論文 95-2-2-1

인쇄체 한글 문서에서의 동적 문자추출에 관한 연구

洪 性 鵬* 正會員 張 喜 敦* 正會員 南宮 在贊*

A Study on Dynamic Extraction of Character from Printed Hangeul documents

Sung Boong Hong*, Hee Done Jang*, Jae Chan Namkung* Regular Members

要 約

본 논문은 인쇄체 한글 문서내에서 문자의 크기나 서체에 무관하게 개별문자를 추출하기 위해 형상 비율이나 피치 비율과 같은 문자의 이미지 정보와 위치 정보를 이용하여 조합문자의 합성을 동적으로 행하고 문자의 구조 정보를 이용하여 접촉 문자사이의 접촉 형태를 파악할 수 있는 투영법을 제안한다.

본 논문에서 제안하는 방법은 문자 블럭의 비례적인 정보를 사용함으로써 크기가 다른 문자가 혼재한 경우에도 합성과 분리 처리가 가능하고 분리 처리시에도 문자의 접촉 위치를 알 수 있는 구조 정보를 사용하여 기호 문자와 접촉해 있는 경우의 문자 추출이 정확하게 이루어질 수 있다. 논문지, 잡지, 교과서 등의 일반 문서를 대상으로 실험한 결과, 99%이상의 문자 추출율을 얻어 본 논문의 유효성을 확인하였다.

ABSTRACT

This paper proposes a dynamic character extraction method for the printed Hangeul documents.

In the proposed method, the isolating compound characters are dynamically merged using the position information and image information such as aspect ratio and pitch ratio for multi-font and multi-size. And touching characters are segmented using the projectional segmentation method. This method is able to know the phase of touching between characters.

The multi-size characters are processed using proportional information of character block such as aspect ratio and pitch ratio. Also, when the touching character being touch with symbolic characters is segmented, the character extraction is applied correctly because the use of structural information is able to know position of touching. Experiments were carried out on document taken from papers, magazines and books. As a result of experiment, we obtain 99% over rate of character extraction.

* 광운대학교 전자계산기공학과
Dept. of Computer Engineering Kwangwoon University
論文番號 : 9440
接受日字 : 1995年 2月 14日

Ⅰ. 서 론

인간에 의해 컴퓨터가 개발된 후, 그 응용 분야는 날로 다양해지고 있다. 그 중에서 정보를 저장하는 수단으로써 컴퓨터가 차지하는 역할은 두드러진다고 할 수 있다. 더구나 이전의 정보 저장수단이었던 문서 (document)를 자동으로 컴퓨터에 입력하여 인식하는 문서 인식 시스템 (document recognition system)은 그 필요성이 크게 대두되고 있다. 이에 부응하여 외국에서는 문서 인식 시스템의 전반에 걸쳐 많은 연구가 진행되어 왔다. 이에 반해서 국내의 문서 인식 시스템에 대한 연구는 문자 인식 (character recognition)에 대한 연구에 치중하여 진행되어 왔을 뿐 문서 인식 시스템의 전처리과정 (preprocessing)에 대한 연구는 상대적으로 미비한 실정이다.

문서 인식 시스템의 전처리과정 중에서도 중요한 부분은 문자 추출 (character extraction)부분이다[1]-[7]. 실제로 개별문자의 인식도가 아무리 높다 하더라도 문자 추출의 정확도가 낮으면 전체적인 문서 인식 시스템의 성능은 크게 떨어지게 된다. 때문에 문서 영상 (document image)에서 문자를 추출하는 것이 중요한 과제가 된다.

본 논문은 인쇄체 한글 문서에서의 개별문자 추출을 다루게 되는 데, 한글 추출에 관한 기존의 연구는 단일 서체 (single font)[7]에 대한 처리만을 다루거나 문자간의 접촉이 발생하지 않는다고 가정하는 등 인쇄 문서에 대한 많은 제약을 두고 있고, 문자의 합성과 분리 처리를 다룬 연구에 있어서도 단순히 두 개의 문자만이 붙었을 경우의 분리 처리를 행하거나[5], 너무 정적(static)인 정보로 일률적인 분리[5][8]를 하는 경우와 문자 추출 처리 자체에 너무 많은 정보가 필요하게 되는 경우 등 해결되지 않은 문제점이 많았다.

본 논문에서는 기존의 한글 추출 알고리즘에서 분리 추출된 문자 블럭의 합성처리시 합성의 조건으로 평균 글자폭과 같은 일률적인 정보를 주었던 것을 한글 문자 블럭의 비례적인 구조 특성을 조사하여 크기 (size)에 무관하게 동적 (dynamic)으로 적용될 수 있는 이미지 정보와 위치 정보를 이용하여 합성하는 방법과 기존의 투영법 (projectional method)[8]으로는 추출하기 어려웠던 두 글자이상의 한글끼리 접촉되어 있는 문자의 분리, 기호 문자와 접촉되어 있는 한글의 분리, 한문과 접촉되어 있는 한글 문자를 분리 추출할 수 있는 투영법을 제안한다. 이 투영법은 한글끼리 접촉되는 경우와 기호, 한문이 한글과 접촉되는 경우의 구조 특성을 파악할 수 있는 위치 정보를 이용하여 한글을 분리 추출한다.

본 논문은 2장에서 문자 추출처리에 필요한 전처리 과정을 다루고, 3장에서 전처리 과정에서 얻어진 문자 블럭에 대해서 논하며, 4장에서 문자 블럭의 합성처리, 5장에서 문자 블럭의 분리에 대해서 논하고, 실험결과와 고찰은 6장에서 논하며, 7장에서 결론을 맺는 것으로 구성된다.

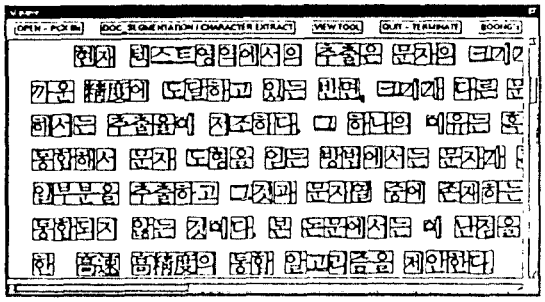
Ⅱ. 문자 블럭 추출

2.1 문자열 추출

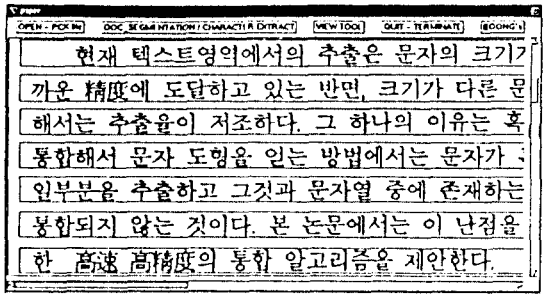
문자 블럭을 추출하기 위해서는 우선 문서 영상내의 문자 영역에서 문자열을 추출하여야 한다. 문서 영상에서 문자 영역을 추출하는 것은 문서 분할 (document segmentation)[1]-[7]과정에서 처리되고, 프로화일링 방법을 문자 영역에 적용하여 문자열을 추출한다. 즉, 1회 수평 투영 프로화일링하여 문자열을 추출한다[9].

2.2 문자 블럭 추출

문자열이 추출되면 추출된 문자열에서 문자 블럭을 추출하게 된다. 마찬가지로 프로화일링 방법을 사용하여 문자 블럭을 추출하게 되는 데 추출된 블럭은 하나의 문자로 구성될 수도 있고, 여러 개의 문자나 문자의 한부분만으로도 구성될 수도 있다. 추출 과정은 앞 절에서 얻어진 각각의 문자열에 대해 프로화일링 방법을 적용하여 문자 블럭을 얻는다. 문자 블럭을 얻기 위해서는 수직, 수평의 2회 프로화일링을 하게 되고 2회째 단계에서 임계치 (threshold)를 두어 문자 블럭이 위, 아래로 분리되는 것을 방지한다. 이것은 차후 문자 블럭 합성시 수직 합성 (vertical merging) 처리가 필요없도록 한다. 그림 1에 문자 블럭의 추출 예를 보였다.



(a) 문자열 추출



(b) 문자 블럭 추출

그림 1. 문자 블럭 추출 예
Fig. 1. An example of character block extraction

Ⅲ. 문자 블럭의 형태와 정보

3.1 문자 블럭의 형태 특성

문자 블럭은 하나의 문자로 구성되어 추출될 수도 있고 여러 개의 문자로 구성되거나 하나의 문자가 분리되어 추출될 수 있으며 블럭의 형태적인 특성은 다음과 같다.

(1) 사각형 (rectangle)의 형태를 가지고 있다.
(2) 가로폭대 세로폭의 비율 (aspect ratio)이 다양하게 나타난다.
(3) 블럭의 최외각선에는 흑화소가 존재한다.
(4) 블럭의 베이스 라인 (base line)이 일정하지 않다.

3.2 문자 블럭 정보

본 절에서는 추출된 문자 블럭에서 이용할 수 있는 문자 블럭의 일반적인 이미지 정보 (image information)와 블럭들 간의 위치 정보 (position information)를 기술한다. 여기서는 블럭 정보의 개념적인 접근을 하고 실제 처리과정에서 세부적인 정보 이용에 대해서 논하기로 하겠다.

블럭의 이미지 정보는 블럭의 가로폭 (width)과 세로폭 (height), 세로폭에 대한 가로폭의 비율 (aspect ratio), 블럭의 대각선 길이 (diagonal), 블럭이 속한 문자열의 세로폭에 대한 블럭의 세로폭 (pitch ratio)이다. 그림 2에 정보의 내용을 도식화 했고, 각 정보의 내용은 다음과 같다.

(1) 블럭의 가로폭 (width)

$$width = xend - xinit + 1$$

xinit : 블럭의 왼쪽 시작 위치값
xend : 블럭의 오른쪽 끝 위치값

(2) 블럭의 세로폭 (height)

$$height = yend - yinit + 1$$

yinit : 블럭의 위쪽 시작 위치
yend : 블럭의 아래쪽 끝 위치

(3) 블럭의 세로폭에 대한 가로폭의 비율(aspect ratio)

$$\text{aspect ratio} = \frac{width}{height}$$

(4) 블럭의 대각선 길이 (diagonal)

$$diagonal = \sqrt{width^2 + height^2}$$

(5) 문자열의 세로폭에 대한 블럭의 세로폭의 비율 (pitch ratio)

$$\text{pitch ratio} = \frac{\text{height of block}}{\text{height of text line}}$$

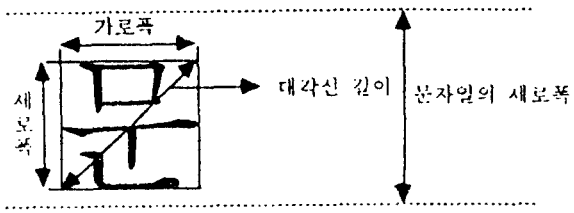


그림 2. 블럭의 이미지 정보.
Fig. 2. A image information of block.

한편 블럭들을 합성할 때 처리될 블럭을 결정하여야 한다. 이때, 이미지 정보외에 블럭들간의 위치 관계를 조사하여 처리되어야 할 블럭을 정한다. 위치 정보로는 블럭들간의 거리 (difference)와 부가적으로

블럭들의 시작점 사이의 거리 (본 논문에서는 이를 constraint difference라 정의한다)를 사용한다. 그림 3에 도식화했고, 각 정보의 내용은 다음과 같다.

(1) 블럭들간의 거리 (difference)

difference = block[i+1].xinit-block[i].xend-1

block[i+1].xinit : 다음 블럭의 왼쪽 시작 위치값

block[i].xend : 현재 블럭의 오른쪽 끝값

(2) 블럭들의 시작점 사이의 거리(constraint difference)

constraint difference $=\sqrt{xdiff^2+ydiff^2}$

xdiff = block[i+1].xinit - block[i].xinit + 1

ydiff = block[i+1].yinit - block[i].yinit + 1

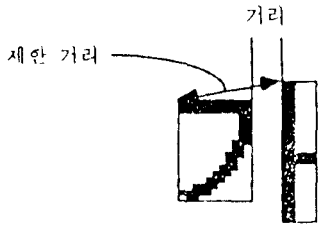


그림 3. 블럭의 위치 정보.
Fig. 3. A position information of block.

Ⅳ. 문자 블럭의 합성

문자 블럭의 합성 처리는 한글 문자 블럭에 대한 규칙과 기호 문자 블럭에 대한 규칙, 한문 문자 블럭에 대한 규칙을 문자 블럭에 적용하면서 처리한다.

4.1 한글 문자 블럭의 합성

본 절에서는 합성 처리의 대상을 좀 더 세부적으로 명시하는데 이것은 합성 처리의 규칙(rule)과 밀접한 관계를 가지고 있다. 그림 4에서 보듯이 한글 문자중 분리되어 추출되는 문자는 스캐너 입력 상태에 의한 경우를 제외하고는 1형식 문자뿐이다.

대상-① 하나의 문자가 두 개의 블럭으로 추출된 경우 (그림 4-(a)(b)(e))

대상-② 하나의 문자가 세 개의 블럭으로 추출된 경우 (그림 4-(c)(d))

대상-③ 한 개 이상의 문자가 붙어 있고 자음이 분리되어 추출된 경우 (그림 4-(f))

대상-④ 하나의 자음블럭과 그 자음에 붙을 모음이 다음 문자의 자음과 붙어서 추출된 경우 (그림 4-(g))

대상-⑤ 한 개 이상의 문자가 붙어 있고 모음이 분리되어 추출된 경우 (그림 4-(h))

대상-⑥ 이전문자의 모음과 다음 문자의 자음이 붙어 있고 그 문자의 모음이 분리되어 추출된 경우 (그림 4-(i))

대상-⑦ 두 개의 블럭이 각기 하나 이상의 문자를 포함하고 자음과 모음을 나누어 가지고 있는 경우 (그림 4-(j))

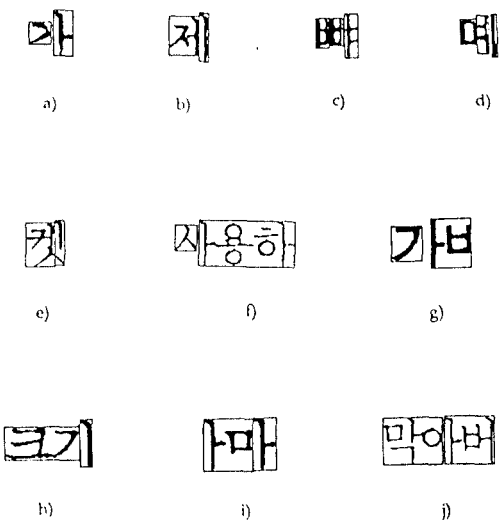


그림 4. 분리 추출된 한글 문자 블럭의 예
Fig. 4. An example of isolating character block for Hangeul

한글 문자 블럭에 있어서 합성 처리의 매개체는 자음 블럭과 모음블럭이 주종을 이루게 된다. 여기서 자음 블럭은 자음으로만 구성된 블럭을 뜻하며 1형식 문자의 초성부분이고, 모음 블럭은 모음으로만 구성된 블럭을 뜻하며 1형식 문자의 중성부분이다. 우선 모음 블럭을 앞 블럭과 합성한 다음, 자음 블럭을 다음 블럭과 합성한다. 그렇게 되면 대상-⑦의 블럭만 남게되고 이를 합성하면 합성 처리는 완료하게 된다.

처리 과정을 좀 더 세부적으로 살펴보면 규칙-Ⅰ, Ⅱ, Ⅲ을 전체 블럭에 적용하여 만족되는 두 개의 블럭을 합성한다. 규칙중에 우선 순위 (priority)는 규

칙-I, Ⅱ, Ⅲ 순으로 주어진다. 처리 예는 각각 그림 5, 6, 7에 보였고 표 1에 모음블럭의 형상비율을 나타내었다.

규칙-I (Rule-I)

대상 — ① ② ⑤ ⑥

조건 I-1 같은 문자열에 존재한다.

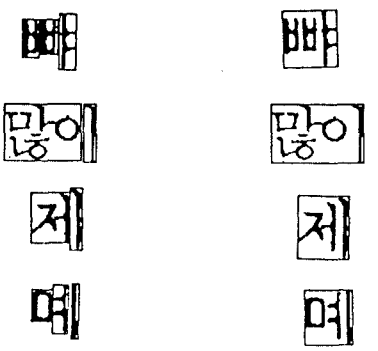
조건 I-2 블럭 사이의 거리가 가깝다.

조건 I-3 피치 비율 (pitch ratio)이 0.5이상이다.

조건 I-4 후행하는 블럭의 형상 비율(aspect ratio)이 0.5보다 작다.

모음블럭 \ aspect ratio	서체		
	명조체 계열	고딕체 계열	궁서체 계열
ㅏ	0.37~0.43	0.18~0.29	0.36~0.42
ㅐ	0.36~0.38	0.31~0.32	
ㅑ	0.37~0.43	0.25~0.26	0.41~0.43
ㅒ	0.35~0.37	0.34~0.35	
ㅣ	0.14~0.21	0.11~0.15	0.21~0.23

표 1. 모음 블럭의 형상 비율
Table. 1. Aspect ratio of vowel block.



(a) 처리전 (b) 처리후

그림 5. 합성 처리의 예(규칙-I)
Fig. 5. An example of merging process. (Rule-I)

규칙-Ⅱ (Rule-Ⅱ)

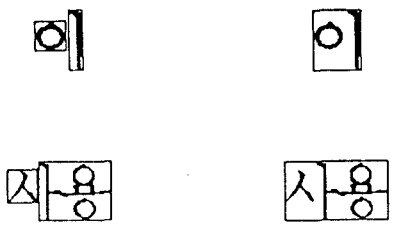
대상 — ① ② ③ ④

조건 Ⅱ-1 조건 I-1, 2의 조건을 만족한다.

조건 Ⅱ-2 블럭의 대각선 (diagonal) 길이가 다음 블럭의 세로폭(height)보다 작다

조건 Ⅱ-3 선행하는 블럭의 중심이 다음 블럭의 중심에서 크게 벗어나지 않는다.

조건 Ⅱ-4 후행 블럭의 피치 비율이 0.5 이상이다.



(a) 처리전 (b) 처리후

그림 6. 합성 처리의 예(규칙-Ⅱ)
Fig. 6. An example of merging process. (Rule-Ⅱ)

규칙-Ⅲ (Rule-Ⅲ)

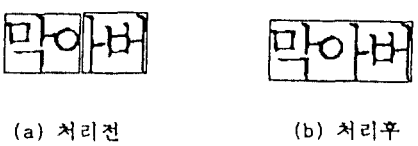
대상 — ⑦

조건 Ⅲ-1 조건 I-1, 2, 3을 만족한다.

조건 Ⅲ-2 선행 블럭의 형상 비율 (aspect ratio)이 1.2보다 크다.

조건 Ⅲ-3 후행 블럭의 형상 비율 (aspect ratio)이 1.2보다 크다.

조건 Ⅲ-4 선행 블럭의 가로폭 (width)이 후행 블럭의 세로폭 (height)보다 크다.



(a) 처리전 (b) 처리후

그림 7. 합성 처리의 예(규칙-Ⅲ)
Fig. 7. An example of merging process. (Rule-Ⅲ)

4.2 기호 문자 블럭의 합성

기호 문자중 하나의 기호 문자를 이루는 블럭이 그림 9처럼 두 개의 블럭으로 나뉘어 추출된 경우가 대상이 되며 다음의 규칙에 부합되면 합성한다.

규칙- Ⅳ (Rule-Ⅳ)

대상 — ①

조건 Ⅳ-1 문자 블럭의 피치 비율(pitch ratio)이 기호 문자 블럭의 특성을 갖는다.

조건 Ⅳ-2 블럭 사이의 거리가 가깝다.

조건 Ⅳ-3 블럭이 나란히 존재한다.

(a) 처리전 (b) 처리후

그림 8. 합청 처리의 예(규칙-Ⅳ)
Fig. 8. An example of merging process.(Rule-Ⅳ)

4.3 한문 문자 블럭의 합성

문서에 존재하는 한문 문자 블럭 특성의 대부분은 한글 문자 블럭의 특성과 유사한 것이 많으며 처리대상은 다음과 같다.

대상-① 하나의 문자가 두 개의 블럭으로 분리되어 추출된 경우 (그림 9-(a))

대상-② 하나의 문자가 세 개의 블럭으로 분리되어 추출된 경우 (그림 9-(a))

상기의 처리대상에 대한 처리과정은 한문 문자 블럭은 거의 한글 문자 블럭의 특성과 유사하기 때문에 규칙-Ⅰ, Ⅱ, Ⅲ을 그대로 적용하여 한문 문자 블럭을 합성한다.

(a) 처리전 (b) 처리후

그림 9. 한문 문자 블럭의 합성 처리 예
Fig. 9. An example of merging process for Chinese

Ⅴ. 문자 블럭의 분리

5.1 분리 처리에 쓰이는 정보

블럭의 분리 처리시에는 단순히 블럭의 이미지 정보와 위치 정보만으로는 정확한 처리를 행할 수 없다. 그런 이유로 이미지 정보와 위치 정보에 추가하여 블럭의 구조 정보를 이용하게 된다. 기존의 연구[8][10][11]에서는 단순히 수직으로 흑화소의 분포를 투영하여 분리점 (break position)을 추정하는 방법을 사용하였는 데 이 방법은 분리점의 후보(candidate)가 여러 개가 존재하게 되고 투영 연결(projectional joint)된 경우는 분리점을 찾기 어렵다. 더욱이 접촉 연결 (touching joint)의 경우 역시 잡음의 영향을 크게 받으며, 두 문자 이상이 접촉된 문자 블럭에 대해서는 각 문자를 정확하게 분리해 내기가 어렵다는 단점[8]을 가지고 있다. 본 논문에서는 흑화소의 밀도 분포만으로는 알아내기 어려운 분리점을 찾아 내기 위하여 흑화소의 분포 위치를 사용하는 새로운 척도 (metric)를 제안한다.

5.1.1 블럭의 선택

분리 처리를 행하기 전에 우선적으로 필요한 것은 문자들이 접촉된 문자 블럭을 찾아내는 것이다.

본 논문에서는 3장에서 다룬 한글 문자 블럭의 특성을 조사하여 분리 처리의 대상이 될 블럭을 찾아내었다.

(1) 한글끼리 접촉된 문자 블럭의 특성
형상 비율 (aspect ratio)이 1.7이상으로 나타난다.

(2) 한글과 기호가 접촉된 문자 블럭의 특성
기호 문자의 크기가 다양하기 때문에 기호와 한글이 접촉된 문자 블럭은 형상 비율이 다양하게 나타나지만 최소 1.0이상으로 나타난다.

(3) 한문끼리 접촉된 문자 블럭의 특성
한문의 형상 비율이 한글과 유사하기 때문에 1.7이상으로 나타난다.

(4) 한문과 한글이 접촉된 문자 블럭의 특성
마찬가지로 1.7이상으로 나타난다.

위의 특성을 조합하여 블럭의 형상 비율이 1.0이상인 블럭을 분리 처리의 대상으로 한다.

5.1.2 블럭의 내용

문자들이 접촉되어 추출된 문자 블럭은 외향적으로는 똑같은 형태를 가지고 있지만 실제 내부의 형태는 차이가 있다. 블럭내의 문자들이 접촉되는 형태는 두 가지로 나타난다. 또한, 본 논문의 방법에 의해 블럭 합성시 나타나는 강제 연결 (force joint)된 경우가 있다. 그림 10-(a)와 같이 문자는 분리되어 있지만 수평 투영값을 계산한 경우 중첩되어 문자가 병합된 것으로 판정되는 경우의 병합을 투영 연결 (projectional joint)이라 한다. 또한 (b)의 경우처럼 문자끼리 직접 연결되는 경우의 병합을 접촉 연결 (touching joint)이라 한다. 부가적으로 본 논문에서는 (c)의 경우를 문자 블럭의 합성 처리시 생기는 것으로 강제 연결 (force joint)이라 정의한다.

(a) 투영연결 (b)접촉연결 (c)강제연결

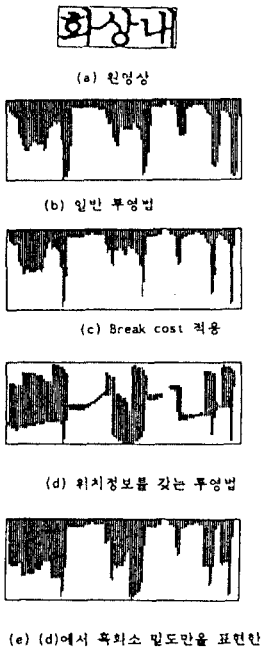
그림 10. 문자간의 접촉 형태의 예
Fig. 10. An example of joint-form between characters.

5.1.3 블럭의 구조 정보

본 논문에서 사용하는 구조 정보는 흑화소의 밀도 분포와 흑화소의 위치이다. 흑화소의 위치를 사용함으로써 흑화소의 밀도만을 이용하는 일반적인 투영법에서 분리점을 찾기 어려운 기호 문자를 정확하게 분리해 낼 수 있다. 또한, 문자 사이의 접촉도를 더욱 분명하게 보여 줄 수 있도록 Tsujimoto가 제안한 브레이크 코스트 (break cost)를 개선하여 적용하였다. 적용 순서는 블럭내의 문자에 대해 브레이크 코스트를 적용하고 블럭내의 구조 정보를 적용한다.

즉 투영 연결된 문자 블럭의 경우에는 두 개의 이웃 쌍에 AND연산을 취하는 것보다 세 개의 이웃 쌍에 대해 AND연산을 취하게 되면, 문자의 접촉도가 더 정확하게 나타난다. 이것은 문자 접촉 정도가 심한 경우에 대응하기 위해서이다.

한편 블럭내의 구조정보는 문자 블럭을 수직 투영 (vertical projection)하면서 처음 흑화소가 나타나는 위치와 마지막으로 나타나는 위치를 정보로 사용한다. 그리고 그 사이 부분을 흑화소 부분으로 간주함으로써 구해진다. 그림 11에 처리의 예를 보였다.



(a) 원영상
(b) 일반 투영법
(c) Break cost 적용
(d) 위치정보를 갖는 투영법
(e) (d)에서 흑화소 밀도만을 표현한 경우

그림 11. 투영된 문자 블럭의 예
Fig. 11. An example of projected character block.

5.2 한글 문자 블럭의 분리

한글 문자 블럭의 합성 처리가 끝난 후에는 한글 문자가 두 개이상 접촉된 문자 블럭만이 남게 되며 본 절에서는 이러한 한글 문자 블럭의 분리 처리에 대해서 논한다.

한글 문자는 문자폭 (가로폭)이 균일한 특성이 있으며 이 특성에 근거하여 블럭의 왼쪽 시작점에서부터 블럭의 세로폭을 변위 (offset)로 하여 변위만큼 떨어진 지점을 기준점으로 하여 이점을 중심으로 분리 처리를 행하게 된다.

처리 순서는 다음과 같다.

(1) 기준점을 중심으로 하여 블럭 세로폭의 25%되는 폭만큼 좌우를 수직 투영한다.
(2) 브레이크 코스트 (break cost)를 적용한다.
(3) 브레이크 코스트를 적용하면서 흑화소의 위치 정보를 구한다.
(4) 위치 정보에 따른 흑화소의 밀도를 구한다.
(5) 흑화소의 밀도가 최소가 되는 점을 분리점으로

결정한다.

(6) 분리점을 기준으로 방금 분리된 문자의 가로폭 만큼을 새로운 변위로 정하여 그만큼 떨어진 지점을 새로운 기준점으로 정한다. 단 차후 변위는 고정되고 변하지 않는다.

(7) (1) ~ (6)과정을 반복한다.

그림 12에 본 방법의 처리예를 보였다.

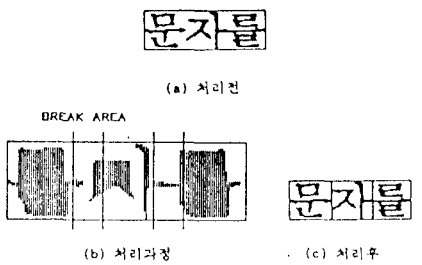


그림 12. 한글 문자 블럭의 분리 처리의 예
Fig. 12. An example of segementation for Hangeul character block.

5.3 한글과 기호가 붙은 문자 블럭의 분리

한글과 기호가 접촉되어 추출된 블럭은 형상 비율이 1.0이상으로 나타나며 분리점을 추정하기 위해서 기준점을 설정하게 되는 데 기호 문자가 한글의 좌측에 위치할 때는 이 기준점을 그대로 사용하게 되면 오류가 생긴다. 이는 기호 문자의 가로폭 (width)이 한글 문자에 비해 작은 경우가 있기 때문이다.

따라서 문자 블럭의 좌측 시작점(xinit)부터 블럭의 세로폭의 절반에 이르는 곳까지 투영하여 분리점을 찾게 된다. 이 때 흑화소의 위치정보에 의해서 기호 문자가 한글과 접촉될 수 있는 경우의 조건을 조사하여 분리를 행하게 된다. 이 과정은 한글 문자 블럭의 분리처리시 병행 처리된다. 분리 처리의 예를 그림 13에 보였다.

또한 한글 문자의 우측에 기호 문자가 접촉된 문자 블럭의 분리 처리시의 기호 문자의 존재 위치는 대부분 블럭의 끝 부분이 된다. 따라서 기준점의 추정은 자연히 이루어진다. 한글의 경우와 마찬가지로 좌우 영역을 투영하여 흑화소 위치 정보에 따라 분리를 행한다. 그림 14에 예를 보였다.

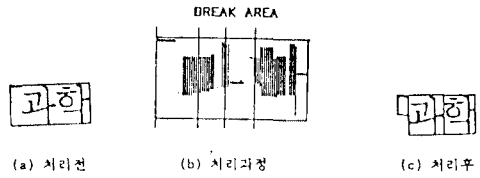


그림 13. 기호 문자 블럭의 분리 처리 예 Ⅰ.
Fig. 13. An example for symbolic character block Ⅰ.

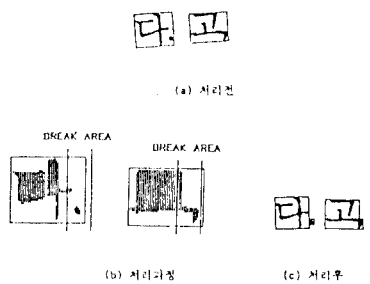


그림 14. 기호 문자 블럭의 분리 처리 예 Ⅱ.
Fig. 14. An example for symbolic character block Ⅱ.

5.4 한문 문자 블럭의 분리

한문 문자 블럭의 특성은 한글 문자 블럭의 특성과 유사하기 때문에 한글 문자 블럭의 분리 알고리즘을 그대로 적용한다. 그림 15에 예를 보였다.

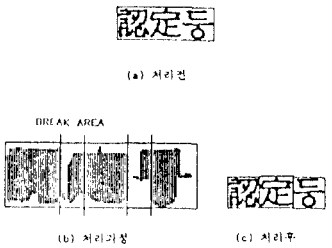


그림 15. 한문 문자 블럭의 분리 처리 예.
Fig. 15. An example for segmentation for chinese character block

Ⅵ. 실험및 고찰

6.1 실험및 결과

본 논문의 실험은 스캐너 (scanner)로 입력받은 문서 데이타에 대해 행하였고 SUN Sparc-Ⅱ workstation을 사용하여 X-window상에서 C-언어로 구현하였으며 그림 16에 시스템의 흐름도를 보였다.

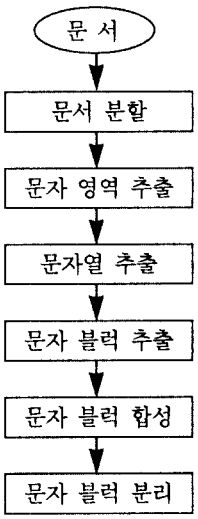
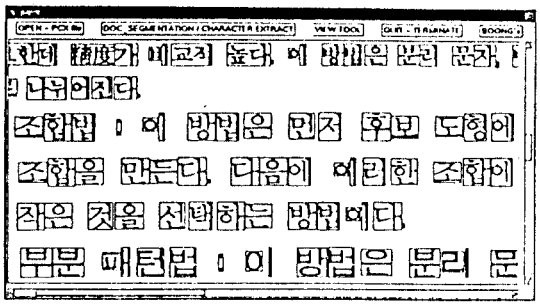
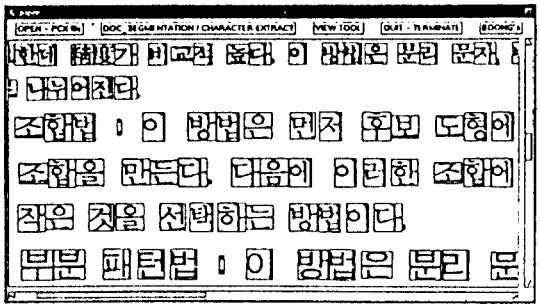


그림 16. 시스템의 흐름도.
Fig. 16. A flowchart of system.

실험 데이타는 인쇄체 한글 문서 (논문지, 잡지, 교과서)이며, 실험 데이타에 쓰이는 서체 (font)는 문서에서 주로 쓰이는 명조체 계열 (명조, 세명조, 견명조등)과 고딕체 계열 (고딕, 중고딕, 견고딕)과 부가적으로 궁서체 등이며 문자 크기는 7 포인트에서 72 포인트까지 처리 가능하며 문서내에 혼재하여도 무관하다. 실험 처리된 예를 그림17 그림18에 보였다.

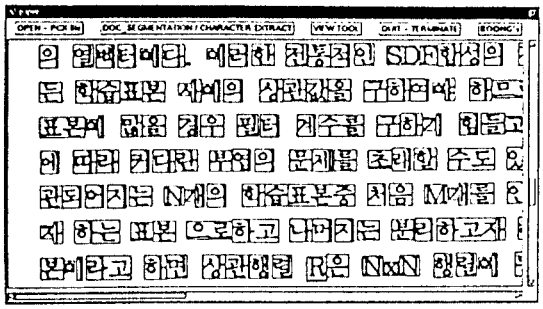


(a) 추출된 문자블럭

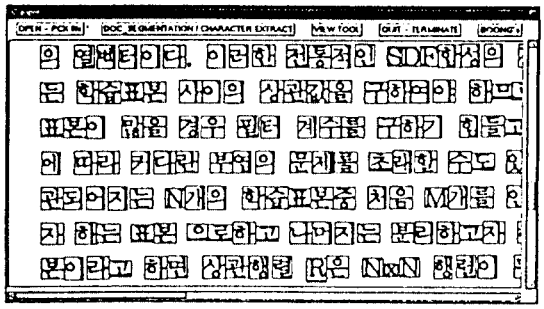


(b) 합성, 분리처리를 거친 후의 문자블럭

그림 17. 실험 예(크기가 다른 문자 혼재)
Fig. 17. An example for experiment.



(a) 추출된 문자블럭



(b) 합성, 분리처리를 거친 후의 문자블럭

그림 18. 실험 예(논문지)
Fig. 18. An example for experiment.

실험 예에서 보듯이 본 논문에서 제안하는 방법이 서체와 크기에 적응할 수 있음을 알 수 있다. 표 2에 실험 결과를 보이고 있다.

6.2 고 찰

본 논문에서 제안하는 알고리즘은 기존의 방법이 정적인 정보 (평균 글자폭)에 의존하던 것에서 벗어나 문자 크기에 상관없이 정보를 자동으로 획득하여 분리 추출된 문자의 합성을 정확하게 수행하였다. 또

처리 / 서체	합성처리후 추출율	합성, 분리처리후 추출율	기존의 방법
명조체 계열	96.8%	99.1%	93.1%
고딕체 계열	96.2%	98.8%	96.3%
궁 서 체	95.5%	99.1%	95.2%

표 2. 실험 결과
Table. 2. Result of experiment.

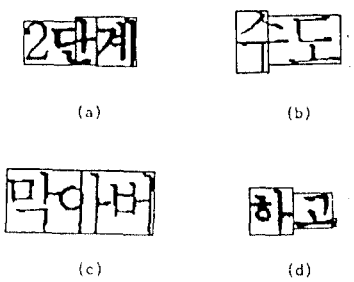
한, 기존의 방법[8]으로는 처리하지 못했던 접촉된 한글의 분리나 한글과 기호의 분리, 한글과 한문의 분리를 행할 수 있는 새로운 투영법을 제시하였다.

표 2에 보이는 기존 방법의 추출율은 크기가 일정한 문서에 대해서 적용한 예이다. 이 경우에도 기호류와 접촉된 문자는 추출해내지 못했으며 두 개이상의 문자가 접촉된 경우에는 추출이 잘못된 경우가 있었다. 일례로 그림 19에서 보듯이 정적인 정보를 사용하여 분리점을 잘못 설정하는 기존의 방법보다 본 논문의 방법이 더욱 정확하게 문자를 추출해 내는 것을 알 수 있다. 그림 19-(a)에서는 평균 글자폭에 의한 기준점 설정이 잘못되고 더욱이단순한 투영에 의한 흑화소 밀도수를 사용한 결과 "화상내"라는 문자가 "회상내"로 잘못 분리된 것을 볼 수 있다. 그러나 그림 19-(b)에서 보듯이 본 논문의 방법에서는 각 문자의 접촉 정도를 명확히 알 수 있고 기준점 추정시 동적으로 기준점을 재설정하기 때문에 추출의 정확도가 높음을 알 수 있다.

실험 결과 99%이상의 추출율을 보였는 데 문서내에 분리되거나 접촉된 문자가 많을 경우는 추출율이 98.6%까지 떨어지나 분리되거나 접촉된 문자가 적은 경우는 추출율이 최대 99.5%이상이었다. 문서내에 분리된 문자나 접촉된 문자가 많은 경우는 실험상에서 전체 문자에 대해 최대 51.4%이었으며 분리되거나 접촉된 문자가 적은 경우는 전체 문자에 대해 최소 10.9%였다. 고딕체 계열의 문자가 추출율이 명조체 계열에 비해 낮은 이유는 합성 처리 대상이 더 많기 때문에 합성 처리의 오류가 더 많았으며, 이에 따라 분리 처리후에도 오류가 추가되었기 때문이다. 대부분의 오류가 생기는 원인으로는 그림 20-(a)와 같이 문서내에 존재하는 한글의 자음 블럭과 같은 특성을 가진 소수의 영, 숫자가 합성 조건에 의해 한글과 잘못 합성되는 경우가 있었으며, 이와 더불어 분리 처리시에 오류가 생겼다. 또한 스캐너 입력시의 에러로 합성 조건에 포함되지 않는 블럭들에 대해 합성 처리가 잘못된 경우가 있었다. 즉, 그림 20-(b)에서 처럼 문자 자체의 구조 특성에 무관하게 입력 상태가 불완전하여 문자의 합성이 제대로 수행되지 않는 경우가 있었다. 그림 20-(c)에서처럼 분리 처리시에도 잡지와 같이 문서 작성시 일반적인 문자 특성에서 벗어난 문자를 사용할 경우 (세로폭이 상대적으로 긴 문자)에는 분리점을 찾기 위해 주어지는 변위값에 의해 잘못 분리되는 경우가 있었고 그림 20-(d)에서 보듯이 기호를 추출하기 위한 조건에 의해 기호가 아닌 문자도 분리되는 경우가 있었다. 또한 한문의 경우, "卜"字와 같이 한글의 모음블럭과 같은 블럭 정보를 갖고 블럭사이의 거리가 가까울 경우에 합성처리시 오류가 생기는 경우가 있었다.

(a) 기존 방법 적용의 예 (b) 본 방법의 적용의 예

그림 19. 기존 방법과의 비교
Fig. 19. Comparision between traditional method and proposed method.



(a) (b) (c) (d)

그림 20. 잘못된 예
Fig. 20. Example of error.

Ⅶ. 결 론

본 논문에서는 문자 인식에 선행되어야 할 문자 추출에 관한 연구를 행하였으며 문자 블럭의 형상 비율과 피치 비율과 같은 이미지 정보와 위치 정보를 사용함으로써 문자 크기에 따라 동적으로 합성 조건이 되는 정보의 획득이 이루어짐으로 다양한 크기의 문자에 대해 합성 처리가 가능하고, 접촉 문자의 분리 처리시 이전의 국내 연구에서 다루지 못했던 기호 문자와 한문의 접촉시 접촉 상태를 조사할 수 있는 혹화소 위치 정보를 알 수 있는 투영법을 제안하였다.

일반 문서에 제안한 방법을 적용한 결과 99%이상의 문자 추출을 행할 수 있었으며, 앞으로의 과제는 영문과 숫자추출에 관한 연구와 병행되어 한글 문서에 존재하는 다양한 문자에 적용될 수 있는 문자 추출법에 대한 연구가 이루어져야 할 것이고 문자의 인식 결과를 이용하여 오류를 수정하며, 문맥적 지식을 사용하여 추출의 정확도를 향상시킬 수 있는 연구가 계속되어야 할 것이다.

참고문헌

1. F. M. WAHL, K. Y. WONG, R. G. CASEY, "Block segmentation and text extraction in mixed text/image documents," Computer, Graphics and Image Processing, No. 20, pp. 375-390, 1982.
2. F. M. WAHL, K. Y. WONG, "An efficient method of running a constrained run length algorithm (CRLA) in vertical and horizontal directions on binary data image," IBM J. Res. Develop. 1982.
3. D. WANG, S. N. SRIHARI, "Classification of newpaper image blocks using texture analysis," Computer, Vision, Graphics, and Image Processing, No. 47, pp. 327-352, 1989.
4. T. SAITOH, T. PAVLIDIS, "Page segmentation without rectangle assumption," 11th IAPR International Conference on Pattern Recognition, Vol. 2, pp. 277-280, 1992.
5. 남궁 재찬, 류 황빈, 남궁 연, "한국어 문서로부터 문자 분리및 도형 추출에 관한 연구," 한국 전자공학회 논문지, Vol. 25, No. 9, pp 1091-1101, 1988.
6. Y. K. HAM, H. K. CHUNG, I. K. KIM, R. H. PARK, "Hierarchical recognition of mixed documents consisting of the Korean /Alphanumeric texts and graphic images," MVA '92 IAPR Workshop on Machine Vision Applications, pp. 287-290, 1992.
7. 신 현관, "문서의 영역 분리와 레이아웃 정보 추출에 관한 연구," 광운 대학교 대학원 석사 학위 논문, 1992.
8. 오 인권, "영문이 혼합된 한글 문서에서의 문자및 특수 문자 추출에 관한 연구," 광운 대학교 대학원 석사 학위 논문, 1988.
9. H. S. BAIRD, H. BUNKE, K. YAMAMOTO, Structured Document Analysis, Springer-Verlag Berlin Heidelberg, pp. 80-81, 1992.
10. S. TSUJIMOTO, H. ASADA, "Major components of a complete text reading system," Proc. of IEEE, Vol. 80, No. 7, pp 1133-1149, 1992.
11. N. SUN, M. SUZUKI, Y. NEMOTO, M. KIMURA, "文字構造情報に基づく高精度な文字切出し處理も用いた文書認識システム," 日本 情報處理學會論文誌, Vol. 33 No. 9, pp. 1083-1091, 1992.

洪 性 鵬 (Sung Boong Hong)

1992년 2월 : 광운대학교 전자계산기공학과 졸업(공학사)
1994년 2월 : 광운대학교 대학원 전자계산기공학과 졸업(공학석사)
1994년 3월~현재 : 포스데이타 주식회사 기술연구소 연구원 재직중
※ 주관심분야 : 패턴인식, 문서이해, 컴퓨터 비전

南宮 在 贊 (Sung Boong Hong) 정회원

1970년 2월 : 인하대학교 전기공학과 졸업(공학사)
1976년 8월 : 인하대학교 대학원 전기공학과 졸업(공학석사)
1982년 3월~현재 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 : 일본 동북대학 객원교수
1979년~현 재 : 광운대학교 컴퓨터 공학과 교수
※ 주관심분야 : 패턴認識, 神經回路網, 文書認識

張 喜 敦 (Hee Done Jang) 정회원

1985년 2월 : 원광대학교 전자계산공학과 졸업(공학사)
1987년 2월 : 광운대학교 대학원 전자계산기공학과 졸업(공학석사)
1991년 3월~현재 : 광운대학교 대학원 전자계산기공학과 박사과정 재학중
※ 주관심분야 : 패턴認識, 神經回路網, 文書認識