한국어 문서로 부터 문자분리 및 도형추출에 관한 연구

(A Study on the Korean Character Segmentation and Picture Extraction from a Document)

南宮 在 贊,* 柳 煌 彬,* 南宮 璉*

(Jae Chan Namkung, Howang Bin Ryou and Yun Namkung)

要 約

본 논문에서는 한국어 문서로부터 문자분리 및 도형추출에 관해 논하였다.

한국어 문서에 대하여 먼저 횡방향에 대한 확대와 축소 조작의 반복과 run-length에 의해 문자열을 추출하고 횡방향과 종방향의 반복조작으로 개별문자 영역을 추출한 다음 정피치 문자를 우선 추출하고 그 위치정보를 이용하여 후보위치를 추가하며 복수회를 반복하므로서 한국어 문서에서 개별문자를 추출하였다.

접촉문자에 대하여는 이미 추출되어진 문자폭의 run-length를 이용하여 서로 접촉하고 있는 문자를 끊어낸 후 접촉문자를 강제적으로 추출하였고 한글의 특수성으로 인하여 반 피치로 갈라진 문자에 대해 융합처리를 하였다.

추출실험은 한국어 인쇄체 활자문자로 쓰여져 있는 월간지를 대상으로 840×600으로 입력되어진 데이타 9개에 대하여 행하였고 추출결과는 문자열에 대한 추출은 100%, 개별문자에 대해서는 98% 의 추출율을 얻어서 본 연구가 문자영역과 개별문자 추출에 대하여 유효함을 보였다.

Abstract

In this paper, a method to segment each character and extract figure from Korean documents is proposed.

At first, each character string is extracted by means of iterative horizontal propagation, shrink algorithm and run-length algorithm. Individual character region is extracted by iterative horizontal and vertical manipulation. Next, characters of right pitch are searched. Each character is segmented by the position information.

Overlapped character is segmented on the ground of the width of already extracted character. The rest are extracted as special characters of half pitch.

Using 9 data input in the form of 840 X 600 from Korean monthly magazine, experiment was simulated. Extraction rate of characte is 100%, and that of individual character is 98%. Judging from these results, efficiency on extracting character region and segmenting individual character is proved.

*正會員, 光云大學校 電子計算機工學科
(Dept. of Comp. Eng., Kwangwoon Univ.)

接受日字 : 1988年 3月 11日
(※ 이 논문은 1987년도 문교부 일반과제 학술연구조성비에 의하여 연구되었음.)

Ⅰ. 서 론

최근 우리 주위의 사회가 급속히 발전하고 다양화되어감에 따라 여기에 대처하기 위한 방향으로 컴퓨터가 도입되고 정보를 다루는 분야의 발전도 빠르게

이루어 지고 있다. 이 때문에 컴퓨터의 사용이 확대됨은 물론 컴퓨터와 인간과의 맨머쉰(man-machine) 인터페이스가 사회적으로 주목을 끌고 있으며 그 중 하나가 문자이다. 컴퓨터에 문자의 입력은 보통 키보드를 사용하는데, 이 수단은 앞으로 급속한 정보량의 증가를 따라 갈 수 없으므로 인간의 수작업이 아닌 컴퓨터 스스로의 문자 추출장치가 절실히 요청된다. 또한, 전화회선 및 facimile을 이용한 통신망의 다양한 서비스와 전자우편이나 문자, 도형의 혼합통신이 이루워지고 있어 컴퓨터에 의한 자동화가 앞으로 필수적인 점을 감안한다면 이미 많이 연구되어진 인쇄체 및 필기체의 인식에 앞서 인식을 하기위한 전처리의 한 과정으로서 문서에서의 개별문자의 추출은 연구해야 할 중요한 과제라 생각된다.

방대한 양의 데이타처리를 필요로 하는 한국어 추출에 관한 연구는 저조한 상태이고, 비교적 문자의 특징인 획(stork) 이나 굴곡이 비슷한 일본어 문서나 문자의 형태가 간단한 영어 문서에서의 문자추출에 관한 연구는 많이 진전되었다. 문자 영역의 추출에 관한 연구는 1979년 Ito Sakatani와 Takai[1] 는 문서상에서 문자 영역을 분할하는 병렬 알고리듬에 관해 연구하여 일본어 문서에서 문자추출의 기반이 되었고 1980년 Scherl, Whal, 그리고 Fuchsbeger[2] 는 인쇄물에서 문장과 그림 영역에 대하여 자동분할에 관한 자동분리법을 발표하였다. 또한, Whal, Wong, 그리고 Caseg[3] 는 혼합된 문서에서 text를 제외한 block과 text를 추출하는 연구를 하였다. 하지만 이러한 연구는 문자의 형태가 간단한 영어 문서에 관하여 연구된 것이 많고 비교적 한국어와 비슷한 구조를 가지고 있는 일본어에 대해서는 많은 연구가 있었다. 1983년 Teruo AKIYAMA와 Isao MASUDA[4] 는 문서지면을 구성하는 문자열등의 요소를, 서식에 관한 정보등을 이용하지 않고 추출하는 방법으로 비교적 문자의 틀이 일정한 신문의 사설란을 대상으로 99%의 추출율을 얻었다. 또한1984년 Osamu NAKAMURA 등[5] 은 그림과 문자가 혼합되어 있는 문서에서 횡방향과 종방향에 대한 확대, 축소 조작과 run-length 분포를 이용하여 100%의 추출율을 얻었고, 1984년 Teruo AKIYAMA등[6] 은 비접촉 문자 우선 추출에 의한 추출을 행하여 피치가 일정하지 않은 문자등을 추정하여 결정한 뒤 반피치 문자까지 추출하는 방법에 대한 연구를 행하였다. 그러나 한국어 문서에서 문자 추출에 대한 연구는 아직까지 상당히 미비한 상태이기 때문에 발전하는 컴퓨터 분야에 부흥하지 못하고 있고 통신분야, 그리고 기계번역 분야등 이용될 수 있는 많은 응용분야는 자연언어에서 문자처리가 강하게 요구되고 있기 때문에 한국어 문서에서 문자추출에 대한 연구가 다양하게 이루어져야 된다고 본다.

본 논문에서는 논문[5] 에서 이용된 종방향과 횡방향에 대한 확대, 축소의 반복조작과 run-length 분포를 고려한 방법을 이용하여 개별문자 등을 추출하는 방법을 한국어 문서의 개별문자 추출에 적용하여 우수함을 입증하였고, 그러나 한글이 가지고 있는 특수성으로 인하여 분리는 되었으나 융합해야 하는 〔가〕나 〔나〕자와 같은 문자를 문자폭의 run-length를 이용하여 융합처리를 행한후 개별문자를 추출하므로서 인식을 위한 전처리를 하였다. 이렇게 추출된 개별문자는 인식 및 통신 시스템의 입력부나 기계번역 시스템, 그리고 다가오는 인공지능 시스템의 자연언어 처리 분야에 유용하게 쓰여질 것이다.

Ⅱ. 한국어 문서에 대한 고찰

실험 대상으로는 표 1에 표시하는 7종의 데이타를 사용하였으며 문서중에 포함되어진 문자열 중의 개별문자 추출을 목적으로 하였다. 한국어 문서화상에서 문자열 중의 개별문자를 추출 할 때는

1) 복수의 부분도형이 하나의 문자를 표시하는 것.
2) 문자가 반드시 횡으로 병행되어 있지 않는 것.
3) 문자가 접촉하고 있는 경우가 있는 것에 유의하지 않으면 않된다.

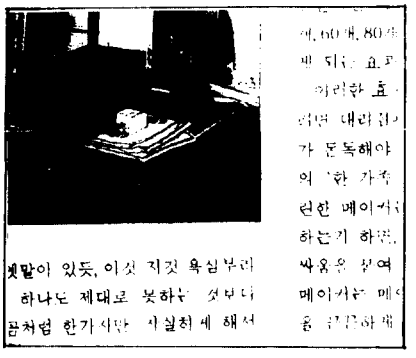
여기에서 분리문자는 1도형 1문자라고 하는 전제조건을 갖는 것이 아니고 횡서 문서화상에서 〔가〕나 〔나〕등의 분리된 문자를 의미한다. 실험에 쓰인 7종의 데이타 중 일반적으로 많이 쓰이는 도형과 표영역에 대하여 예를 들었다.

표 1. 문서의 구성요소
Table 1. Component of document.

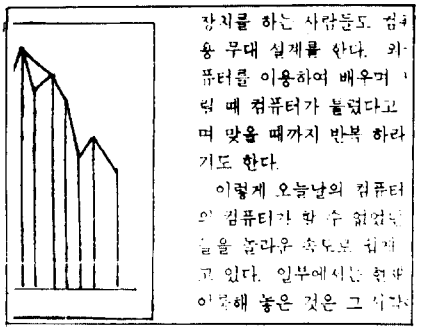
지 면		
이 치 영 역		gray-level
문 장 영 역	도 형 영 역	사 진
문 자 열 단 어 개 별 문 자	도 형 표 서 명	

문서화상은 여러가지 방식에 따라 전기적 신호로 변환될 수 있는데 이것을 처리하기 위해서는 양자화(quantization) 되어야 한다. 본 논문에서는 입력장치로서 image scanner를 사용하였고 240 dpi(dot per

inch) 의 밀도에서 840×600의 해상도로 2치 양자화 되어 입력하였다. 입력된 화상은 문자의 절단이나, hole 같은 잡음을 제거하기 위하여 2치 평활화 하였다. 그림 1에 그림영역과 표영역이 포함된 양자화된 화상에 대한 예를 보였다.



(a) 그림영역이 포함된 화상



(b) 표영역이 포함된 화상

그림 1. 양자화된 입력화상
Fig. 1. Input data of quantized image.

본 논문에서는 개별문자를 추출하는데 있어 한국어 문서에 대하여 문자열의 성질을 이용하였다. 한국어에 대한 문자열의 성질을 조사하려면 대상이 어느 것이냐에 따라 많이 틀리는데 본 논문에서는 인쇄체 활자로된 한국어 주간잡지를 예로서 설명하였다.

지면을 구성하는 한국어 문서에서 문자영역은 아래에 나타난 성질을 가지고 있다.

1) 문자영역은 문자열에 의해 구성되고 문자열의 간격은 대체로 일정하다.
2) 문자열은 단어에 의해 구성되고 문자열의 간격은 대체로 일정하다.
3) 단어는 문자로 구성되고 거의 일정간격으로 나란하게 있다.
4) 문자와 문자의 간격은 단어와 단어의 간격에 비해 작다.
5) 통상의 한국어 문자는 폭과 길이가 거의 같은 비율로 구성되어 있다.

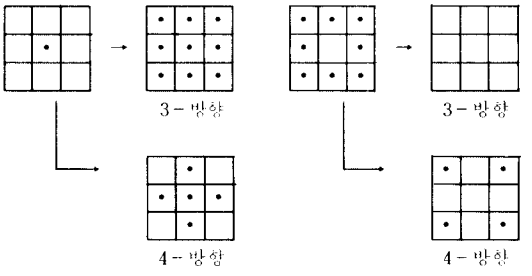
이상의 성질을 이용하여 지면을 문장영역과 도형영역으로 분리한 다음 문장영역으로부터 문자열의 추출과 개별문자의 추출을 차례로 행하였다.

Ⅲ. 문자영역과 도형영역의 추출

1. 확대, 축소조작과 Run-Length 의 원리

본 논문에서는 확대와 축소조작을 통하여 문자열을 추출하였다. 문서상의 그림이나 문자를 도형F라 가정했을때 확대와 축소의 조작은 도형 F를 융합시키는 조작을 행한다. 일반적으로도 도형F에 대한 확대와 축소는 주목점 f(i,j) 의 8근방에 대하여 그림2와 같이 기본적으로 행한다.

$$\text{단, } f(i,j) = \begin{cases} 1 : f(i,j) \in F \\ 0 : f(i,j) \notin F \end{cases}$$

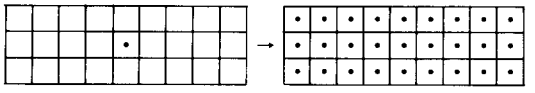


(a) 확 대　　(b) 축 소

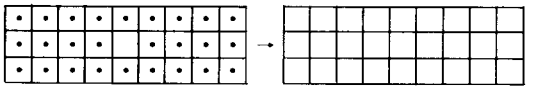
그림 2. 확대와 축소조작의 기본원리
Fig. 2. Basic theory of propagation and shrink manipulation.

본 논문에서는 위의 확대와 축소조작의 기본 원리를 이용하여서 이와 같은 조작을 반복하여 행하였다. 그 반복수를 t라 하면 t회 확대 조작을 F^t, t회 축소조작을 F^{-t}라 표시한다. 한편 이들의 조작을 조합하여 실행하는 경우, 예를 들면 $(F^t)^{-t}$로 표시한다. 이 조합에 의해 복수의 도형을 융합하는 일도 가능하며

어떤 경우에 대해서는 그 도형 융합 조작이 유효하게 되는 경우도 있다. 이상의 조작은 종방향과 횡방향의 거리비가 1 : 1 의 비율로 행하여 졌지만 본 논문에서는 종방향과 횡방향의 거리를 변화시키면서 행하였다. 그림 3에 방향성을 동반하여 종방향과 횡방향의 거리비를 변화시키면서 확대와 축소조작 한 예를 보였다.



(a) 확대조작



(b) 축소조작

그림 3. 방향성에 의한 도형의 확대와 축소
Fig. 3. Propagation and shrink with a direction.

다시 말하면 도형 F 를 1회 조작으로 횡방향의 거리비를 n 화소 종방향의 거리비를 m 화소로 확대 또는 축소하는 경우 $({}_mF_n)^t$ 또는 $({}_mF_n)^{-t}$ 로 하여(그림 4(a) 의 경우 $({}_1F_4)^1$ 로 표시한다) 이 조작을 t 회 반복 실시하는 경우

$$({}_mF_n)^{\pm t} \quad \begin{cases} +t : t\text{회 확대조작} \\ -t : t\text{회 축소조작} \end{cases}$$

으로 표시한다.

또 이 조작은 t 회 확대후, t 회 축소 조작을 반복 실시하는 경우를 예로들면

$\{({}_mF_n)^t\}^{-t}$ 단, t는 정수

로 표시한다.

이와 같은 확대와 축소 조작의 예를 순서적으로 실행한 예를 그림 4 에 보였다.

본 논문에서는 문자열을 추출하기 위하여 한 문자가 연결되어 있으면 몇개가 연속해서 있는가를 고려하는 run-length 방법을 이용하였다. 방향성을 가진 문서화상의 경우에는 그 방향에 따르는 화소의 갯수가 어느 정도 연결되어 있다. 이렇게 동일 화소가 연결되어 있는 화소의 수를 run-length 라고 한다. 그림 5 에 2치화된 문서상의 run-length 를 나타내었다.

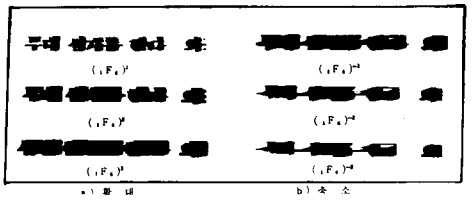


그림 4. 화대와 축소 조작의 예
Fig. 4. The example of propagation and shrink operation.

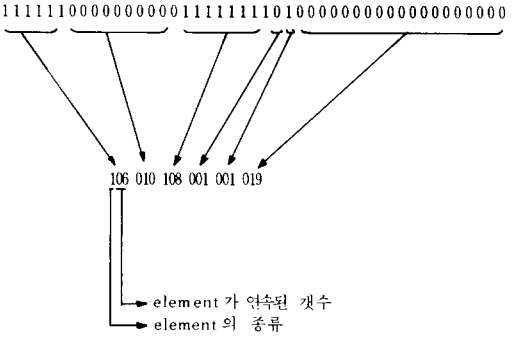


그림 5. Run-Length의예
Fig. 5. The example Run-length.

2. 문자열의 추출

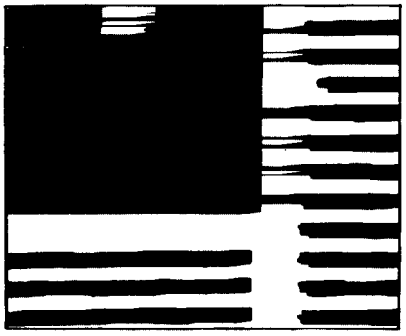
본 논문에서 문자열의 추출은 확대와 축소조작과 run-length 분포를 이용하여 아래의 순서로 행한다.

〈순서 1〉

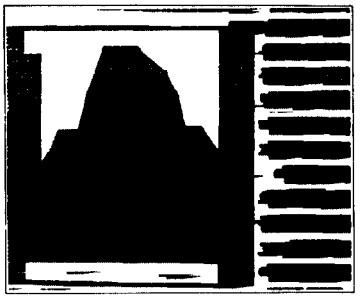
본 논문에서 문자열을 추출하기 위하여 먼저 입력지면 전체 그림 1 에 $\{({}_1F_8)^4\}^{-4}$ 의 조작을 실시한다. 여기서 m과 n의 값은 최대 50화소 정도인 것에 근거하여 m=1,n=8로 결정했다. 이 조작에 의해서 단어뿐 아니라 문자로 구성된 열이 완전히 융합되고 문장영역의 문자열이 하나의 연결영역으로 된다. 본 논문에서 이 조작은 문자열 뿐만 아니라 문서화상 내의 도형 F 에 대하여 모두 융합됨을 그림을 통해서 알 수 있다. 그림 6에 이 조작의 예를 보였는데 (a) 는 그림영역이 있는 처리이고 (b) 는 표영역이 있는 처리의 예를 나타내고 있다.

〈순서 2〉

위 순서 1 의 처리결과에 포함된 전체 연결영역에 대하여 종방향의 주사를 행한다. 여기에서 종방향의 주사란 문서화상의 위에서 아래로 run-length 를 의

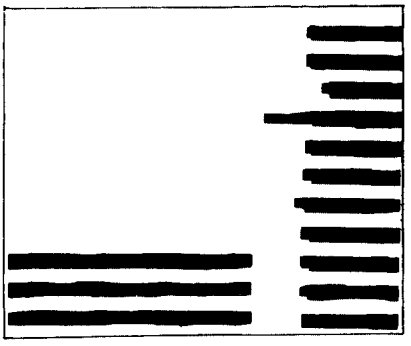


(a) 그림영역의 처리결과

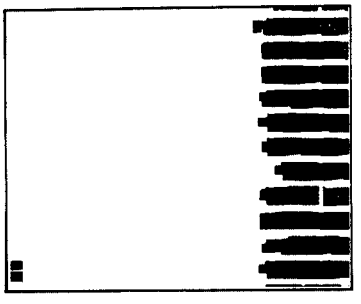


(b) 표영역의 처리결과

그림 6. $\{({}_1F_8)^4\}^{-4}$ 의 처리결과
Fig. 6. The result of operation $\{({}_1F_8)^4\}^{-4}$.



(a) 그림영역에 대한 예



(b) 표영역에 대한 예

그림 7. 추출된 문장영역의 예
Fig. 7. The example of extracted text field.

미한다. 이때 흑화소의 run-length 분포를 $H^K(i)$: k =1,2,…(단, k는 연결영역의 라벨, i는 length) 라 할때

$$\frac{\sum_{i=\alpha}^{\beta} H^K(i)}{\sum_{i=1}^{600} H^K(i)} > Th$$

를 만족하는 연결영역을 전체 문자열로 판단한다. 본 논문에서 위식을 만족하지 못하는 경우에는 run-length 값이 16에서 38사이의 run-length를 문자열로 판단하였다. 그리고 threshold값 Th는 5장의 실험 및 고찰에 근거하여 그림영역인 경우에는 Th=0.850으로 표영역인 경우에는 Th=0.625로 결정하였다. 또한 α와 β는 실험적 고찰에 의하여 16에서 38로 정하였다. 추출한 예를 그림 7의 (a)와 (b)에 나타내었다.

3. 도형영역의 추출

본 논문에서는 도형영역추출에 있어서 문장영역 이외의 부분을 도형영역으로 판단하였다. 따라서 입력화상에 문장영역과 같은 성질의 것이 아닌 모든 부분에 대하여 도형영역으로 취급한다. 도형 영역만 추출되어진 예를 그림 8에 나타내었다.

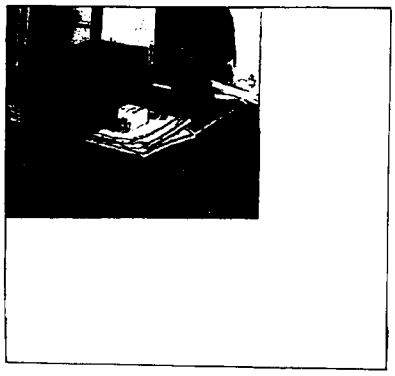
4. 개별문자의 추출

본 논문에서는 문자열의 추출에 의하여 개별문자를 추출한다. 개별문자를 추출하기 위해서는 원래 문서화상 중 그림이나 표영역이 빠진 부분인 문자열만이 추출 되어진 화상만을 입력대상으로 한다. 그림 9는 그림과 표영역이 빠진 데이타를 보였다. 본 논문에서의 개별문자 추출의 순서는 아래에 나타낸 순서로 행한다.

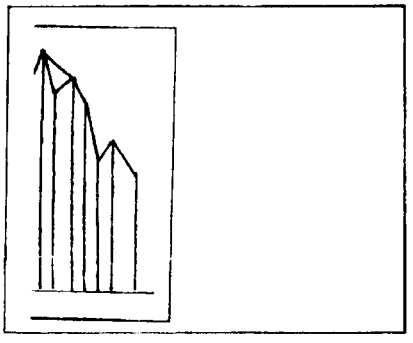
〈순서 1〉

문장영역 전체 그림 9에 대하여 $\{({}_1F_4)^3\}^{-3}$의 조작을 실시하여 그 결과를 W_w라 한다. 그림10에 처리결과를 나타내었다. 여기에서 문자열이 모두 융합되어진 것을 볼 수 있다.

이 조작은 개별문자들의 융합을 목적으로 한다. 여기에서 m, n은 실험적 고찰을 기초로 2치화의 Threshold에 의한 문자 간격의 변동을 고려해서 m=1, n =4로 정하였다.

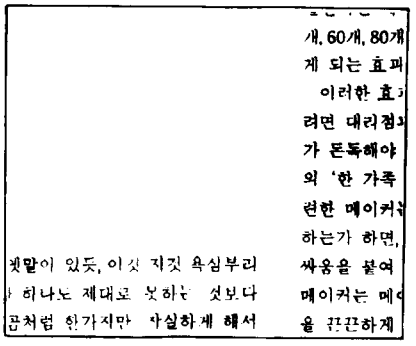


(a) 그림 영역이 추출된 예

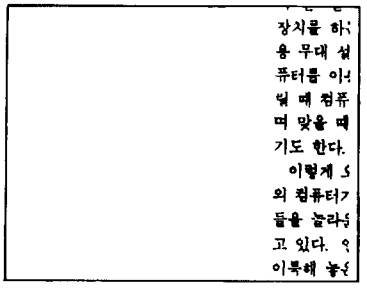


(b) 표 영역이 추출된 예

그림 8. 도형영역의 추출 예
Fig. 8. The example of extracted graphic field.



(a) 그림 영역이 빠진 원래 데이타



(b) 표 영역이 빠진 데이타

그림 9. 도형영역이 빠진 원래 데이타
Fig. 9. The original data expected graphic field.

〈순서 2〉

문장영역 전체에 대하여 $(_4F_0)^3$의 조작을 실시하고 이 결과를 W_c라 한다. 이 조작에 의해서 횡으로 나란한 문자간의 융합을 방지하는 것과 함께 종으로 분리된 문자, 예를들면 〔느〕와 〔으〕 같은 문자를 종으로 연결하는 일이 가능하게 된다. 그림11에 처리된 결과를 보였다.

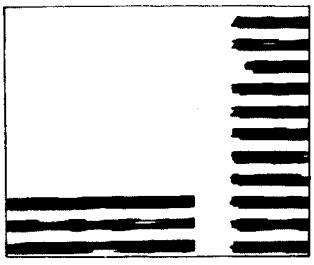
〈순서 3〉

위의 순서 1과 순서 2에서 처리된 결과인 W_w 와 W_c에 대하여 $W_w \cap W_c$인 공통부분을 추출한다. 이 공통부분의 추출은 한국어 문서에서 개별문자의 영역을 분리되어 있는 정도에 따라 비교적 충실하게 추출되는데 분리되어 있는 문자의 분리된 형태로 추출된다. 여기에서 공통부분이란 횡방향에 의한 융합조작과 종방향에 의한 융합조작의 공통부분이므로 한국어의 개별문자 특성이 일정한 비율의 크기 즉 직사각형의 모형을 갖고 있으므로 문자가 있는 부분은 충실하게 추출되었다. 그림12에 추출 되어진 공통부분을 나타내었다.

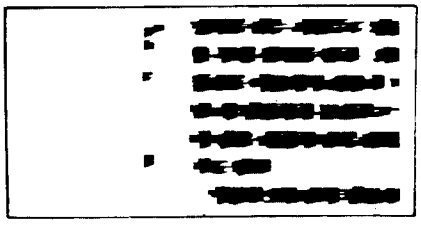
5. 개별문자의 윤곽선 처리와 문자영역의 모형화

상기의 처리결과에 의하여 추출되어진 공통 부분에서는 몇가지의 문제점이 발생한다. 첫째 확대조작에 의해서 추출 되어진 개별문자영역은 실제 문자가 있는 영역보다 큰 영역으로 되고, 둘째 한국어의 특징인 획(strok) 이 자유롭게 있어 개별문자 추출부분이 접속하는 경우가 생긴다. 그리고 입력대상의 문자가 일반적으로 촘촘히 들어선 형태이므로 해상도가 낮은 입력장치로 입력을 받으면 문자간의 접촉이 일어난다. 본 논문에서는 이러한 문제점을 해결하기 위해서 윤곽선을 추출하여 개별문자 영역의 모형화를 행한다. 여기에서 윤곽선 추출 알고리듬은 화상처리 수법중 3×3 마스크를 사용하여 윤곽선을 찾는 방법을 이용하였다. 간단히 알고리듬을 기술하면 Pn−1 점의 좌표를 〔I(N−1), J(N−1)〕, Pn 점의 좌표를 〔I(N), J(N)〕 이라고 하면 Pn−1점이 Pn점에 대하여 어느 방향이었는가는 다음식으로 구해진다.

$$IX = I(N-1) - I(N)$$
$$IY = J(N-1) - J(N)$$



(a) 그림 영역의 처리결과



(b) 표 영역의 처리결과

그림10. $\{(_1F_4)^3\}^{-3}$ 조작의 처리결과
Fig. 10. The result of operation $\{(_1F_4)^3\}^{-3}$.

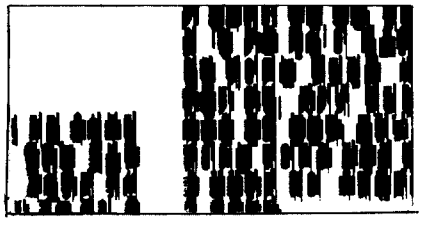


그림11. $(_4F_0)^3$ 의 처리결과
Fig. 11. The result of operation $(_4F_0)^3$.

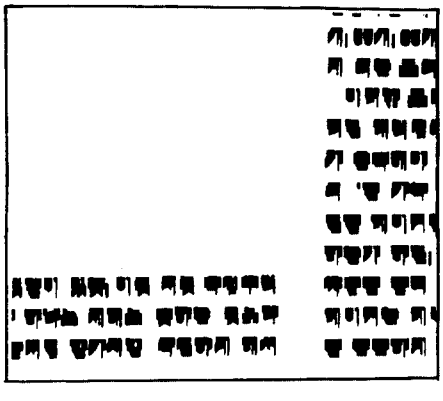


그림12. 공통부분의 추출결과
Fig. 12. The result extracted individual symbol field.

만일 그림 13에 표시하는 상태로 되어 있으면 IX $=-1$, IY $=-1$로 되고 왼쪽 위에 Pn-1이 있는 것을 알 수 있다. 이때 시계반대방향 즉 표2에 표시하는 순서로 흑 영역인가 백 영역인가를 조사하면 된다.

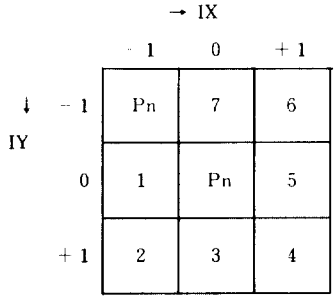


그림13. 3×3 마스크의 표시
Fig. 13. Representation of 3×3 mask.

표 2. 마스크 이동 순서표
Table 2. Order table of mask moving.

순 서	IX	IY
시 점	−1	−1
1	−1	0
2	−1	1
3	0	1
4	1	1
5	1	0
6	1	−1
7	0	−1

이것은 IX와 IY 의차와 합을 구하고 그 값이 0 이나, 또는 -1이면 각각 -1,1로의 치환에 의하여 그대로 다음의 탐색점 방향이 차례로 구해진다.

$$IX = IX + IY$$
$$IY = IY - IX$$

모형화는 윤곽선을 따라 가면서 X, Y 좌표의 최대 최소치를 구하여 한국어 문자의 형태가 사각형 형태의 구조인 점을 고려하여 사각형의 형태로 모형화를 행하였다. 그림 14에 윤곽선 추출 예를 보였고 그림 15에 모형화한 예를 보였다.

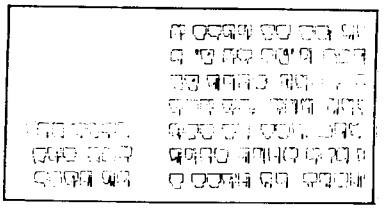


그림14. 윤곽선 처리한 결과
Fig. 14. The result of edge detected.

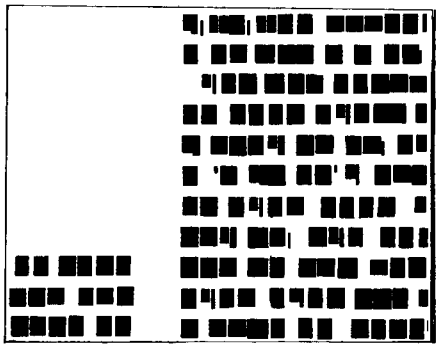


그림15. 모형화 처리한 결과
Fig. 15. The result of symbol segmentation.

6. 융합 되어진 문자의 절단

본 논문에서는 위의 5 절에서 기술한 문제점을 해결하기 위하여 모형된 데이타를 이용하였다. 먼저 여백 제거에 대하여는 추출된 데이타가 문자인식의 기법으로 인식되는 데는 아무런 상관이 없으므로 처리하지 않았고 융합되어진 문자의 절단에 대하여는 한국어 문서의 예비 조사결과 240 dpi로 입력 했을때 한문자의 폭이 거의 20에서 25화소 정도이고 두 문자가 결합 했을때의 횡방향에 대한 run-length 는 40 화소 이상이 되었다. 그래서 본 연구에서 문자의 절단은 40화소 이상이 되는 것만 고려하였다. 그림 16에 융합되어진 문자의 절단에 대한 예를 보였다.

Ⅳ. 정피치 문자 우선 추출에 의한 개별문자의 추출

본 논문에서는 추출의 마지막 단계로서 정피치 문자 우선 추출에 의한 방법으로 문자를 추출하였다. 본 방법은 먼저 횡 방향의 run-length 분포가 가장 많은 문자부터 추출하고 다음으로 분리된 문자를 융합하여 추출한 다음 특수문자와 같은 반피치 문자를 추출하는 방법이다. 그림 17의 (I) 는 문자열을 구성

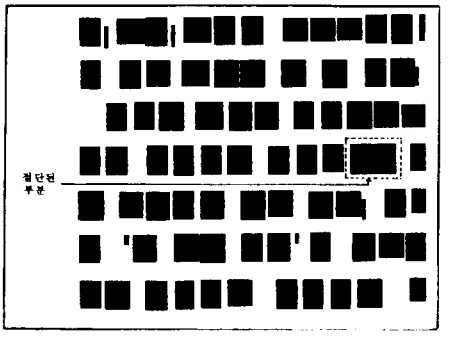


그림16. 융합되어진 문자의 절단
Fig. 16. The truncate result of overlapping field.

하는 개별문자를 표시하고 있다. 다음에 문자열을 수평 방향으로 주사하여 문자열에서 가장 많이 나타난 피치 a를 구한다. (Ⅱ) 에서 a가 구해진 부분을 나타내었다. 그래서 a를 정피치라 하고 (A) 에 나타난 것과 같이 정피치 문자를 우선추출한다. 여기에서 * 표시는 새로 추출 되어지는 문자이다. (B) 에서는 정피치 문자와 피치가 같은 정도의 영역을 정한 다음 추출한다. (C) 에서는 문자열 중의 반피치 문자를 추출한 예이다.

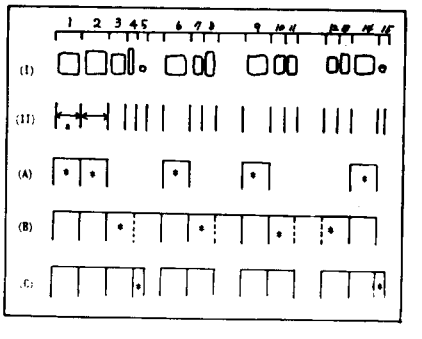


그림17. 기본 추출처리
Fig. 17. Basic extraction operation.

본 처리는 한국어와 같이 한 문자가 분리되어 있는 분리 문자나 반피치문자 그리고 특수문자를 추출하는데 매우 좋은 결과를 나타내지만 한국어의 문자 특성상 분리되어진 초성, 중성, 종성은 결합하여야 한 문자가 되므로 실험적 고찰에 의하여 얻은 한 문자의 횡방향에 대한 run-length가 20에서 25인 점과 주위의 분포와 관계를 고려하여 문자를 융합하였다.

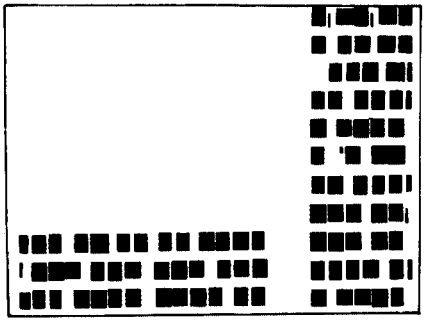


그림18. 추출되어진 결과
Fig. 18. The result of extracted sysbol.

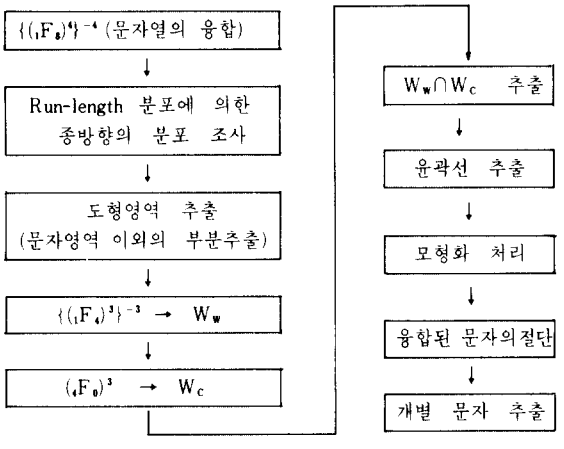
그림 18에 이상의 처리 결과인 최종적으로 추출 되어진 문자를 나타내었다.

V. 실험 및 고찰

1. 전체 추출시스템의 흐름도

본 연구는 지금까지의 처리에 의하여 문서상에서 개별문자를 추출하였다. 처리의 순서를 간략히 표시하기 위하여 표3에 추출시스템에 대한 전체 흐름도를 표시하였다.

표 3. 전체 흐름도
Table 3. Whole flow-chart.



2. 추출 시스템

실험에 사용된 데이타는 A4 크기의 문서에서 1/6 정도의 한 부분을 image scanner에 입력시켜 양자화를 행하였다. 이 시스템의 주사방식은 840×600이고 각 비트당 gray-level은 2이다. 이와같이 2치화된 도면은 RS-232C inferface 를 통해 IBM/AT와 호환성이 있는 마이크로 컴퓨터 시스템에 저장했으며 모든 처리는 XENIX-V OS와 MS-DOS 상에서 C와 ASSEMBLY 언어를 사용하여 행하였다. 그림 19에 추출 시스템을 보였다.

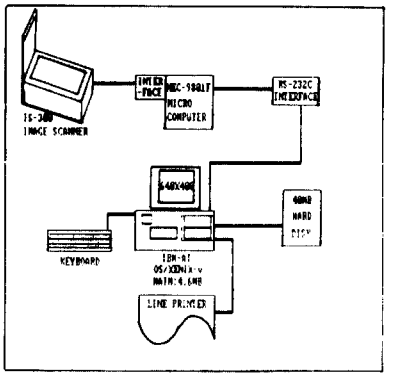


그림19. 추출 시스템
Fig. 19. Extraction system.

3. Threshold 결정

여기에서 문자열을 추출하는데 있어 Th에 관해 검토한다. 실제적으로 조사결과 그림 영역에서 문자영역의 run-length 비율이 문서가 840×600의 해상도인 점을 고려하면 약 11개 정도의 문자열이 입력되므로 그림영역에 대하여 종방향 각각의 run-length 를 조사하면 그림 영역이 있는 부분에 대하여 종방향의 run-length 중 문자열 즉, run-length가 16에서 38이 가지는 비율은 80%에서 90%정도였다. 그 이유는 일반적으로 그림에 있어서 확대축소 조작이 그림 전체의 영역을 차지하고 있기 때문이다. 한편, 그림 영역이 가는선인 표 같으면 문자영역이 차지하는 비율은 60%에서 65%가 일반적 이었다. 본 논문은 이러한 조사를 근거로 그림 영역이 사진같은 그림에 대하여는 th=0.850으로 하였고 표 영역에 대하여는 0.625로 하였다.

4. 추출 능력

추출 능력을 검토하는데 있어서 대상이 정확하게 식별되고 제대로 추출되는가를 추출율로 평가 하였다. 여기에서 대상으로는 최종적으로 추출되어진 개별문자이다. 여기서 추출율이 98% 였는데 그 주된 원인 단어의 전후에 위치하는 콤머등의 반피치 문자가 개별문자와 융합되어 이 과정에서 같이 추출되어

버린 점이다. 그러나 이러한 문제는 글자의 인식처리[7],[8],[9] 에 의해서 정확히 추출될 수 있다고 생각된다.

본 연구의 대상으로 추출되는 문자가 각각의 개별문자이므로 해당하는 개별문자의 추출도가 어떻게 되느냐에 따라서 알고리듬의 좋고 나쁨이 결정된다. 그리고 입력대상으로 하는 문자가 비교적 복잡한 한글 패턴을 대상으로 하므로 본 논문에서 추출알고리듬의 주종이 되는 확대와 축소조작에 대한 문자의 융합정도에 따라서 추출도에도 상당한 영향을 미친다.

보통 융합도를 $\{({}_mF_n)^t\}^{-t}$로 나타나는데 t 값에 따라 전체 추출시스템에 미치는 영향을 보면, 먼저t값의 증가는 한개의 흑화소에 대하여 확장을 해나가는 과정이므로 t 값이 너무 커버리면 문서에 있는 모든 문자가 융합되어 버리고 너무 작으면 문서의 문자열에 대한 융합이 되지 않는다. 그러므로 각국의 글자 형태가 틀리고 문자의 크기도 틀리므로 해당하는 문자가 어느 정도 융합이 되어야 하는가가 문제로 된다. 이 문제는 실험에 의하여 t의 값을 추정할 수 있는데 한국어에 대하여는 당하는 문자열의 백 부분 영역없이 대체적으로 완전히 융합되는 범위내에서 결정하였고 종방향의 융합에서는 윗 부분 문자열과 아래부분 문자열이 융합되지 않는 범위내에서 t값을 결정하였다.

본 연구의 추출시간을 고려하면 데이타의 모형화까지 대부분이 확대와 축소의 반복조작이기 때 문에 소요시간은 48분 정도가 소요되었다. 각 부분별로 시간의 진행과정을 살펴보기 위하여 대상으로 하는 문서데이타를 살펴보면 640×400의 Bi-level 영상으로 용량은 약 512K 정도가 된다. 이 방대한 양의 데데이타를 처리하기 위하여는 시간이 많이 소요된다. 각 부분에서 소요되는 시간을 비교하기 위하여 표 4에 소요시간량을 나타내었다.

본 논문에서는 일반적인 문서가 그림과 표가 주종을 이루기 때문에 이 두 경우를 고려해서 충실한 결과를 얻을 수 있었고 또한 한국어와 복잡도가 비슷한 한문이나 문자가 하나로 되어있는 영문문서에 대해서도 좋은 결과를 얻을 수 있으리라 사료된다.

그리고 본 논문에서 처리된 형태가 대부분 병렬처리이기 때문에 하드웨어로 치환하면 실시간의 처리가 가능 해질 것이다.

표 4. 소요시간의 비교표
Table 4. Comparison table of exhausted time.

단 계	수 행 시 간
문자열의 융합	12 분
Run-length 분포에 의한 횡방향의 분포조사	3 분
도형영역 추출	2 분
$\{({}_1F_4)^3\}^{-3} \rightarrow W_w$	9 분
$({}_4F_0)^3 \rightarrow W_c$	6 분
$W_w \cap W_c$ 추출	1 분 30 초
윤곽선 추출	2 분
모형화 처리	7 분 30 초
융합된 문자의 절단	2 분
개별 문자 추출	3 분
총	48 분

Ⅵ. 결 론

본 논문은 다른 언어에 비하여 자형의 변형이 다양한 한국어 문서에 대하여 개별문자를 추출한 결과를 보였다. 문자, 도형 혼합통신의 효율적 방법으로 주목되고 있는 인식 통신 시스템을 실현하는데 요구되는 기술인 문장영역과 그림영역의 분리, 개별문자의 추출에 관해 본 논문에서는 도형을 포함한 한국어 문서를 대상으로 하여 입력지면에 대해 확대축소조작에 의한 도형의 융합처리와 run-length분포를 이용한 방법을 한국어에 적용한 것과 정피치를 추정한 다음 그 정보를 기초로 우선 추출하는 알고리듬을 한국어 문자에 맞게 변형시켜 적용한 결과를 보였다. 또한 점차 지능화 시대로 가는 컴퓨터 과학분야의 자연언어 처리나 기계 번역 그리고 미래의 통신인 인식통신 시스템의 기초가 되는 입력부분의 중요한 역활을 할 것이다. 이후의 연구과제로는 그림과 문자열을 추출하는데 필요한 Threshold는 약간의 변화에도 추출율을 크게 변화할 수 있으므로 Threshold의 자동 결정법이 중요한 연구과제로 연구되어져 할 것이며 한자, 영어, 특수문자등이 포함되어 있는 다양한 문서도 연구대상으로 남아있다.

參 考 文 獻

[1] Ito Sakatani and Takei, "parallel segmentation algorithm of document," ICCS, pp. 519-523, 1979.

[2] Scherl, Whal and Fuchsberger: "Automatic separation of text, graphic and picture segments in printed material," Pattern recognition in practice, North-Holland, pp. 213-221, 1980.

[3] Whal, Wong and Casey, "Block Segmentation and Text Extraction in Mixed Text,"

Computer Graphics and Image Processing, vol. 20, pp. 375-390, 1982.

[4] Teruo NAKAYAMA, and Isao MASUDA, "A Segmentation method for Document Image without the Knowledge of Document Formats," 일본전자통신학회 논문지, vol. J66-D, no. 1, 1983.

[5] Osamu NAKAMURA, Norivosh OKAMOTO, Tsuyoshi NIWATA and Toshi MINAMI, "A Symbol Segmentation Algorithm from Free Format Document," 일본 전자통신학회 논문지, vol. J66-D, no. 4, 1983.

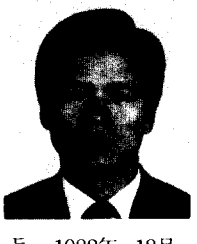
[6] Teruo AKIYAMA, Seiichrio NAITO and Isao MASUDA, "A Metho' of Character Extraction from Printed Documents Guided by Positions of Non-Overlapping Characters," 일본 전자통신학회 논문지, vol. J67-D, no. 10, 1984.

[7] J.K. Lee, "Korea Character Display Variable Combination and Its Recognition by Decomposition Method," Ph. D. dissertation in Keio University, Japan, 1972.

[8] 이주근, 남궁재찬, 김영건, "한글 pattern 에서 subpattern 분리와 인식에 관한 연구," 전자공학회 논문지, vol. 18, no. 3, 1983.

[9] 남궁재찬, "Index-Window 알고리듬에 의한 한글 pattern의 부분 분리와 인식에 관한 연구," 인하대학교 박사학위 논문, 1982. *

著 者 紹 介



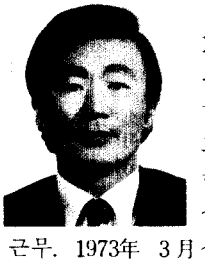
南 宮 在 贊(正會員)

1947年 6月 13日生. 1970年 2月 인하대학교 전기공학과 공학사 학위 취득. 1976年 8月 인하대학교 대학원 전자공학과 공학석사학위 취득. 1982年 2月 인하대학교 대학원 전자공학과 공학박사 학위취득. 1982年 12月～1984年 1月 일본 동북대학교 객원 연구원. 1979年 3月～현재 광운대학교 전자계산기공학과 부교수. 주관심분야는 Pattern recognition, Computer Vision, 한글인식 등임.

●

柳 煌 彬(正會員)

1975年 2月 인하대학교 전자공학과 졸업. 1977年 8月 연세대학교 산업대학원 전기·전자공학과 졸업 1987年 9月 경희대학교 대학원 전자공학과 박사과정 수료. 1981年～현재 광운대학교 이과대학 전자계산학과 부교수. 주관심분야는 데이타통신, 영상처리 등임.



南 宮 璉(正會員)

1947年 7月 17日生. 1971年 2月 광운대학교 응용전자공학과 학사학위 취득. 1988年 2月 광운대학교 산업정보대학원 전자계산기공학과 석사학위 취득. 1971年 4月～1973年 2月 춘천공업고등학교 근무. 1973年 3月～1976年 7月 영월공업고등학교 근무. 1977年 5月～1977年 8月 구미전자공업고등학교 근무. 1977年 9月～현재 국립구미전자공업고등학교 통신설비과 과장 및 교사 재직. 주관심분야는 데이타통신 및 영상처리 등임.