논 문 논문번호 15-5-5

한국 신문 영상의 구조 분석을 통한 기사의 추출

(Extracting Articles from Structural Analysis of Korean Newspaper Image)

김 형 훈* 이 성 환** 김 진 형***
(Hyung Hoon Kim) (Seong Whan Lee) (Jin Hyung Kim)

요 약

본 논문에서는 한국 신문 영상으로부터 기사를 자동적으로 추출할 수 있는 한국 신문 처리 시스템을 기술한다. 이 시스템은 크게 3개의 구성 요소, 즉 신문을 표제, 기사 본문, 그림과 같은 영역으로 분할하는 영역 명시 단계와 명시된 영역으로부터 기사를 구성하는 기사 추출 단계 그리고 마지막으로 텍스트 열(text line)로부터 문자 단위 영상을 추출하는 문자 영상 추출 단계로 구성된다.

신문 기사의 각 영역을 명시하기 위해서 영역의 통계적 특성과 분리선을 이용하였다 분리선의 기능을 이용함으로서 영역의 통계적 특성만을 사용하는 것보다 빠르고 정확한 유형 명시가 가능하였다. 기사 추출 단계에서는 한국 신문 기사의 일반적인 지면 배열 원칙에 대한 정보를 이용하였다. 신문에 사용되는 문자의 크기가 다양함으로 문자 단위 영상을 추출하기 위해서 일반적인 높이/너비 비율을 이용하여 주어진 텍스트 열의 너비 또는 높이로부터 추출하고자 하는 문자의 크기를 계산하였다. 실험을 통하여 본 시스템의 성능의 우수성이 입증되었다

ABSTRACT

In this paper, we describe a Korean newspaper processing system which can extract article images automatically from Korean newspaper image. This system consists of three modules; 1) newspaper region identification module which segments newspaper image into regions such as titles, article bodies, pictures and figures, 2) newspaper article tracing module which form an article by connecting separated newspaper regions and 3) character extraction module which segments text into character images.

Both of statistical characteristics of regions and ruled lines are used in the region identification process. By utilizing the information of ruled lines, regions are identified much faster and more accurately than those based on only statistical information. In the article extraction process, the knowledge about general layout principles of Korean newspapers is utilized. Since characters may appear in different size, a single character size is calculated from given line width or line height considering the typical height/width ratio of Korean fonts. The performance of the system was assured through a set of experiments.

* 정 회 원 기아산업 중앙연구소 연구원
** 종신회원 한국과학기술원 전산학과 박사과정
*** 종신회원 한국과학기술원 전산학과 교수
접수일자 1988 3 15

1. 서 론

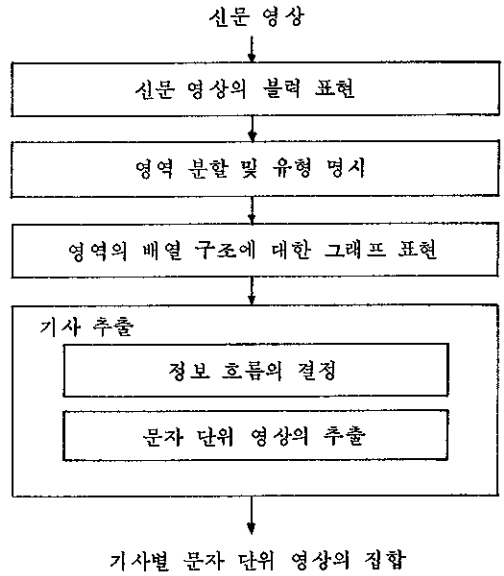
오늘날 일상적인 정보의 전달에는 서류, 잡지, 교과서 등과 같이 종이가 필수 불가결한 매체로 사용되고

있다. 컴퓨터의 발달과, 정보 통신망의 확충으로 인하여 전자 수단에 의한 정보의 전달이 보다 실용적이고 경제성이 높아지자 사람들은 종이에 이미 기록된 정보를 컴퓨터에 사용하기 적절한 형태로 변환할 필요가 생겼으며 사무 자동화에 있어서도 이러한 요구는 높은 수준의 자동화를 위해서 필수적인 것으로 인식되고 있다. 그러나 사람이 키보드를 사용하여 정보를 입력하는 방식은 많은 비용이 들어갈 뿐만 아니라 문자와 그림이 혼용된 문서의 입력이 불가능하다는 문제점을 안고 있다. 또한 이미 실용화 단계에 있는 광문자 인식기(Optical Character Reader)는 미리 정해진 형태와 크기의 문자만을 인식할 수 있다는 한계성을 갖고 있다[1, 2,3,4]. 대부분의 문서가 문자와 그림을 혼용하고 있고 다른 크기의 활자체를 혼용한다는 점을 고려해 볼 때 새로운 입력 방식의 요구는 당연한 귀결이라 하겠다.

문서 자동 입력 시스템은 문서로부터 사용자가 얻고자하는 정보를 자동적으로 추출하여 컴퓨터에 입력시켜주는 시스템이다[1,3]. 문서 자동 입력 시스템의 가능한 응용분야는 발간된 안내서, 교과서의 재편집 그리고 제출된 논문을 잡지에 발표하기 위한 일정한 형식으로의 변환, 문서를 저장하기 위한 효율적인 방법등이 있다[2] 일반적으로 문서 자동 입력 시스템은 스캐너 또는 카메라를 통해서 얻어진 문서에 대한 영상을 입력으로 하고, 그 구성은 문서 영상으로부터 문자 단위의 영상을 추출하는 전처리 단계와 추출된 문자 영상을 인식하는 문자 영상 인식 단계로 이루어진다

신문은 현대 사회에서 정보 전달의 중요한 위치를 차지하고 있다 따라서 신문을 자동적으로 입력하고 관련된 정보를 검색할 수 있는 시스템을 개발하는 것은 사회적인 관점에서 볼 때에도 중요한 의미를 갖는다[1,2, 3,4,5]. 최근들어 일본에서 영상 처리 기법을 이용하여 신문을 자동적으로 입력하기 위한 연구[5,6,7]가 일부 대학 및 연구소를 중심으로 활발히 진행되고 있으나 국내에서의 이와 관련된 연구는 전무한 실정이다. 본 논문에서는 이러한 신문 자동 입력 및 검색 시스템의 전처리 시스템으로서 신문 영상으로 부터 기사의 문자 단위 영상을 자동적으로 추출하는 시스템이 소개되는데, 이 시스템은 영역 분할 및 유형 명시, 기사 추출, 문자 단위 영상의 추출을 구성 요소로 가지며 시스템의 전체 구성도는 (그림 1)과 같다

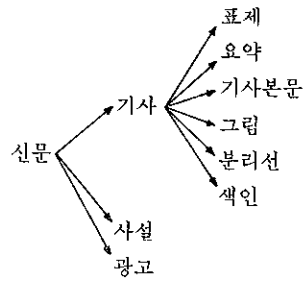
본 논문의 구성을 살펴보면 2장에서는 한국 신문 의 구조적 특성이 분석되고, 3장에서는 2장에서 분석한 구조적 특성을 이용하여 신문 영상으로부터 기사를 추출하는 시스템이 소개된다 4장에서는 본 논문에서 소개된 시스템을 이용한 실험 및 결과 분석을 보이며 끝으로 5장에서는 결론 및 향후 연구 방향이 다루어진다.



(그림 1) 시스템의 구성도

2. 신문의 구조적 특성 분석

이 장에서는 한국 신문에 사용되는 유형의 종류를 정의한 다음 이들 각각의 특성을 분석하고 신문 지면에서의 각 유형에 대한 지면 배열 원칙을 조사한다. 한국 신문의 내용은 구조적 특성상 (그림 2)와 같이 기사, 사설, 광고로 나눌수 있다.



(그림 2) 일반적인 한국 신문의 구성

(그림 3)에서 한국 신문의 한 예를 볼 수 있는데 일반적으로 한국 신문의 경우에 신문의 각면은 14개의 같은 크기를 갖는 수평 단으로 나뉘어지며 기사는 몇개의

수평 단에 분리되어 배치된다. 신문의 각 기사는 몇개의 수평 단에 분포되어 있는 영역으로 구성된다고 볼 수 있으며 이러한 영역은 아래에 설명되는 유형 중 어느 하나에 속한다. 영역에 담겨 있는 내용에 따라서 유형을 다음과 같이 구분할 수 있다.

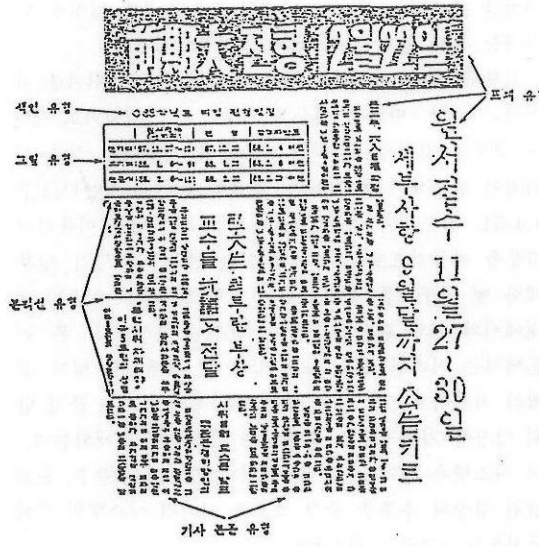
1) 표제 유형(Title type) : 기사의 표제어
2) 기사 본문 유형(Article body type) : 기사 본문 내용
3) 요약 유형(Abstract type) : 기사 내용의 요약
4) 그림 유형(Picture and figure type) : 기사 내용에 삽입된 그림, 표, 그래프
5) 색인 유형(Index type) : 그림을 설명하는 글
6) 분리선 유형(Ruled line type) : 신문의 지면 구성에 사용되는 것으로서 주로 수직선과 수평선을 사용함

이러한 6가지 유형의 종류에 따라 (그림 3)에 예시된 신문에서의 유형을 표시하면 (그림 4)와 같다.

신문은 스캐너를 통하여 입력되는데 본 연구에서 사용된 스캐너는 A4용지 크기의 신문을 입력할 수 있으며 신문에 대한 이진 영상을 출력한다. 스캐너(인치당 200화소의 해상도)에서 얻은 이진 영상의 화소를 크기의 기본 단위로 할 때, 한국 신문의 경우에 위에서 정의한 각 유형별로의 통계적 특성은 다음과 같이 분석되었다.

1) 표제 유형 :
   a) 요약 유형과 기사 본문 유형에 사용되는 문자보다 더 큰 문자를 사용하며 문자의 크기가 다양하다.
   b) 밑바탕에 간단한 무늬를 갖는 경우도 있다.
   c) 밑바탕에 무늬를 갖는 표제 유형의 높이와 너비 간에 비율이 보통 3보다 크다.
   d) 수직으로 쓰여지는 경우와 수평으로 쓰여지는 경우가 있다.
2) 기사 본문 유형
   a) 사용된 문자는 너비가 약 21화소 그리고 높이는 약 15화소의 크기를 갖는다.
   b) 기사 본문 유형의 영역은 보통 180화소의 높이를 갖는다.
3) 요약 유형 :
   a) 기사 본문 유형에 사용된 문자와 같은 크기의 문자를 사용한다.
   b) 수직으로 쓰여지는 경우와 수평으로 쓰여지는 경우가 있다.

(그림 3) 한국 신문의 한 예



(그림 4) (그림 3)에 대한 유형의 종류

c) 수직으로 쓰여진 경우에는 더 큰 높이를 가지며 수평으로 쓰여진 경우에는 더 큰 너비를 갖는다.

4) 그림 유형 :

a) 그림 유형의 높이와 너비는 각각 100화소 이상의 크기를 갖는다.

b) 그림 유형의 높이와 너비간의 비율은 보통 3보다 작다.

5) 색인 유형 :

a) 그림 유형에 근접하여 존재한다.

b) 수평 성분과 수직 성분의 두 종류가 있다.

c) 사용된 문자의 크기가 50화소보다 작다.

6) 분리선 유형 :

a) 수평 성분과 수직 성분 두 종류가 있다.

b) 분리선 유형의 높이와 너비간의 비율은 보통 20보다 크다.

기사는 한 개 이상의 표제 영역, 요약 영역, 기사 본문 영역, 그림 영역과 같은 여러 유형의 영역으로 구성되며 각 기사를 구성하는 영역은 어떤 규칙에 의해서 배열되어 있다. 이러한 규칙은 정확하게 문서화된 공식적인 규칙은 아니고 신문을 작성하는 사람과 읽는 사람이 갖고 있는 경험적 지식을 말한다. 본 논문을 통하여 분석된 한국 신문 기사의 지면 배열(Page layout)의 원칙은 다음과 같다.

1) 표제 유형의 영역은 기사 본문 내용이 시작되는 기사 본문 유형의 영역 오른쪽 또는 윗쪽에 위치한다.

2) 요약 유형의 영역은 기사 본문 내용이 시작되는 기사 본문 유형의 영역과 표제 유형의 영역 사이에 위치한다.

3) 기사 본문 유형의 영역은 신문의 수평 단 내에서는 수평 단의 오른쪽에서 왼쪽으로 배열된다.

4) 신문의 한 수평 단에서 다른 수평 단으로 기사 본문이 배열될 때는 수평 분리선이 수평 단간에 위치하고 있는 그 수평 분리선 아랫 부분에 근접된 영역중 제일 오른쪽 영역으로 연결된다.

5) 기사 본문 내용은 표제 유형, 수직 분리선 유형, 그림 유형을 가로질러서 배열되지 않는다.

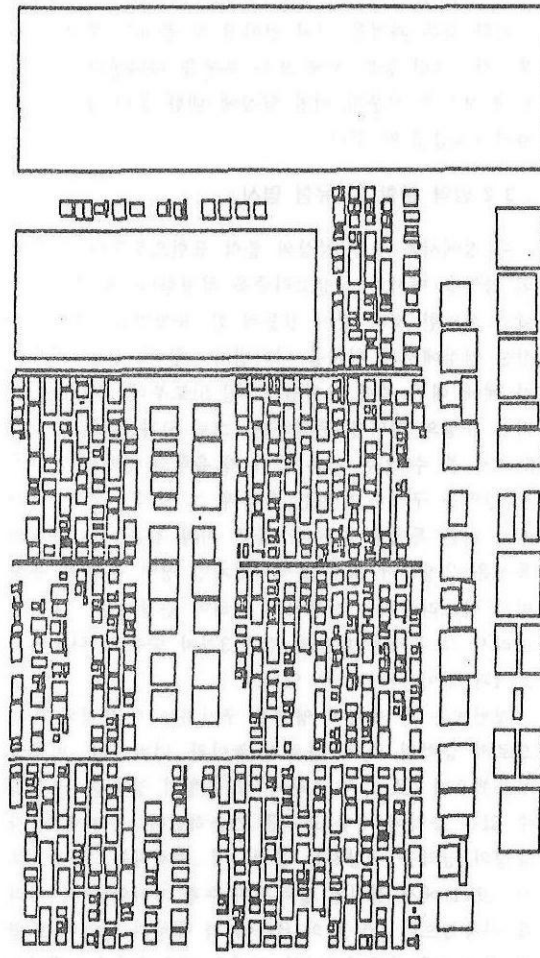
여기서 분석된 각 유형의 통계적 특징 및 지면 배열에 대한 여러가지 원칙은 3장에서 설명될 영역의 유형 명시 및 기사 추출과정에서 사용된다.

3. 신문의 구조적 특성 분석을 통한 기사의 문자 단위 영상 추출

3.1 신문 영상의 블럭 표현

스캐너를 통하여 입력되는 신문 영상은 이진 영상으로서 신문의 검정 부분은 1인 화소 값으로 흰 부분은 0인 화소 값으로 표현된다. A4용지 크기의 문서 영상은 보통 10^6-10^7개의 화소로 구성되므로 전처리 단계의 전과정에서 이러한 문서 영상을 계속 반복하여 사용할 때는 엄청난 처리 시간을 요구하게 된다. 따라서 본 논문에서는 신문 영상을 효율적으로 표현하기 위하여 신문 영상으로부터 연결된 1인 값을 갖는 화소의 블럭 표현을 구한다.

신문 영상에 대한 블럭 표현을 얻기 위해서 신문 영상의 왼쪽 윗모서리를 원점(x=0, y=0)으로 하고 다음과 같은 처리를 한다. 연결된 화소는 8방향 연결성에



(그림 5) (그림 3)의 신문 영상에 대한 블럭 표현

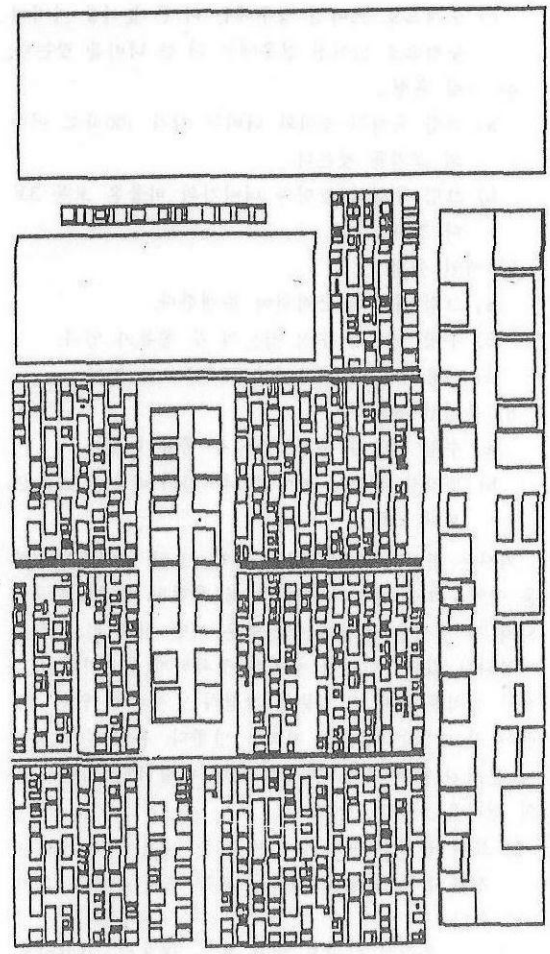
만족되는 화소의 집합을 의미하며, 이의 블럭 표현은 연결된 화소의 집합을 포함하는 가장 작은 크기의 직사각형을 나타낸다[6]. 이 직사각형의 왼쪽 윗모서리의 좌표를 (x_s, y_s)라 하고 오른쪽 아랫모서리의 좌표를 (x_e, y_e)라 할 때 블럭은 (x_s, y_s, x_e, y_e)로 표현된다. 이러한 블럭의 리스트는 이진 영상을 행별로 살펴봄으로써 얻을 수 있다. 즉, 현재 관찰되는 행으로부터 얻어진 연결된 1인 화소의 블럭과 그 전 행까지 해서 얻어진 블럭 중에서 서로 연결된 블럭이 있으면 이 두 블럭을 한 블럭으로 병합한다. 예를 들어 만약 블럭 $(x_{is}, y_{is}, x_{ie}, y_{ie})$와 블럭 $(x_{js}, y_{js}, x_{je}, y_{je})$가 서로 연결되었다면 블럭 $(\min\{x_{is}, x_{js}\}, \min\{y_{is}, y_{js}\}, \max\{x_{ie}, x_{je}\}, \max\{y_{ie}, y_{je}\})$으로 병합한다. 이와 같은 작업을 이진 영상 전체에 대해서 수행하여 입력된 신문 영상에 대한 블럭 표현을 얻는다.

위와 같은 과정을 거쳐 얻어진 각 블럭은 문자 영상, 분리선, 그림 등의 전체 또는 부분을 나타낸다. (그림 3)에 보여준 신문의 이진 영상에 대한 블럭 표현을 구하면 (그림 5)와 같다.

3.2 영역 분할 및 유형 명시

이 절에서는 신문 영상의 블럭 표현으로부터 각 유형의 영역을 명시하는 알고리즘을 설명한다. 영역은 2장에서 설명한 것과 같이 신문의 한 유형만을 포함하는 신문 영상에서의 위치를 나타낸다. 한국 신문 기사의 한 예에 대한 블럭 표현인 (그림 5)로부터 2장에서 정의한 유형의 영역을 분할하면 (그림 6)과 같다. (그림 6)에서 볼 수 있는 것과 같이 각 유형의 영역에 따라서 그 영역을 구성하고 있는 블럭과 그 영역 자신은 각각 서로 다른 특징을 갖는다. 물론 이와 같은 각 유형의 특징은 2장에서 모두 분석되었지만 영역 분할 및 유형 명시 알고리즘에서 사용하기 위하여 관찰되는 대상에 따라서 그 특징을 다음과 같은 3가지 종류로 나누어서 표 1과 같이 나타낼 수 있다.

첫번째는 각 유형의 영역을 구성하는 각 블럭의 특징으로서 블럭의 크기, 블럭의 높이와 너비간의 비율을 사용했으며 〈표 1〉의 A란에 각 유형의 영역을 구성할 수 있는 블럭에 대한 조건을 기술하였다. 두번째는 각 유형의 영역을 구성하는 블럭간에 존재하는 특징으로서 〈표 1〉에서 정의된 블럭간의 수평 거리와 수직 거리를 사용했으며 〈표 1〉의 B란에 각 유형의 영역에 포함될 블럭간에 만족되어야할 조건을 기술하였다. 세번째는 각 유형의 영역 자체에 대한 특징으로서 영역의 크기, 영역의 높이와 너비간의 비율을 사용했으며 각 유형의 영역에 대한 조건을 〈표 1〉의 C란에 나타내었다.



(그림 6) (그림 5)의 블럭 표현에 대한 영역의 분할

본 논문에서는 〈표 1〉과 같은 각 유형의 계통적 특성을 주로 이용하는 Toyoda 등의 유형 명시 알고리즘[5]를 기본적으로 사용했으나 분리선 유형의 지면 분리 특성을 고려하여 보다 정확하고 빠른 유형 명시 알고리즘을 개발하였다. 분리선은 신문의 구조를 나타내는 유형으로서 신문 지면을 효율적으로 사용하기 위해서 신문 지면을 몇개의 구간으로 분리한다. 이러한 분리선의 특성을 이용하여 보다 정확하고 빠른 유형 명시가 가능하였다. 유형 명시 순서는 분리선, 기사 본문, 표제, 그림, 색인순이며 임의의 유형 T_i를 명시하기 위해 유형 명시 알고리즘은 항상 〈표 1〉을 참조한다.

알고리즘에 대한 입력을 이진 영상에 대한 블럭리스

트 P로 할 때 각 유형 T_i를 명시하는 알고리즘을 간략하게 기술하면 아래와 같다.

1) 블럭 리스트 P로부터 〈표 1〉의 A조건을 만족하는 블럭을 리스트 R에 모은다
2) 분리선 유형이 명시되었다면, 분리선의 지면 분리에 의해서 신문 영상이 몇개의 구간으로 나누어질때 각 구간에 해당되는 부분 리스트 S_j를 R로부터 구성한다.
3) 분리선 유형이 명시되지 않았다면, 리스트 R을 한개의 부분 리스트 S_1로 한다.
4) 만약 부분 리스트 S_j에 〈표 1〉의 B조건을 만족하는 블럭이 없다면, 7)번 단계로 간다.
5) 만약 부분 리스트 S_j에 〈표 1〉의 B조건을 만족하는 한쌍의 블럭이 있다면, 이 두 블럭을 병합하여 S_j에 넣고 S_j로부터 이 두 블럭을 삭제한다.
6) 다시 단계 4)로 간다.

〈표 1〉 신문의 각 유형에 대한 통계적 특성(유형의 일부분)

유 형	A	B	C
수평 분리선	(3<높이<10) or (10<너비/높이)	(DX_{ij}<11) and (DY_{ij}=0)	(높이<30) and (너비/높이>20)
수직 분리선	(3<너비<10) or (10<높이/너비)	(DY_{ij}<11) and (DX_{ij}=0)	(너비<30) and (높이/너비>20)
기 사 본 문	(3<너비<22) and (3<높이<180)	(DX_{ij}<10) and (DY_{ij}<20)	(너비>19) and (140<높이<200)

단 임의의 블럭 B_i와 블럭 B_j가 아래와 같이 표현될때

$$B_i=(x_{0i}, y_{0i}, x_{1i}, y_{1i})$$
$$B_j=(x_{0j}, y_{0j}, x_{1j}, y_{1j})$$

블럭 B_i의 너비(width)와 높이(height) 및 블럭 B_i와 블럭 B_j간의 수평거리(horizontal distance) DX_{ij}와 수직 거리(vertical distance) DY_{ij}는 아래와 같이 정의된다

$$\text{너비}=x_{1i}-x_{0i}+1$$
$$\text{높이}=y_{1i}-y_{0i}+1$$

$$DX_{ij}=\begin{cases} x_{0j}-x_{1i}, & \text{if } x_{0j}>x_{1i} \\ x_{0i}-x_{1j}, & \text{if } x_{0i}>x_{1j} \\ 0, & \text{그 외의 모든 경우} \end{cases}$$

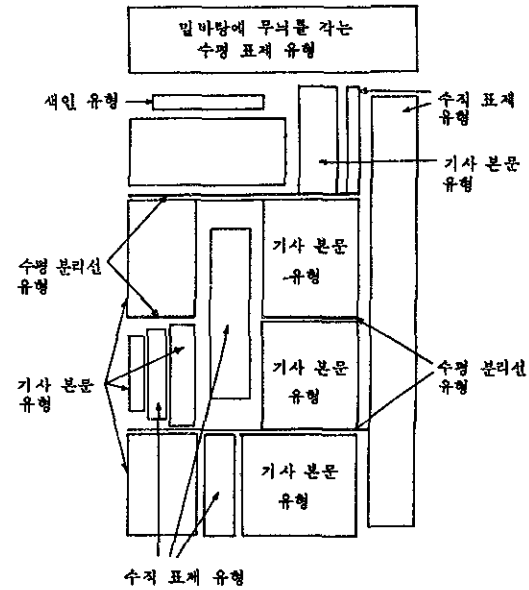
$$DY_{ij}=\begin{cases} y_{0j}-y_{1i}, & \text{if } y_{0j}>y_{1i} \\ y_{0i}-y_{1j}, & \text{if } y_{0i}>y_{1j} \\ 0, & \text{그 외의 모든 경우} \end{cases}$$

7) 모든 부분 리스트 S_j에 대해서 단계 4)부터 단계 6)의 작업을 한다.
8) 각 S_j의 블럭에 대해서 〈표 1〉의 C조건을 만족하면 그 블럭을 유형 T_i로 명시하고 P로부터 이 블럭에 포함된 블럭을 삭제한다.
9) 모든 유형에 대해서 위의 작업을 한다.

Toyoda 등의 알고리즘에서는 단계 1)에서 모아진 블럭에 대해서 바로 루프(loop) 4)-6)을 처리하였다. 루프 4)-6)에서는 블럭간의 거리를 조사하여 〈표 1〉의 조건 B에 만족하는 블럭을 한 영역으로 병합한다. 본 논문에서는 단계 2)에서 분리선을 이용하여 단계 1)에서 모아진 블럭을 몇개의 블럭의 리스트로 분리하고 분리된 각 블럭의 리스트에 대해서 루프 4)-6)을 처리함으로서 Toyoda 등의 방법처럼 모아진 블럭에 대해서 바로 루프 4)-6)을 처리하는것 보다 계산량을 줄였다. 또한 각 영역이 밀집되어 있을때 분리선을 이용하지 않고 블럭간의 거리만을 이용해서 루프 4)-6)에서 블럭을 병합하게 되면 잘못된 영역 분할이 발생한다. 신문 영상 (그림 3)에 대하여 분리선을 이용한 영역 명시 결과는 (그림 7)과 같다.

3.3 영역의 배열 구조에 대한 그래프 표현

신문 영상으로부터 기사를 추출하기 위해서는 입력된



(그림 7) (그림 3)의 신문 영상에 대한 영역 명시 결과

신문 영상의 구조에 대한 정보가 필요하다. 신문 영상은 3.2절에서 명시한 영역으로 구성된다고 할 수 있으며 신문 영상의 구조는 결국 이러한 영역의 배열 형태에 의해서 나타낼 수 있다. 그러므로 이 절에서는 3.2절에서 명시된 영역으로부터 신문 영상에서의 영역의 배열 형태를 구하는 방법을 설명한다. 영역의 배열 형태를 분석하기 위해서 먼저 각 영역의 상하 좌우 방향을 정의한다 각 영역은 4개의 면을 갖는 블럭이므로 각 면에 그 영역의 상하 좌우와 4방향을 대응시킨다. 이와 같은 영역의 방향성이 정의되었을 때 입력된 신문 영상에 있는 각 영역의 상하 좌우 방향에 근접된 영역에 대한 정보가 영역의 배열 형태를 나타낸다고 볼 수 있다. 신문 영상은 영역을 원소로 갖는 집합 $P(\{R_l \mid R_l=(x_{0l}, y_{0l}, x_{1l}, y_{1l}), l=1...n\})$로 가정할 수 있으며 신문 영상 P에 있는 임의의 영역 $R_i(x_{0i}, y_{0i}, x_{1i}, y_{1i})$와 영역 $R_j(x_{0j}, y_{0j}, x_{1j}, y_{1j})$의 근접성은 다음과 같이 정의될 수 있다[7]. 두 영역 R_i와 R_j간의 거리는 3.2절에서 정의한 블럭간의 거리에 대한 개념을 사용한다.

모든 영역 $R_k(\{R_k \mid R_k \in P-\{R_i, R_j\}, DY_{ik}=0, DY_{jk}=0\})$에 대해서 식 (1)을 만족하는 영역 R_k가 존재하지 않는다면 영역 R_i와 영역 R_j는 수평 근접성을 갖는다

$$1<x_{0k}<x_{1k}<r \quad \cdots\cdots \text{식 (1)}$$

단 $l=\min\{x_{1i}, x_{1j}\}$

$r=\max\{x_{0i}, x_{0j}\}$

또한 모든 영역 $R_k(\{R_k \mid R_k \in P-\{R_i, R_j\}, DX_{ik}=0, DX_{jk}=0\})$에 대해서 식 (2)을 만족하는 영역 R_k가 존재하지 않는다면 영역 R_i와 영역 R_j는 수직 근접성을 갖는다.

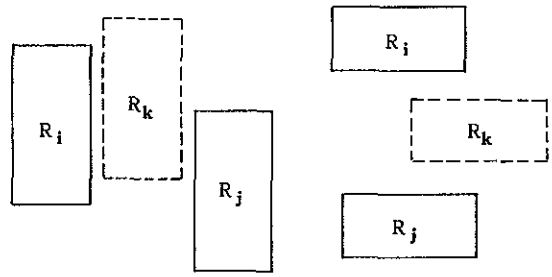
$$u<y_{0k}<y_{1k}<d \quad \cdots\cdots \text{식 (2)}$$

단 $u=\min\{y_{1i}, y_{1j}\}$

$d=\max\{y_{0i}, y_{0j}\}$

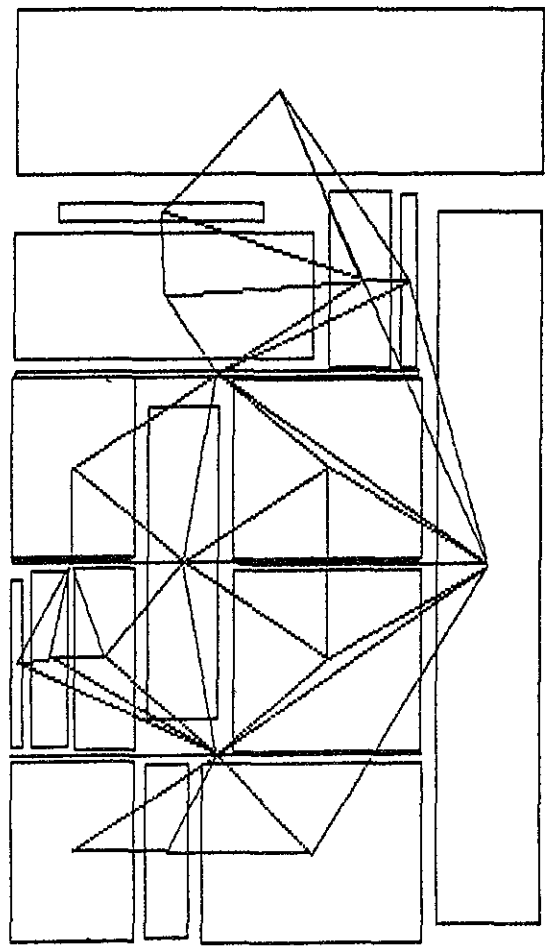
(그림 8)에서 두 영역 R_i와 R_j가 위에서 정의한 근접성을 갖기 위해서는 영역 R_k가 존재하지 않아야 한다

이러한 영역의 배열 구조에 대한 정보는 그래프 형태로 간직하는데 그래프의 노드(node)는 각 영역에 대한 정보(유형, 위치)를 나타내며 가지(edge)는 근접성 관계를 갖는 노드를 연결한다 가지는 상하 좌우 방향에 의해서 4종류로 나누어진다. (그림 7)에서 보여준 명시된 영역에 대해서 이와 같은 분석을 통하여 신문 영상의 구조를 나타내는 그래프를 구하면 (그림 9)와 같다.



(a) 수평 근접성 (b) 수직 근접성

(그림 8) 영역 R_i와 R_j의 수평 근접성 및 수직 근접성



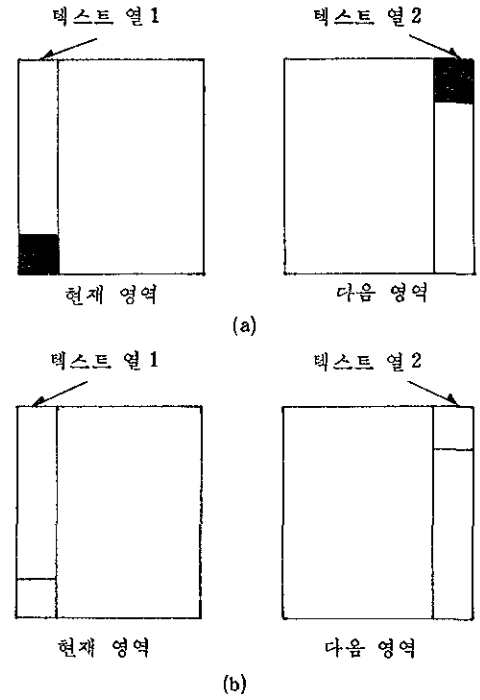
(그림 9) (그림 7)의 영역의 배열 형태에 대한 그래프 표현

3.4 기사 추출

신문 영상으로부터 명시된 각 유형의 영역은 신문 영상의 각 부분에 해당되므로 영역 그 자체만으로는 신문의 올바른 정보를 나타내지 못한다. 따라서 이러한 영역이 적절한 순서로 연결되었을 때 그 신문 영상에 대한 올바른 정보를 나타낸다. 이와 같이 문서 영상으로부터 올바른 정보를 얻기 위해서 명시된 영역을 적절한 순서로 연결하는 것을 정보 흐름의 결정이라고 한다. 본 논문에서는 문자 영상을 인식하지 않고 신문의 지면 배열 특성만을 이용하여 정보의 흐름을 결정하는 방법을 설명한다.

교과서와 같이 간단한 구조를 갖는 경우에는 많은 노력을 들이지 않고 정보 흐름을 결정할 수 있으므로 종래 문서 인식 시스템은 문서 자동 입력 시스템의 전처리 단계에서 입력된 문서에 대한 정보 흐름의 결정을 중요시하지 않았다 그러나 신문은 한 면에 여러개의 기사가 복잡한 구조로 배열되어 있으므로 신문의 정보 흐름 결정 단계는 신문 자동 입력 시스템에 있어서 중요한 역할을 담당하게 된다 기사 영역의 추출은 신문 영상의 구조를 나타내는 그래프와 신문의 지면 배열 원칙을 기반으로 한다 기사 영역을 추출하기 위해서 필요한 구성 요소는 기사의 시작 부분을 결정하는 단계와 기사 본문 내용을 찾기 위해서 기사 본문 유형의 현재 영역으로부터 기사 본문 유형의 다음 영역을 결정하는 단계, 그리고 결정된 기사 본문 유형의 현재 영역과 다음 영역의 정보 연결성을 조사하여 기사의 종료를 결정하는 단계로 구성된다.

기사 추출은 한 기사의 표제 유형을 찾고 이로부터 기사 본문의 시작부분에 해당되는 영역을 찾는일로 시작된다. 표제 영역의 왼쪽 또는 아랫쪽에 기사 본문의 시작 부분이 배열된다는 신문의 지면 배열 원칙을 이용하여 기사 본문 내용의 시작 부분에 해당되는 기사 본문 유형의 영역을 찾는다. 기사의 표제 및 기사 본문의 시작 부분을 찾은 다음에는 기사 본문의 시작 영역을 기점으로 해서 그 기사에 포함되는 기사 본문 영역을 정보 흐름에 따라서 결정해간다 기사 시작 부분의 기사 본문 영역을 현재 영역이라 하고 이로부터 신문의 지면 배열 원칙과 신문 영상의 구조를 나타내는 그래프를 이용하여 기사 본문 유형의 다음 영역을 결정하기 위해서 아래와 같은 처리를 한다 기사 본문 유형의 영역에 있는 각 열은 수직으로 쓰여져 있으며 신문의 지면 배열 특성상 수직으로 쓰여진 열은 신문 영상의 오른쪽에서 왼쪽으로 정보 흐름이 존재하므로 현재 영역의 왼쪽에 근접된 영역을 조사한다. 만약 현재 영역의 왼쪽에 기사 본문 유형의 영역이 존재하면 이 영역을 다음 영역으로 결정한다. 그러나 현재 영역의 왼쪽에 기사 본문 유형이 아닌 다른 유형이 존재하면 현재 영역의 아랫쪽에 근접된 영역을 조사한다 현재 영역의 아랫쪽에 수평 분리선 유형 이외의 다른 유형이 존재하면 현재 영역에서 본 기사 내용의 종료를 결정한다. 현재 영역의 아랫쪽에 수평 분리선 유형이 있는 경우에는 이 분리선 영역의 아랫쪽에 근접된 영역 중에서 제일 오른쪽에 있는 영역을 조사하여 기사 본문 유형이면 이 영역을 다음 영역으로 결정하고, 그렇지 않으면 현재 영역에서 기사의 종료를 결정한다 이와 같은 방법으로 기사 본문 유형의 현재 영역으로부터 지면 배열 특성을 이용하여 기사 본문 유형의 다음 영역을 결정하거나 기사의 종료를 결정할 수 있다.



(그림 10) 두 기사 본문 유형의 영역이 연결될 수 있는 조건

또한 신문의 지면 배열 특성에 의해서 기사 본문 의 현재 영역으로부터 기사 본문의 다음 영역이 결정되었을 때 현재 영역과 다음 영역의 정보 연결성을 조사하

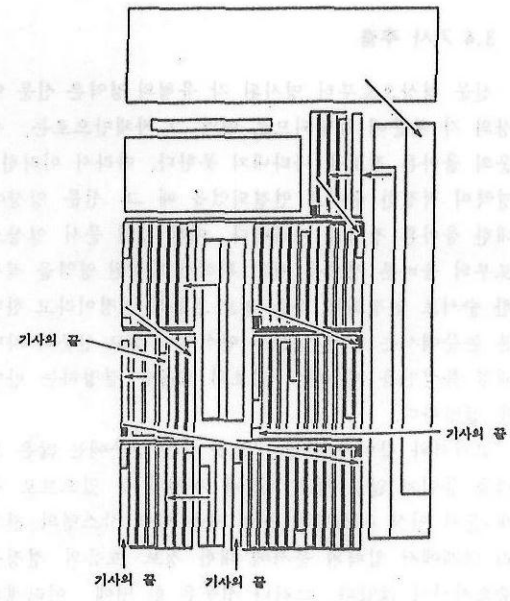
여 현재 추출되는 기사의 종료를 결정한다. 이 단계에서는 문자 영상을 인식하지 않기 때문에 현재 영역과 다음 영역의 상태를 조사하여 정보의 연결성을 결정하는데 다음과 같은 방법을 사용한다. 현재 영역의 제일 왼쪽에 있는 텍스트 열(text line) 1과 다음 영역의 제일 오른쪽에 있는 텍스트 열 2가 현재 영역과 다음 영역간의 정보 연결성에 대한 고려 대상이 된다. 이 두 영역이 연결될 수 있는 조건을 그림으로 보면 (그림 10)과 같다. (그림 10)의 (a)는 텍스트 열 1과 텍스트 열 2가 한 문장을 구성할 수 있는 경우 또는 텍스트 열 1은 한 문장의 끝 부분 그리고 텍스트 열 또 다른 문장의 시작 부분이지만 이 두 문장이 같은 절에 속해 있는 경우에 열 1과 열 2 사이에 흰 공간이 별로 없다는 것을 나타내며 연결될 수 있는 또하나의 경우인 (그림 10)의 (b)는 텍스트 열 1과 텍스트 열 2가 각기 다른 절에 포함되어 있는 문장의 끝 부분과 시작 부분인 경우에 텍스트 열 1과 텍스트 열 2 사이에 많은 흰 공간이 존재하는 것을 나타낸다. 이와 같은 방법으로 지면 배열 원칙에 의해서 결정된 현재 영역과 다음 영역으로부터 정보 연결성을 조사하여 기사의 종료를 결정할 수 있다. 신문의 지면 배열 원칙을 사용한 기사 추출 알고리즘을 간략하게 기술하면 다음과 같다.

1) 표제 유형을 찾는다.
2) 만약 표제 유형이 없으면 기사 추출 작업을 마친다.
3) 근접된 모든 표제 유형을 찾는다.
4) 표제 유형으로부터 기사 본문의 시작부를 찾는다 (이 영역을 현재 영역이라고 하자).
5) 현재 영역을 문자 영상 추출 단계로 보낸다.
6) 지면 배열 원칙을 이용하여 현재 영역으로부터 다음 영역을 찾는다.
7) 만약 다음 영역이 없으면 현재 기사의 종료를 알리고 1) 단계로 간다.
8) 현재 영역을 다음 영역으로 대치하고 단계 5)로 간다.

(그림 7)의 영역에 대해서 신문의 지면 배열 원칙을 사용한 기사 추출 알고리즘을 적용했을 때 (그림 11)과 같이 각 기사 영역을 추출하게 된다.

3.5 문자 단위 영상의 추출

우리 나라 신문에 주로 사용되는 문자의 종류는 한글, 한문, 영문, 숫자, 특수 문자 등이 있다. 문자 단위 영상의 추출은 텍스트 열 영상으로부터 이러한 각 종류의 문자를 추출한다. 신문의 경우 표제 유형에 사용되는 문자의 크기가 가변적이므로 텍스트 열 영상으로부터 그 텍스트 열에 사용된 문자의 크기를 구해야 정확한 문자 단위 영상의 추출이 가능하다.

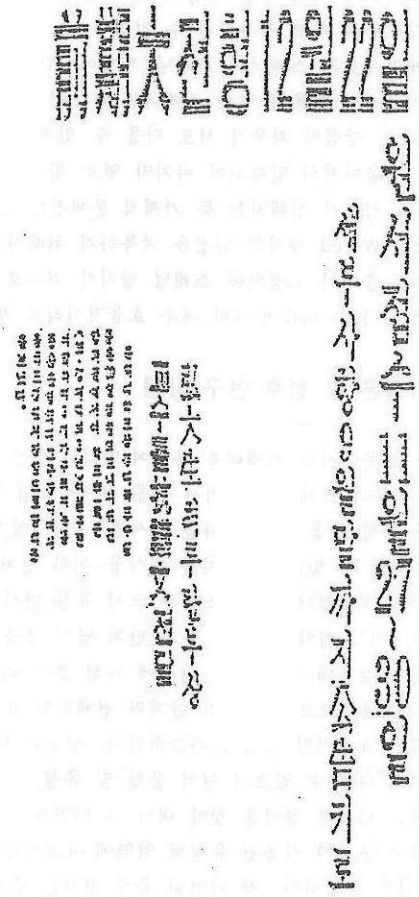


(그림 11) (그림 7)의 명시된 영역으로부터 각 기사의 영역 결정

본 논문에서는 신문의 텍스트 열 영상으로부터 문자 단위 영상을 추출하기 위해서 텍스트 열에 대한 투사와 일반적인 문자의 높이와 너비간의 비율을 이용했다. 텍스트 열이 수직으로 쓰여진 경우에는 수평 방향으로 투사하고 수평으로 쓰여진 경우에는 수직 방향으로 투사한다. 텍스트 열 영상의 투사 결과로부터 어떤 임계치 보다 큰 두께로 연결된 1인 화소에 대한 텍스트 열 영상 내에서의 위치를 구한다. 이 결과는 텍스트 열 영상 내에서 일반적으로는 문자 영상의 전체 또는 부분에 대한 위치를 나타내지만 스캐너를 통하여 얻은 텍스트 열 영상의 경우에는 여러개의 문자가 붙어있는 경우도 있으므로 여러 문자 영상의 위치를 포함하기도 한다. 주어진 텍스트 열 영상으로부터 문자의 크기를 구하는 방법은 다음과 같다. 수직으로 쓰여진 텍스트 열은 텍스트 열 영상의 너비가 그 텍스트 열 영상에사용된 문자의 너비를 가장 잘 나타내주므로 이를 문자의 너비에 대한 근사치로 사용한다. 수직으로 쓰여진 텍스트 열의 경우에 문자 영상을 추출하기 위해서 필요한 정보는 문자의 높이에 대한 값이므로 문자의 너

비에 대한 근사치로부터 문자의 높이/너비 비율을 이용하여 문자의 높이를 구한다. 수평으로 쓰여진 경우도 마찬가지 방법으로 문자 영상의 크기를 구할 수 있다.

문자 크기에 대한 값과 투사 결과를 이용하여 문자 영상을 추출한다. 텍스트 열 영상의 투사 결과는 문자 영상의 일부분이거나 문자 영상 전체, 또는 여러개의 문자 영상을 포함하는 텍스트 열 영상 내에서의 위치를 나타내므로 텍스트 열 영상으로부터 예측한 문자의 크기를 이용하여 텍스트 열 영상 내에서 정보 흐름에 맞는 순서, 즉 수직으로 쓰여진 텍스트 열이면 텍스트 열 영상의 위에서 아래로 그리고 수평으로 쓰여진 텍스트 열이면 텍스트 열 영상의 왼쪽에서 오른쪽으로 투사 관찰한다. 관찰되는 투사 결과가 예측된 문자 크기와 비교해서 더 작은 크기이면 이 위치는 문자 영상의 일부분을 포함한다는 것을 알고서 그 다음 순서에 조사할 투사 결과와 연결해서 다시 예측된 문자 크기와 비교하고 만약 예측된 문자의 크기 보다 두 배 이상 크면 이 위치는 여러개의 문자 영상을 포함하는 위치라는 것을 알 수 있으므로 예측된 문자의 크기로 분리할 수 있다. 위와 같은 과정을 거쳐 (그림 3)의 신문으로부터 추출된 기사별 문자 단위 영상의 일 부분을 같은 크기로 정규화하여 출력하면 (그림 12)와 같다.



(그림 12) (그림 3)의 신문 영상으로부터 추출된 기사별 문자 단위 영상의 일부분

4. 실험 및 결과 분석

본 시스템은 IBM-PC/AT에서 LISP 언어를 사용하여 구현되었다. 신문 영상은 CompuScan 광학 스캐너를 통하여 입력되는데 입력 가능한 신문의 최대 크기는 A4용지 크기이며 결과로서 1600x1600 크기의 이진 영상을 출력한다. 본 논문을 통하여 개발된 기사 영상 추출 시스템을 이용한 실험을 통하여 아래와 같은 결과를 얻었다.

1) 본 시스템은 한국 신문의 구조적 특성을 이용하여 신문 영상으로부터 기사별 문자 단위 영상을 추출할 수 있었다.
2) 이진 신문 영상에 대한 블럭 표현의 사용은 작업을 용이하게 할 뿐만 아니라 처리 시간을 단축할 수 있었다.
3) 영역 분할 및 유형 명시 단계에서 분리선 유형의 기능을 사용하지 않을때는 잘못된 결과를 얻게되는 경우도 있었으며 분리선 유형의 지면분리 기능을 이용하였을 때는 정확하고 빠른 유형 명시가 가능하였다.
4) 신문 영상중에서 기사 본문에 있는 문자 영상은 서로 다른 문자 영상과 붙어 있는 경우가 있었다. 이러한 경우에는 예측된 문자의 크기로 붙어 있는 문자 영상들을 분리하지만 약간씩 문자 영상이 훼손되는 경우가 있다. 이와 같이 문자 영상이 붙게 되는 현상은 원래 신문 자체에서 몇개의 문자가 약간씩 겹치게 인쇄되었거나 또는 원래 신문에서는 떨어져 있는 문자 영상들이 스캐너의 해상도의 한계에 의해서 발생할 수 있다.
5) 시스템 구현시 사용한 스캐너의 해상도로는 신문 기사의 본문 내용으로부터 추출된 문자 영상을 인식하기 어려운 상태다. 하지만 표제와 같이 큰 문자의 인식은 가능하다.
6) 스캐너로부터 얻은 1600×1600 신문 영상을 본 연

구를 통하여 구현된 시스템으로 IBM-PC/AT 상에서 수행시켰을 때 각 단계별 수행시간을 〈표 2〉와 같다. 효과적이며 편리한 시스템의 개발을 위하여 LISP언어로 시스템이 구현되었지만, 보다 빠른 처리 속도를 위하여 C언어나 어셈블리 언어를 사용하여 시스템을 재 구현하면 현재의 속도에 비해 10~100배의 속도 향상을 기대할 수 있다[8].

7) 한 라인 영상에 서로 다른 크기의 문자가 사용되었을 때는 잘못된 문자 영상이 추출되기도 하는 문제점을 안고 있다

〈표 2〉 구현된 시스템의 처리 속도
(1600×1600 이진 영상)

단 계	수행시간
신문 영상에 대한 블럭 표현	20분
영역 분할 및 유형 명시	8분
영역의 배열 구조에 대한 그래프 표현	0.5초
기사 영역 추출	1초
문자 영상 추출	25분

대부분의 영상 처리 시스템이 안고 있는 공통적인 문제점이기도 한 처리속도의 한계성 및 방법론의 제한성은 본 연구를 통하여서도 명확히 드러났다 한 예를 들자면, 기울어진 신문 영상을 바로 잡는 문제는 수학적으로는 매우 잘 정의된 문제이지만 실제 문제의 해결에 있어서는 여러가지 문제점을 노출시킨다. 우선 기울어진 각도를 계산하여야 하는데 여러가지 방법론이 제안되어 있지만 시간이 너무 소비되거나 일반적이지 못한 경우가 대부분이다 또한 각도를 알고 난 후 신문 영상을 바로 잡는 과정에서의 문제점은 다음과 같다 첫째는 속도상의 문제로 1600×1600 크기의 영상을 θ 만큼 회전시키는데 걸리는 시간은 IBM-PC/AT상에서 알고리즘을 LISP 언어로 프로그래밍했을 때 2시간 이상 걸리고, C 언어로는 10분 이상 걸리며, 어셈블러 언어로는 1분 이상 걸린다[8] 특수 목적의 메모리 시스템을 고밀도 집적 회로로 개발하여 1초 이내에 수행했다는 일본에서의 연구결과[9]가 최근에 발표되었다. 둘째는 θ 만큼 회전시켰을 때 영상 배열의 첨자를 계산하면서 발생하는 오차 때문에 영상에 변형이 생기게 된다. 한 문자 영상을 형성하는 화소가 32×32일 때 한 두 화소의 변형은 복잡한 문자 영상을 손상시키는 효과를 유발한다.

처리속도와 시스템의 성능은 서로 상충적인 관계에 있다 빠른 처리속도를 위해서 단순한 알고리즘을 사용한다면 시스템의 성능이 떨어지게 되고 견실한 알고리즘을 적용하려면 많은 시간이 소요된다. 또한 높은 해상도의 입력장치를 사용하면 보다 나은 시스템의 성능을 기대할 수 있다. 그러나 해상도가 높으면 그에 따라 처리시간도 증가하게 되는 문제점이 발생한다. 따라서 개발하고자 하는 시스템의 요구에 맞도록 적절히 시스템의 성능과 처리속도간에 균형을 맞추어야 한다.

본 연구에서 사용된 영상 입력장치인 CompuScan 광학 스캐너는 인치당 200개의 화소를 스캔할 수 있는 능력이 있다 따라서 보통 사용하는 24-포인트 활자가 32×32정도의 해상도를 갖고 있다. 이러한 환경 하에서는 노이즈에 의해 신문 영상이 많이 손상되어서 문자 인식 시스템과 직접 접속시킬 수 없었다. 또 다른 기계적인 제약으로는 스캐너의 스캐닝 방법에 기인한 것인데 본 연구에서 사용된 영상 스캐너는 Wired 방식으로서 Roller의 말림이 좌우가 서로 다를 수 있기 때문에 신문이 기울어져서 입력되어 마지막 열로 갈수록 글자의 변형의 정도가 심해지는 등 기계적 문제점이 뒤따른다. 이러한 Wired 방식의 단점을 극복하기 위해서는 복사기처럼 종이가 고정되어 스캐닝 장치가 좌우로 움직일 수 있는 Flat-bed 방식이 매우 효율적이라고 생각한다.

5. 결론 및 향후 연구 방향

본 논문에서는 스캐너를 통하여 입력된 이진 신문 영상으로부터 표제 유형, 기사 본문 유형, 그림 유형, 분리선 유형 등을 구별한 다음 기사를 문자 영상 별로 추출해낼 수 있는 신문 자동 인식을 위한 전처리 시스템이 소개되었다 본 시스템은 크게 유형 명시 및 문장 분할, 기사 영역의 추출, 문자 단위 영상 추출의 구성 단계를 갖는데 이진 신문 영상에 대한 블럭 리스트 표현을 사용함으로써 전처리 단계의 전체적인 처리속도를 높였고 효율적인 작업이 가능하였다 신문은 몇개의 영역으로 나눌 수 있으며 영역 분할 및 유형 명시 단계에서는 이러한 영역을 찾아 내서 그 영역의 유형을 명시하고 문자가 사용된 유형의 영역에 대해서는 각 텍스트 열을 분할한다 각 영역의 유형 명시는 주로 각 유형의 통계적 특성을 이용하였으며 분리선 유형의 지면 분리 특성은 보다 빠르고 정확한 유형 명시를 가능케 하였다 입력된 신문 영상의 구조를 나타내는 영역의

배열 형태를 명시된 영역으로부터 구한다 즉 명시된 각 영역에 대해서 그 영역의 상하 좌우에 근접된 영역을 조사하여 그래프 형태로 나타내는데 그래프의 노드(node)는 명시된 영역을 나타내며 그래프의 가지(edge)는 근접된 영역을 연결한다. 기사 추출 단계는 신문의 구조를 나타내는 그래프와 신문의 지면 배열 원칙을 이용하여 명시된 영역으로부터 각 기사의 영역을 결정하고 이를 정보 흐름에 따라서 연결한다 문자 단위 영상 추출은 문자의 높이와 너비간의 비율과 텍스트 열 영상의 투자 결과를 이용했다 문자 영상 추출에 대한 이 방법은 추출할 문자의 크기가 가변적인 경우에도 정확한 문자 단위의 영상을 추출하였으며 여러개의 문자 영상이 붙어 있을때는 이를 분리하여 문자 단위의 영상을 추출하였다. 본 논문에서 소개된 신문의 구조 분석을 통한 기사의 문자 단위 영상 추출 시스템은 처리할 문서의 구조적 특성만을 고려해 줌으로써 신문 뿐만 아니라 논문 및 기타 다른 문서 영상의 전처리에 쉽게 사용될 수 있으리라 기대한다.

앞으로 더 연구되어야 할 부분은 참고문헌[4]에서 지적된 바와 같은 기울어진 신문 영상의 교정에 관한 연구, 방법론의 일반화에 관한 연구, 처리속도의 개선에 관한 연구 등이다. 특히 신문의 경우에는 지면에 복잡한 구조로 여러개의 기사가 배열되어 있으므로 투사를 사용하여 기울기를 찾기 어렵고 참고문헌[10]에서와 같은 Fourier 변환을 사용한 방법은 엄청난 계산량 때문에 그 실용성에 문제가 있다 또한 밀바탕에 무늬를 갖는 표제 유형의 처리에 대해서도 연구가 더욱 필요하다.

일반적으로 영상처리는 동일한 작업을 각 화소에서 반복해야 하기 때문에 전체적으로 많은 계산량이 필요하지만 이 계산은 병렬처리하기에 적합하다 따라서 실용화의 필요성에 의해 많은 전문 병렬 프로세서와 메모리 시스템이 개발되고 있다[11]. 특히 최근에 고밀도 집적회로 기술의 발달로 병렬처리 알고리즘을 쉽게 하드웨어화할 수 있기 때문에 신문 자동 입력 시스템을 위한 고밀도 집적회로 칩의 개발도 연구해 볼만한 과제이다

참 고 문 헌

1 K Y. Wong and R. G Casey, "Document Analysis System," IBM Journal of Research and Development, Vol. 26, No 6, 1982.

2. F. M. Wahl, K. Y. Wong and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," Computer Graphics and Image Processing, Vol. 20, pp 375-390, 1982.

3 이성환, 강희중, 김형훈, 박진규, 심원태, 이승호, 김진형, "문서 인식 및 검색을 위한 전처리 시스템의 설계 및 구현," 한국정보과학회 추계 학술 발표회 논문집, pp. 503-509, 1986 년 10 월.

4. 이성환, 조창제, 김진형, "실용적 한글 문서 자동 인식 시스템 개발의 문제점 및 개선 방향," 한국정보과학회 춘계 학술 발표회 논문집, pp.127-130, 1988 년 4 월.

5 J. Toyoda, Y. Noguchi and Y. Nishimura, "Study of Extracting Japanese Newspaper Article," Proc IEEE Conf. on Computer Vision and Pattern Recognition, pp.1113-1115, 1982.

6. H. Makino, "Representation and Segmentation of Document Images," Proc. IEEE Conf. on Pattern Recognition and Image Processing, pp.291-296, 1983.

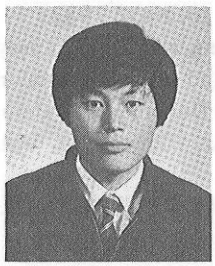
7. S. Ito and S. Sakatani, "Field Segmentation and Classification in Document Image," Proc. IEEE Conf. on Pattern Recognition and Image Processing, pp.492-495, 1982.

8. 한국과학기술원 전산학과 인공지능 연구실 메모, "Transputer 를 사용한 영상 처리," AI-TR-88-01, 1988 년 2 월.

9. A Tanaka, M. Kameyama, S. Kazuma and O Watanabe, "A Rotation Method for Raster Image using Skew Transformation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.272-277, 1986.

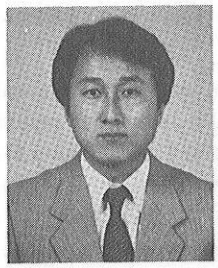
10 M. Hase and Y. Hoshino, "Segmentation Method of Document Image by Two-Dimensional Fourier Transformation," Systems and Computers in Japan, Vol. 16, No 3, pp.38-47, 1985.

11 K Inagaki, T. Kato, T. Hiroshima and T. Sakai, "MACSYM A Hierachical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents," Pattern Recognition, Vol. 17, No.1, pp. 85-108, 1984.



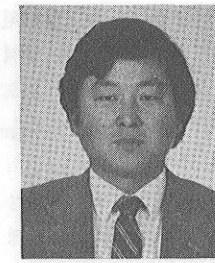
김 형 훈

1986 년 전남대학교 계산통계학과 졸업.
1988 년 한국과학기술원 전산학과 석사학위 취득.
1988 년 9 월 현재 기아산업 중앙연구소에 근무중.
주 관심분야는 사무자동화, 패턴 인식등임.



이 성 환

1984 년 서울대학교 계산통계학과 졸업.
1986 년 한국과학기술원 전산학과 석사학위 취득.
1987 년 네덜란드 Delft 공과대학초청연구원.
1988 년 9 월 현재 한국과학기술원 전산학과 박사과정 재학중.
주 관심분야는 패턴 인식, 전문가 시스템, 지능형 Man-Machine Interface 등임.



김 진 형

1971 년 서울대학교 공과대학 졸업.
1979 년 UCLA 전산학 석사학위 취득.
1983 년 UCLA 전산학 박사학위 취득.
1973 년～1976 년 KIST 전산실 연구원.
1976 년～1977 년 미국 California 도로국 연구원.
1981 년～1985 년 미국 Hughes 인공지능 센타 선임연구원.
1985 년～현재 한국과학기술원 전산학과 부교수.
1986 년～1988 년 본 학회 산하 인공지능 연구회 위원장
주 관심분야는 패턴 인식, 전문가 시스템, 지능형 Man-Machine Interface 등임.