. 이중 텍스트 영역에 대하여 문자의 분리및 합성을 수행하기 위하여, 앞서 추출한 블럭 정보 및 구조적인 정보를 이용하여 문자영역과 그림영역을 분리하였다.

보통 문자들은 그림보다 작은 영역을 차지하고, 몇개의 문자열들이 모여서 문자영역을 구성한다. 본 논문에서는 이들 문자영역의 구조적인 배열 및 위치적인 특성을 이용하여 앞서의 각 부분 영역별로 문자영역을 추출하였다. 특징으로는 부분영역에서의 열정보를 이용하여 이들 특징된 블럭의 원래 이미지에서 다음의 정보를 추출한다.

- 블럭의 높이
- 블럭의 크기
- 블럭내의 흑화소의 갯수
- 블럭내의 흑백의 교차수
- 블럭의 크기대 흑화소의 비율
- 블럭의 크기대 교차수의 비율

본 논문에서는 각 문자열 별로 추출한 위의 정보를 이용하여 문자영역과 그림영역을 분리하는데 이용하였다. 보통은 문자영역과 그림영역의 사이는 문자열과 문자열 사이의 간격보다 작다. 이들 그래픽 블럭과 문자열의 특징은 실험적으로 다음과 같다.

- 그래픽 블럭
  - 블럭의 높이 > 200
  - 블럭의 크기대 흑화소의 비율 < 0.1
  - 블럭의 크기대 교차수의 비율 < 0.02
- 문자열 블럭
  - 블럭의 높이 < 80
  - 블럭의 크기대 흑화소의 비율 > 0.1
  - 블럭의 크기대 교차수의 비율 > 0.02

이들 특징을 이용하여 먼저 블럭의 높이가 200 이상인 것은 그림영역으로 분류하였다. 그리고 블럭내의 흑백의 교차수, 흑화소의 수 등을 사용하여 전체 문서영역에서 문자열 블럭과 그림영역을 나누었다.

표 1은 이들 블럭들의 특징을 나타내며 그림 11은 이들 특징들을 이용하여 각 부분 영역별로 문자영역과 그림영역을 분리한 것이다.

3.4 문자 추출

앞서의 레이블링 처리후에 획득한 블럭정보들은 몇개가 모여서 하나의 문자정보를 이루게 되는데, 이 절에서는 먼저 부분영역에서 문자영역에 대하여 추출된 문자열에 대해서 정렬된 블럭 정보를 가지고 개개의 문자를 추출한다.

3.4.1 문자 추출

각 개개의 문자를 추출하기 위하여 먼저 문자열의 블럭표현에 대하여 문자와 문자사이의 간격을 추출하였다. 문자와 문자 사이의 간격 추출은 레이블링 후의 블럭표현 정보를 이용하였으며, 이들 블럭표현에서 각 블럭의 x축 시작점과 △x의 정보를 사용하여 한 문자열에서 세로축으로 빈 공간을 추출하였다. 문자영역에서 개개의 문자를 추출한 결과를 그림 12에 보였다.

3.4.2 블럭의 특징

레이블링 처리 후에 결과로서 나타난 블럭은 대개의 경우 하나의 문자가 여러개의 블럭으로 나누어져 있으며 다음과 같은 특징이 있다.

1) 한 문자열에서 세로축으로 분리된 문자는 심한 경우에 5-6개의 블럭으로 분리되어 있다. 본 논문에서는 각 문자열별로 세로축으로 분리되어 있으면서 겹쳐진 블럭에 관해서는 한 문자요소의 블럭으로 간주 하였다.(그림 13-a)
2) 실험적인 결과에 따라 한글의 경우 하나의 문자로 추출될 수 있는 경우에 블럭의 가로-세로의 비는 0.75 이상이다. 따라서 가로-세로 비가 0.75 이하의 블럭은 두개의 문자가 하나로 블럭화된 것으로 가정한다.(그림 13-b)
3) 한글의 경우 그림13과 같은 관계가 존재한다.(그림 13-c)

1)과 3)의 경우에는 블럭의 합성대상이되고, 2의 경우에는 분리작업의 대상이 된다.

표 1. 블럭의 특징
Table 1. Characteristic of Block

num	heignt	size	pixels	count	ratio	count/size	status
1	40	55309	6005	1159	0.109	0.021	T
2	37	31958	5231	1020	0.164	0.032	T
3	38	32838	5333	1100	0.162	0.033	T
4	37	31844	4893	1120	0.154	0.035	T
5	37	31958	4457	977	0.139	0.031	T
6	39	33720	4972	1144	0.147	0.034	T
7	38	33799	5071	1109	0.155	0.034	T
8	38	32877	5131	1098	0.156	0.033	T
9	38	32994	4622	973	0.140	0.029	T
10	39	33600	5123	1088	0.152	0.032	T
11	39	33800	5139	1045	0.152	0.031	T
12	39	33880	5137	976	0.152	0.029	T
13	38	32994	5131	1038	0.156	0.031	T
14	38	32799	5148	1174	0.157	0.036	T
15	37	20330	3729	797	0.183	0.039	T
16	36	27713	8538	991	0.308	0.036	T
17	35	29268	6168	1173	0.211	0.040	T
18	35	30384	5858	1069	0.193	0.035	T
19	37	32262	6294	1249	0.195	0.039	T
20	37	32110	5379	936	0.168	0.029	T
21	482	355488	96735	6831	0.272	0.019	P
22	37	30210	5688	976	0.188	0.032	T
23	34	3360	798	190	0.237	0.057	T
24	304	203740	42960	4652	0.211	0.024	P
25	36	28083	6081	939	0.217	0.033	T
26	31	20768	2784	228	0.134	0.011	P
27	44	25785	878	169	0.034	0.007	P
28	34	22645	4356	284	0.192	0.013	P
29	37	10450	1979	413	0.189	0.040	T
30	32	21384	3008	216	0.141	0.010	P
31	48	28812	910	168	0.032	0.006	P
32	33	22134	5249	233	0.237	0.011	P
33	77	50466	4217	374	0.084	0.007	P
34	35	10116	1996	456	0.197	0.045	T
35	36	18944	3516	530	0.186	0.028	T
36	547	304688	21108	3605	0.069	0.012	P
37	37	6422	1309	252	0.204	0.039	T
38	74	50475	9719	403	0.193	0.008	P
39	73	46324	2485	476	0.054	0.010	P
40	35	6048	1245	250	0.206	0.041	T
41	37	20596	4235	642	0.206	0.031	T
42	22	9039	2817	575	0.312	0.064	T
43	21	4466	1491	321	0.334	0.072	T
44	302	226341	34121	2950	0.151	0.013	P
45	36	18759	4151	577	0.221	0.031	T
46	32	2211	821	176	0.371	0.080	T

※T : 문자영역(Text)　※P : 그림영역(picture)

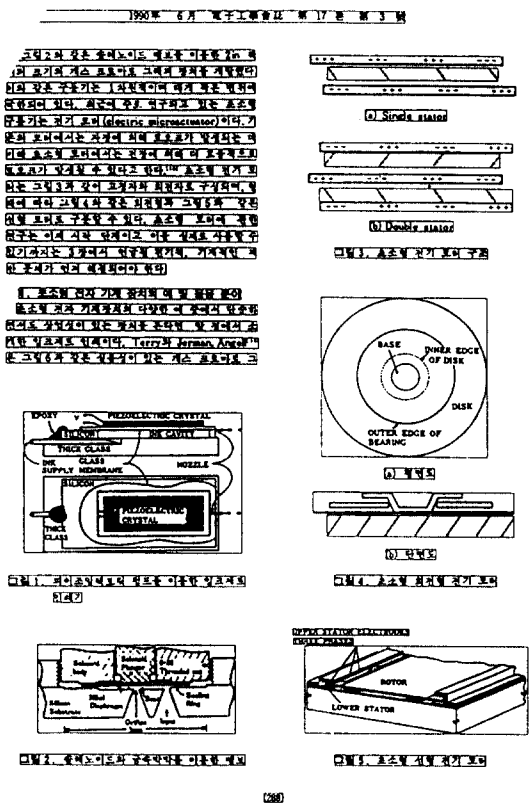

그림 11. 문자영역과 그림영역의 분리
Fig. 11. Seperation of text and image region

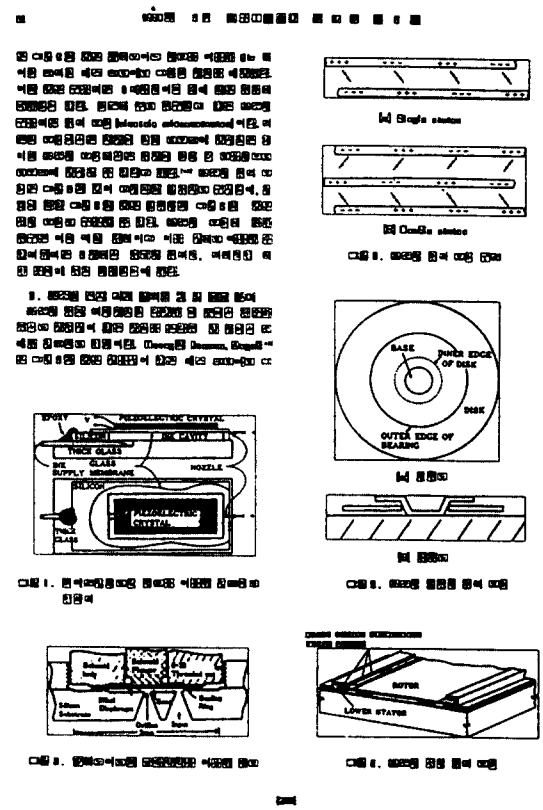

그림 12. 문자 추출
Fig. 12. Extract of Character

3.4.3 블럭의 분리 및 합성

블럭의 분리 및 합성에 대한 조건을 한글의 경우 그림 13과 같은 블럭 관계가 존재할 때 이들 블럭들의 분리 및 합성을 행한다.

블럭의 분리 및 합성에 대한 알고리즘은 오인권 「13」의 블럭의 분리 및 합성 알고리즘을 수정하여 사용하였다. 본 알고리즘은 한글의 자소 특성을 고려하여 광운대학교 화상공학 연구실에서 개발한 것으로서 한글의 블럭 비율을 이용하여 블럭의 분리 및 합성을 하는 알고리즘이다. 이러한 경우는 한 블럭에 두개 이상의 블럭이 겹쳐지는 경우와 한개의 문자가 두개의 블럭으로 분리되는 경우가 존재하게 된다. 이러한 경우를 적용하여 처리하면 하나의 문자는 하나의 블럭으로 분리 및 합성하여지게 된다. 이와 같이 처리한 결과를 그림 14에 나타내었다.

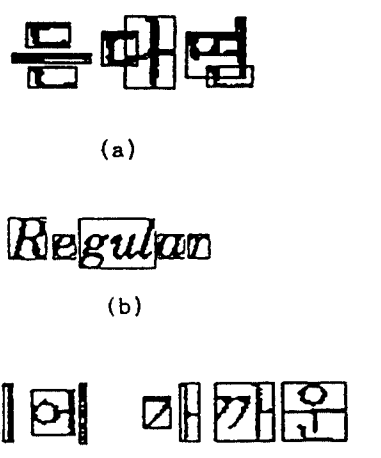

그림 13. 블럭의 특징
Fig. 13. Characteristic of blocks

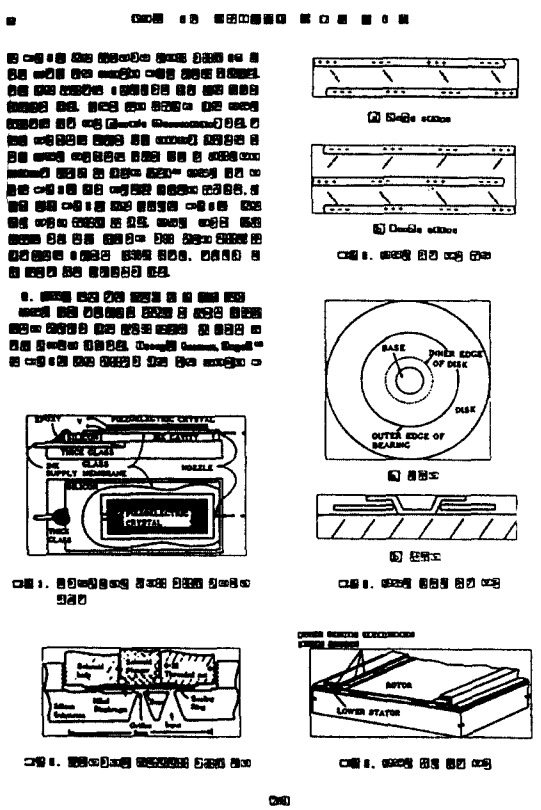

그림 14. 분리 및 합성된 데이타
Fig. 14. Examples of merged and seperated data

Ⅳ. 레이아웃 정보의 추출

본 장에서는 추출되어진 블럭을 대상으로 문서를 이루는 구조적 정보를 이용하여 레이아웃 인식을 위한 정보를 추출한다. 본 논문에서는 문서의 레이아웃 정보의 요수를 헤더, 풋터, 문자 블럭, 그림 블럭, 캡션(caption), 페이지 번호로 크게 분류하였으며, 앞장에서 추출된 특징들을 이용하여 처리하였다.

4.1 문서의 레이아웃

어떤 문서든지 그 문서가 가지는 고유의 형태가 존재하게 된다. 본 논문에서 사용하는 문서 데이타는 전문 학술 논문지로서 일정한 형태를 가지고 있다. 앞서의 처리에서는 이러한 레이아웃 정보를 알기위하여 수행한 블럭화 처리이고, 본 장에서는 이렇게 분리된 블럭 정보를 합성해 가면서 문서가 가지는 레이아웃 인식을 위한 정보를 추출한다. ISO에서는 모든 레이아웃 정보는 프레임(frame)이라는 정보와 한 프레임 속에 몇개의 블럭이 속하게 된다. 본 논문에서는 문서에서 문자와 그림을 분리하여 각각의 블럭으로 나누는 과정을 수행하였다.

블럭의 분류는 그래픽 블럭(Graphic block)과 텍스트 블럭(Text block)으로 분리되어지고, 하나의 텍스트 블럭은 몇개의 텍스트 소블럭(sub-block)으로 나누어진다. 논리적인 문서의 블럭과 프레임 개념을 그림 15에 도시하였다.

4.2 레이아웃 정보의 추출

앞 장에서의 문서의 부분영역 추출과, 각 영역별로 추출한 특징을 이용하여 다음의 문서의 레이아웃적인 요소의 특징 정보를 추출한다. 각 요소들의 특징은 다음과 같다.

1)hearder :1)문서의 제일 처음 라인으로 나타난다.
:2)문자 블럭으로 이루어진다.
:3)헤더와 본문사이엔 커다란 공백 라인이 존재한다.

2)footer :1)문서의 제일 마지막 라인에 나타난다.
:2)문자 블럭으로 이루어져 있다.
:3)풋터와 본문사이엔 커다란 공백 라인이 존재한다.

3)페이지 번호 :1)헤더와 풋터사이에서 나타남
:2)짧은 문자 블럭으로 이루어져있다.
:3)문자열중 가운데 아니면 왼쪽 끝이나 오른쪽 끝에 존재한다.
:4)블럭의 크기가 매우 작다.

4)문자 블럭 :1)헤더와 풋터사이에 존재하는 몇개의 문자 블럭으로 이루어진다.
:2)각 문자열의 간격이 일정하다.

5)그래픽 블럭 :1)헤더와 풋터사이에 존재하는 몇개의 그래픽 블럭으로 이루어진다.
:2)크기가 매우 크다.

6)caption :1)그래픽 블럭의 위나 아래에 짧은 문자열로 존재.
:2)문자열의 갯수는 2개 이하이다.

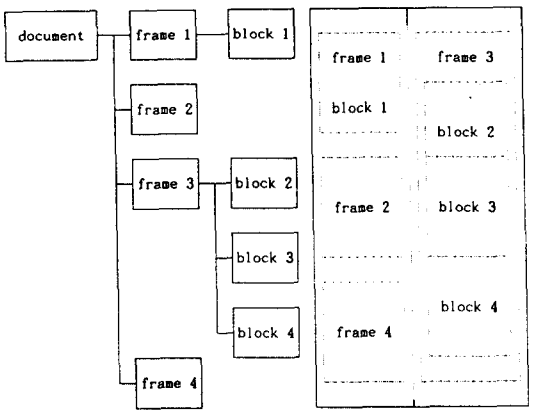

그림 15. 문서의 블럭 및 프레임 개념
Fig. 15. Block and frame conception of document

이들의 특징들을 이용하여 본 논문에서는 문서의 레이아웃적인 정보를 추출한다. 레이아웃 정보의 추출에서 헤더와 풋터의 인식은 3장에서의 문서를 부분 영역별로 나누면서 각 열을 추출할 때 제일 처음 열과 제일 마지막 열에 표준 편차 이상의 문자열과 구분이 되어진다.

그리고 문자 블럭과 그래픽 블럭의 구분은 결과적으로 각 부분영역에서 문자열에 대한 표준편차이상의 간격을 추출함으로써 구분되어 이를 점으로 본 논문에서 제안한 문서의 부분 영역별 추출 방법이 문서의 레이아웃적인 요소를 추출하는데 효과적임을 알 수 있었다. 그림 16와 표 2에 위의 개념과 특성을 이용하여 블럭으로 분류 및 정보를 추출한 문서의 예를 보여준다.

표 2. 레이아웃 정보 추출의 결과
Table 2. The results of layout information extraction

block number	status	group	Layout
1	T	1	page neumber
2	T	1	header
3	T	2	text
.	.	.	.
.	.	.	.
17	T	2	text
18	T	2	text
19	T	2	text
20	T	2	text
21	P	3	picture
22	T	4	caption
23	T	4	caption
24	P	5	picture
25	T	6	caption
26	P	7	picture
27	P	7	picture
28	P	7	picture
29	T	8	caption
30	P	9	picture
31	P	9	picture
32	P	9	picture
33	P	9	pciture
34	T	10	caption
35	T	10	text
36	P	11	picutre
37	T	12	caption
38	P	13	picture
39	P	13	picture
40	T	14	caption
41	T	14	text
42	T	14	caption
43	T	14	caption
44	P	15	picture
45	T	16	caption
46	T	17	page number

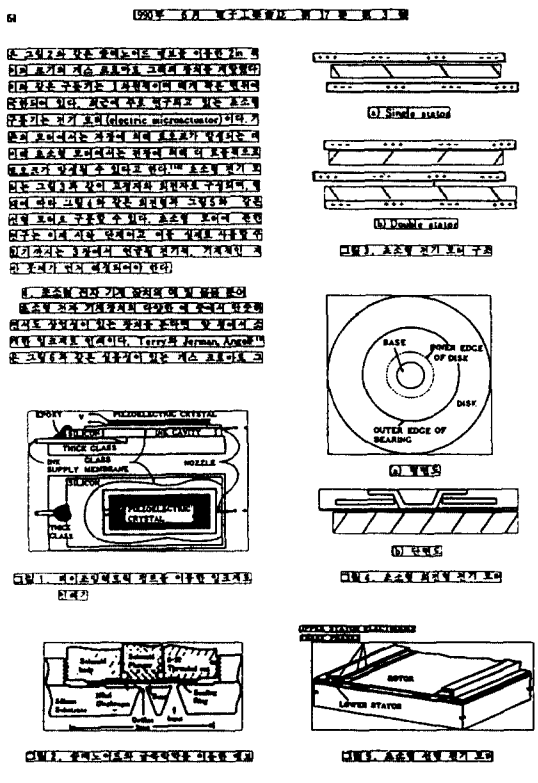

그림 16. 레이아웃 정보의 추출
Fig. 16. The extraction of layout information

V. 실험 및 고찰

5.1 실험 시스템

실험에 사용된 입력문서에는 5개의 국내학회지를 대상으로 하였으며 IBM-PC/AT에서 MS-WINDOW 3.0의 SDK(SoftWare Development Kit)를 사용하여 구현되었다.

문서 영상은 Hewlett Packard의 이미지 스캐너(Image scanner)를 이용하여 인치당 300 화소의 해상도로 입력받았다. 입력한 문서의 최대 크기는 A4 용지의 크기이며 결과로써 2432×3460 크기의 이진 영상과 레이아웃 정보 추출표를 출력한다. 그림 17에 본 연구에서 사용한 실험 시스템의 구성도를 보였다.

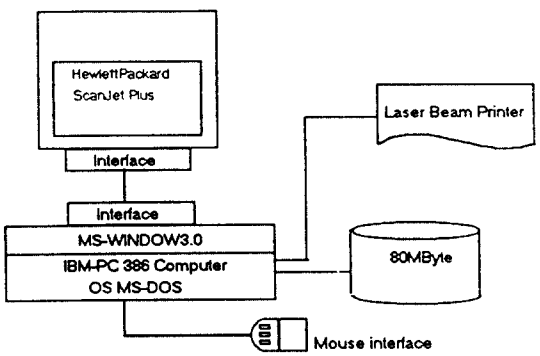

그림 17. 실험 시스템
Fig. 17. Experimental system

5.2 실험 결과

본 논문은 문서에서 영역 분리와 레이아웃을 위한 정보의 추출을 수행하였다. 문서의 영역 분리는 문서에서 문자영역과 그림영역을 분류하는 것 이외에도 문서를 부분영역으로 나누어 처리함으로써, 레이아웃 요소의 인식 수행시에 필요한 정보 추출에 유효하였다. 그리고 문서영역을 1차 레이블링 처리 후에 블럭 정보를 추출함으로써 2차 레이블링 처리와 이후에 계속 문서를 처리하기 위해서 조사하는데 소요되는 시간을 단축하였다.

본 논문은 일정한 형식을 갖춘 한국어 학회지 5개를 대상으로 30개의 데이타를 실험하였으며 이에 대해서 부분 영역 추출 및 문자영역과 그림영역의 추출에 대해서는 98.5%의 추출률을 얻었다.

개별 문자의 추출은 문자를 가로·세로비로 추출함으로써 한글에 대해서는 100%의 추출 결과를 얻었으나, 영문 및 특수 문자의 추출에 대해서는 많은 오차가 있었다. 이에 대해서는 문자의 분류단계를 거쳐 영문 및 특수 문자에 대해서 따로 분리하여야 할 것이다.

레이아웃을 위한 정보의 추출은 레이아웃의 복잡성과 프레임의 애매성때문에 커디란 블럭단위의 레이아웃 정보를 추출 하였으며, 일정한 형식을 갖춘 학회논문지에 대해서 98%의 정확성을 얻었다.

5.3 고찰

본 논문은 문서의 영역 분리와 레이아웃 정보의 추출은 문서 처리 및 인식의 전단계로서 수행하는 일이다.

본 논문에서 제안한 부분 영역별로 문서를 나누어 사람이 읽는 순서대로 문서를 처리함으로써 하나의 문서를 분류하는데 보다 효과적으로 사용하였다.

전체 이진 영상에 대해서 1차 레이블링 후의 블럭 정보만을 가지고 처리함으로써, 2차 레이블링과 계속적으로 문서를 처리해야 하는 번거로움을 없앴으나, 부분영역별로 나누면서 각 부분 영역들에 해당하는 블럭 정보들을 모으고, 다시 순서에 맞게 정열하는 과정을 거쳐야한다.

본 논문에서는 문서를 영역 분리하는데 제한 사항을 두어 일정한 형식을 갖춘 학회를 대상으로 하였다. 문서를 부분 영역으로 나누는 과정에서 행과 열이 뚜렷히 구분되는 문서에 대해서는 완벽하게 추출되었으나, 행과 열이 삐뚤어 지거나 형식이 없는 복잡한 문서에 대해서는 많은 오차가 발생하였다. 앞으로는 좀 더 복잡하고 다양한 문서에 대해서도 확장되어져야 할 것이다.

문서를 부분적으로 나누어 처리하는 것은 나중에 다시 세분화 할때 매우 유용한 정보로 이용되었으며 커다란 레이아웃적인 요소의 추출로 유용한 정보로 사용되었다.

그림영역 및 문자영역 추출 부분은 K.Y.WANG [9]에서 사용한 특징들을 이용하여 추출하였으며, 영역의 추출면에서는 모두 완전하게 추출되었으나, 그림 22과 같이 사각형내의 문서의 제목, 반전된 색깔로 표시된 문자(배경 : 흑, 문자 : 백)에 대해서는 그림으로 판정되었다. 사각형 내의 문자들이 그림으로

판정된 것은 사각형내의 영역이 그림으로 판정되면서 안에 속해있는 문자영역을 포함하기 때문이다. 이와같이 블럭들이 그림영역의 특징들을 가지는 영역에 대해서는 그림영역내의 문자를 추출하는 단계에서 추출하여 이러한 문자들에 대해서도 모두 완전하게 문자로 추출되어 인식되어 질 수 있도록 많은 연구가 이루어져야 할 것이다. 문제점은 문서를 부분 영역별로 나누는 과정에서 구조적인 행정보가 잘못 추출된 결과이다. 이렇게 잘못 추출된 결과로 인한 문서의 구조가 잘못 인식되고, 심한 경우엔(열과 행으로 구분되어지지 않는 문서) 문자열 추출에 실패하는 경우도 있다. 이미지의 아래에 캡션이 아닌 데이타에서는 그림과 표를 설명하는 캡션부분이 일반적인 문자 영역으로 추출되어진 것이다. 본장에서는 그림과 그림을 설명하는 캡션의 관계를 그림과 문자 사이의 간격으로 추출하였다. 이러한 이유로 인하여 그림과의 간격이 넓은 캡션에 관해서는 그림과 같이 문자열로 추출되어진다. 그림영역을 설명하는 캡션은 그림영역에 속해있는 문자열 이므로 이러한 문자들을 그림 영역과 함께 추출할 수 있는 연구가 또한 진행되어져야 할 것이다.

본 논문에서는 1차 레이블링 처리 후의 블럭정보를 이용하여 문서의 열과 행을 추출함으로써, 일반적으로 1차, 2차 레이블링 처리의 이용과 계속적인 문서영역의 조사에 따르는 시간을 단축하였다. 아래 표3에 본 연구의 단계별 처리시간을 나타내었다.

표 3. 단계별 처리시간
Table 3. Processing time for each step

블럭정보 추출	부분영역 추출	문자및 그림 영역 분리	개별문자 추출	레이아웃 정보추출	총처리 시간
8분	1분	10분	30초	30초	20분

위의 표에서 알 수 있는 것과 같이 문서를 조사하는데 걸리는 시간이 전체 처리시간의 80% 이상을 차지한다. Wang이 제안한 알고리즘[1][9]의 경우는 문서를 최소한 4번 이상 조사하는데 A4 크기의 문서를 한번 조사하는 경우 약 10분 정도가 소요된다. 본 논문에서는 문서의 조사를 처음에 블럭정보의 추출과, 문자 및 그림영역의 분리에서 특징을 추출할 때 두번 조사함으로서 조사의 빈도수를 줄였다. 그리고 문서의 처리를 부분영역 별로 나누어 처리함으로써 나중에 따로 리딩 오더(reading order)의 순서를 결정할 필요가 없으며, 이러한 부분영역은 레이아웃 정보의 추출 및 인식의 단계에서 유용한 정보로 이용될 것이다.

레이아웃의 인식을 위한 정보의 추출은 다음에 레이아웃 인식을 위해서 필요한 정보를 커다란 블럭별로 추출하였다. 문서의 헤더, 풋터, 페이지 번호, 문자블럭, 그림블럭들은 비교적 양호하게 추출되었으나, 그림을 설명하는 캡션 부분에 대해서는 다소의 에러가 있었다. 이러한 부분은 문자를 인식하면서 추출하여야 할 것이다. 복잡한 문서에서의 정보의 추출과 문서의 레이아웃 인식이 앞으로도 계속 연구되어져야 할 것이다. 그림 18에 본 연구의 전체 흐름도를 나타내었다.

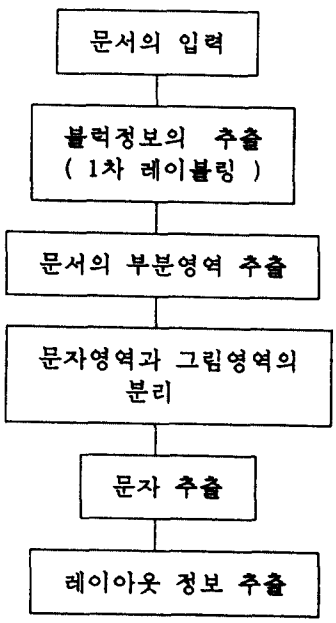

그림 18. 전체 블럭도
Fig. 18. Total Block Diagram

VI. 결 론

본 논문에서는 문서인식을 위한 전처리 연구로서, 문서 데이타 정보를 레이블링하여 생성된 블럭데이타만을 가지고 문서의 구조적인 정보와 레이아웃 인식을 위한 정보의 추출에 관한 연구를 하였다.

또한 한글에서 나타나는 블럭 분리화 합성을 처리

하여 문서의 레이아웃 요소인 그림과 문자 정보를 추출하였으며, 이러한 정보는 문서의 블럭정보를 추출하는데 좋은 방식으로 인식되었다. 기존의 문서인식 방법에 사용된 방법보다 속도면에서 많은 향상을 가져왔으며 추출 및 인식도 한국어 문서에 적절하게 이용될 수 있다는 점이 실험을 통하여 입증되었다. 문서의 레이아웃 인식을 위한 정보의 추출에서는 문서의 레이아웃을 인식하기 위하여 필요한 정보들을 추출하였으며, 이러한 정보를 이용하여 문서의 레이아웃 인식이 계속적으로 연구되어져야 한다.

앞으로 좀더 복잡한 문서에 대하여 영역의 분리 및 정보의 추출에 대한 연구가 이루어져야 할 것이며, 추출된 문자정보의 인식도 연구되어져야 할 것이다.

참 고 문 헌

1. Dacheng Wang and Sargur N. Shrihari, "Classification of Newspaper Image Block Using Texture Analysis," *Computer Vision, Graphics, and Image Processing 47*, 327-352, 1989.
2. Stephen W. Lam, Dacheng Wang and Sargur N. Srihari, "Reading Newspaper Text," *International Conference on IEEE*, 1990.
3. Baird H., "Bird Feature Identification for Hybrid Structural/Statistical Pattern Classification," *Computer Vision, Gaphic, and Image Processing, 42* : 318-333, 1988.
4. Duda R.O. and Hart P.E., "Pattern Classification and Scence Analysis," Wiley, 1973.
5. Fletcher L.A. and Kasturi R.A., "Robust Algorithm for Text String Separation from Mixed Text/Graphics Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6) : 910-916, 1988.
6. Pavlidis T.A. "Vectorizer and Feature Extractor for Document Recognition," *Computer Vision, Graphics, and Image processing*, 35 : 111-127, 1986.
7. Srihari R. and Rapaport W.J., "Extracting Visual Information From Text : Using Captions," *In. Proc. of the 11th Annual Conference of the Cognitive Science Society*, pp.364-371, 1989.
8. S. N. Srihari, "Document Image Understanding," *Proc. of the IEEE Computer Society Fall Joint Computer Conference*, pp.87-96, 1986.
9. K. Y. Wong, R. G. Casey and F. M. Wahl, "Document analysis system," *IBM J. Res. Develop.*, vol.6, pp.642-656, Nov. 1982.
10. G. Nagy, S. C. Seth, and S. D. Stoddard, "Document Analysis with an Expert System," *Proc. of ACM Conference on Document Processing Systems*, pp.169-176. 1988.
11. K. Inakagaki, T. Kato, T. Hiroshima and T. Sakai, "MACSYM : A Hierarachical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents," *Pattern Recognition, Vol. 17, No. 1*, 1984.
12. H. Kida, O. Iwaki and K. Kawada, "Document Recognition Ststem for Office Automation," *International Conf. on IEEE*, 1986.
13. 오인권, "영문이 혼합된 한글 문서에서 문자 및 특수문자 추출에 관한 연구," 광운대학교 대학원 석사학위 논문, 1988.
14. 남궁연, "한국어 문서로 부터 문자분리 및 도형 추출에 관한 연구," 광운대학교 산업정보대학원 석사학위 논문, 1986.

南宮在贊(Jae Chan Namkung) 正會員
1947년 6월 13일생
1970년 : 인하대학교 전기공학과 졸업(공학사)
1976년 8월 : 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 2월 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년 ~ 1984년 : 일본 Tohoku대학 객원교수
1979년 ~ 현재 : 광운대학교 전자계산기공학과 교수
※관심분야 : 패턴인식, 컴퓨터비전, 인공지능

趙 鎔 周(Yong Joo Cho) 正會員
1964년 3월 15일생
1986년 2월 : 원광대학교 전자계산공학과 졸업(공학사)
1988년 8월 : 광운대학교 전자계산기공학과 졸업(공학석사)
1991년 ~ 현재 : 광운대학교 대학원 전자계산기공학과 박사과정 재학중
※관심분야 : 패턴인식, 컴퓨터비전