論 文

인쇄체 문서의 문자영역에서 한글과 한자의 구별에 관한연구

正會員 沈 相 完* 正會員 李 成 範** 正會員 南宮 在 贊*

A Study on Classification into Hangeul and Hanja in Text Area of Printed Document

Sang Ouan Sim*, Sung Bum Lee**, Jae Chan NamKung* *Regular Members*

要 約

본 논문에서는 문서인식시스템의 문자인식부에서 각 문자를 인식하기 위한 전처리 단계인 한글과 한자를 구별하는 알고리즘을 제안한다. 본연구에서는 문자의 구별에 큰 영향을 미치고, 쓰기형태와 글자체에 따라서 변동을 흡수할 수 있는 9가지의 한자 특성을 제안하고, 문자의 크기에 영향을 받지 않고 문자를 구별할 수 있도록 문자 크기에 따른 비율을 제안된 각 특성에 반영하여 문자의 구별을 행하였다. 입력된 문서의 문자영역에 블럭화를 행하여 각각의 문자를 분리해내고, 다음으로 분리된 문자에 대하여, 본 논문에서 제안한 9가지의 한자 구조적 특성을 조사하여, 한글과 한자로 구별한다.

KS-C5601의 한글 2350자와 한자 4888자의 고딕, 명조체에 대하여, 실험결과는 인쇄 표본, 신문, 학회지, 잡지, 교재에서 각각 98.8%, 92%, 96%, 98%, 98%을 얻었다.

ABSTRACT

This paper propose an algorithm for preprocessing of character recognition, which classify characters into Hangeul and Hanja. In this study, we use the 9 structural chacteristics of Hanja which isn't affected by deformation of size and style of characters and rates based on character size to classify characters. Firstly, we process the blocking to segment each characters. Secondly, on this segmented characters, we apply algorithm proposed in this paper to classify Hangeul and Hanja. Finaly, we classify characters into Hangeul and Hanja, respectively.

An experiment with 2350 Hangeul and 4888 Hanja printed Gothic and Mincho style of KS-C 5601 are carried out. We experiment on typeface sample book, newspapers, academic society's papers, magazines, textbooks and documents written out word processor to obtain the classifying rates of 98.8%, 92%, 96%, 98% and 98%, respectively.

*光云大學校 電子計算機工學科
**大宥工業專門大學
論文番號 : 93-80

I. 서 론

급속도로 변환하는 정보화 사회에서 인간이 생활

하고 있는 모든 분야에 걸쳐 정보처리 시스템이 광범위하게 응용되고 있으며, 매우 빠른 속도로 개발, 확산되어가고 있다. 이에따라 사용되어지는 정보의 매체도 다양하고, 정보량도 방대하게 되었다. 이렇게 발생한 정보를 처리할 컴퓨터 하드웨어의 기술은 매우 빠르게 발전되어오고 있으나, 정보의 입력에 있어서 대부분 인간의 수고를 요하는 기존 입력 방식인 키보드(key board)를 통해 데이타로 입력되어져 왔다.

따라서 기존의 입력 방식을 사용하여 정보를 입력하는 것은 막대한 데이타를 보다 신속하게 처리 요구되는 현재 정보화 시대에서 점점 자리를 잃어가는 것이 자명한 일이다. 기존 정보 매체에 있어서 거의 대부분이 문서로 작성되어져 있고, 현재는 컴퓨터의 발달로 전자출판 시스템이 보편화 되었고, 문서의 질도 매우 향상되어졌다. 이러한 문서 영상을 자동인식할 수 있다면, 문서의 편집, 수정 그리고 보안에 있어서 시간적인 단축과 업무의 신속성을 기할 수 있을 것이다. 이러한 시대적 조류에 부응하기 위해 문서의 자동입력에 대한 연구의 필요성이 일찌기 대두되었다. 기존 문자의 인식은 단일 자종 즉 한글, 한자만을 인식하였으나, 문서에는 여러가지의 문자가 다른 크기로 동시에 존재하기 때문에 이렇게 다양한 문자를 동시에 처리하기 위해서는 문자의 종류가 무엇인가를 구별해야 한다. 본 논문에서는 이러한 점을 고려하여 문자의 크기가 비슷한 한글과 한자를 자동으로 구별하여 문자를 인식하기 위한 집단(cluster)을 줄였다. 문서인식에 있어서 이러한 기능은 필수적이며 이렇게 하지 않으면 문서인식의 속도를 감당하기 매우 어렵다. 따라서 본 논문은 문서인식시스템에서 인식을 효과적으로 하기 위하여 한글과 한자를 자동으로 구별하는데 목적을 둔다.

문서인식시스템은 스캔너(scanner) 또는 카메라(camera)를 통하여 얻어진 문서에 대한 영상을 입력하고, 입력 문서 영상으로 부터 전처리 후 그래픽영역(graphics) 및 문자영역(text) 부분을 분리한 후, 그중 문자영역 부분에만 단어별 블럭화를 취하여 각 블럭에 대한 문자인식이 필요하다. 우리나라가 한자문화권에 속하고 거의 모든 문서가 한글과 한자가 혼용되어 사용되고 있으므로 문자의 인식시에 문제로 되는 문자수를 줄이기 위하여 한글 또는 한자로 자동구별하여 각 인식 시스템으로 자료를 넘겨주면 문서인식시스템에서는 문자를 인식하는데 빠르고 효율적으로 처리할 수 있다.

본 논문에서는 한글과 한자가 포함된 문서에서 한글과 한자를 자동으로 구별을 시도하였다. 지금까지의 문서영상의 연구 상황을 살펴보면 김진형[1], Dacheng Wang[2] 등은 신문에 대하여 신문의 구조적 분석과 문자 부분을 블럭화하여 신문 기사의 추출에 관한 연구를 하였다.

남궁연[3]등은 그림과 문자가 혼합되어 있는 문서영상에서 그림영역과 문자영역을 추출하는 연구를 하였으며, 신현관[4]등은 문서의 영역분리와 레이아웃 정보 추출에 관한 연구를 하였다. 또한 오인권[5]은 복합된 문서에서 문자를 블럭화하는 DOWN-UP 알고리즘을 제안하여 문자의 블럭화를 용이하게 하였다. 그리고 한자 자체에 대한 구조적 분석을 위한 연구를 보면 이주근[6]은 한글을 6가지 형식으로 분류하였고, 남궁재찬[7][8]등은 한글의 자소를 추출하는 Index-Window 알고리즘을 제안하여 인식을 용이하게 하였고, 또한 남궁재찬[9]은 한글이 가지는 특성의 분석에 관한 연구도 하였다. 한자 자체에 대한 연구를 보면 Yasuaki Nakno[10]등은 한자 구성요소의 구조적 특성과 주변분포에 따른 문자 히스토그램의 특징성에 관한 연구를 하였다. 김학성[11]은 한자의 구조 분석에 관한 연구를 하였다. 특히 이승형[12]은 한글과 한자의 구별과 한글의 형식 분류에 관한 연구를 하였는데, 주로 한글의 특징을 중점으로 구별 하였기 때문에 구별률이 약간 저조하여 본 연구에서는 한자 특징 요소를 중심으로 알고리즘을 개선하여 바람직한 구별률을 얻게 되었다.

본 논문에서는 문서 인식을 용이하게 하기위하여 인쇄체 문서를 영상 입력받아 문자에 대하여 한글과 한자를 자동으로 구별하도록 하였다. 문자를 구별하기 위하여 먼저 고딕체와 명조체의 한글과 한자를 영상 입력을 통해 한글과 한자의 구조 특징을 추출하였고, 실험대상 문서는 일반 인쇄체 문서와 학회지로 하였다. 입력된 영상중에 문자영역 부분만을 블럭화한 다음, 블럭화를 거쳐서 얻어진 단위 문자에서 본 논문에서 제안한 한자의 특징을 적용하여 한글과 한자를 구별하는 연구를 하였다.

본 논문은 제 1장 서론, 제 2장 문서내의 요소 특징, 제 3장 한글과 한자의 구조 특성, 제 4장 한글과 한자의 구별, 제 5장 실험및 고찰, 제 6장 결론의 순으로 구성되어 있다.

Ⅱ. 문서내의 요소 특징

문자구별을 하기 위한 전단계로, 문서를 구성하고 있는 기본구조를 분석하고, 특히 쓰기형태에 따라 문자 구조적 특성이 달라지므로 문자영역을 이러한 관점으로 분석한다.

2.1 문서내의 구조

일반적으로 문서를 이루고 있는 구성 요소는 문자영역(text)과, 도표, 그림, 사진 등을 포함한 그래픽 영역(graphic)으로 그림 1과 같이 분류할 수 있다. 본 논문에서는 그림영역을 제외한 문자영역 만을 대상으로 한다.

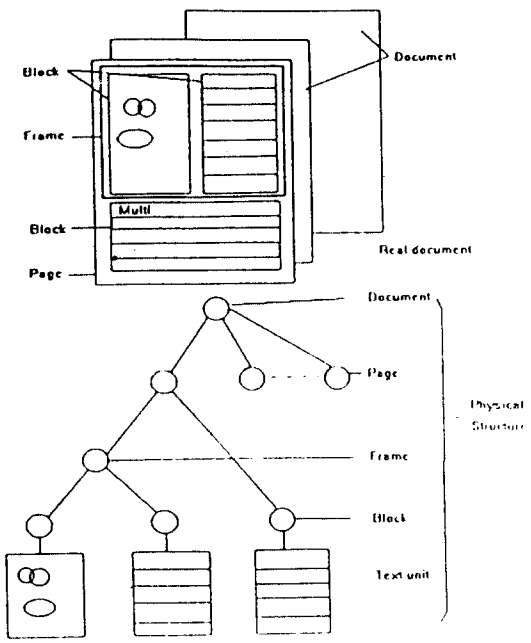


그림 1. 문서의 구조
Fig. 1. Layout structure of document

2.2 문서내의 문자열의 서식 유형 조사

현재 통용이 되고 있는 문서의 서식을 살펴보면 그림 2와 같이 가로쓰기와 세로쓰기형태로 구분되어지고 있다. 이러한 정보는 문자의 블럭화시 문자의 놓인 위치가 고려되어 블럭화되어야 한다. 신문, 논문, 교과서를 대상으로 문자열 서식 유형과 글자체를 조사해 본 결과 표 1과 같다. 신문이나 논문 그리고 교과서에서 쓰이는 문자는 한글만이 쓰이는 것이 아니라 다양한 문자 즉, 한자, 영자, 숫자, 특수 문자등이 쓰이고 다양한 문자의 크기가 사용되는 것을 알 수 있다.

표 1. 문서의 서식 유형과 글자체 분석
Table 1. Analysis of writing form and font in document

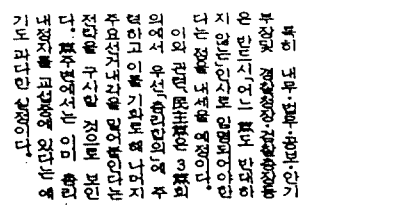
문서 종류		서식	글자체
신문	조선일보	세로	명조
	한겨레신문	가로	명조
	한국일보	세로	명조
	동아일보	세로	명조
논문	전자공학회	가로	명조
	정보과학회	가로	명조
	한국통신학회	가로	명조
교과서	대학교국어	가로	명조
	고등학교국어	가로	명조

2.3 문서의 쓰기 형태 비교

본 연구에서는 가로쓰기 형태를 중심으로 연구를 행하였으므로 이를 중점적으로 살펴보도록 한다. 가로쓰기형태는 그림 2에서 볼 수 있듯이 한글의 특성[9]에서 연구한 것처럼 한글의 굵기의 변화보다는 한글 자체의 크기(body size)에 따라 자소의 위치가 매

대상물은 光源으로 부터의 빛이 없이는 感知될 수 없다. 光源은 그 자체의 物理的 特性을 가지고 視測者의 色感知에 영향을 준다. 프린터가 製作되어 畵面을 놓고 볼 때 光源으로 부터 照射되어 畵面에서 反射될 때 對相物인 畵面의 物理化學的 特性이 또한 영향을 준다. 視測者는 機械的 카메라가 아니다. 관측자는 눈과 신경계통과 뇌의 유기적인 생리적 심리적 작용에 의하여 대상물과 칼라를 인식한다. 따라서 칼라를 논할 때 光源과 照射對相物과 視測者가 모두 고려 되어야 한다.

(a) 가로형식 (a) Horizontal writing



(b) 세로 형식 (b) Vertical writing

그림 2. 서식 유형
Fig. 2. Writing form

우 민감하게 변하는 것을 알 수 있다. 세로쓰기 형태는 가로쓰기 형태와 달리 문자의 높이(height)가 많이 줄어든 형태를 가지고 있다. 그러나 본 논문에서는 쓰기 형태에 따른 문자의 변형을 흡수할 수 있는 특징점도 고려하여 문자의 구별에 이용하였다.

Ⅲ. 한글과 한자의 구조 특성

문서 인식 시스템은 전처리 단계에서 한글과 한자를 구별할 수 있도록 구조 특성에 따라 각 문자의 구별 시스템을 설계해야 한다. 따라서 본 장에서는 한글과 한자의 용어를 알아본 다음, 한글과 한자의 기본적 구조를 분석하고 구조적 특성에 대하여 알아본다.

3.1 한글의 기본 줄기와 용어

문자의 구조에 있어서 글자꼴을 이루고 있는 가장 기본이 되는 줄기(획, stroke)를 기본 줄기라 말하고, 이러한 기본 줄기들이 여러 형태로 놓임에 따라 하나의 낱글자가 구성된다. 기존의 한글과 한자구별에 대한 연구에서는 글자체가 미치는 영향이 매우 컸다.

본 연구에서는 글자체가 주는 영향을 최소화 시키기 위하여 현존하는 글자체중 가장 많이 이용되고 있는 고딕체와 명조체를 대상으로 하였고, 1979년에 한글 시각 문제연구 단체인 글꼴 모임[13]에서 정의한 기본 줄기 용어를 사용하였다. 그림 3은 글꼴 모임에서 정의한 기본 줄기와 용어를 보였다.

그림 3. 한글의기본 줄기와 용어
Fig. 3. Primary stem and vocabulary of Hangeul

3.2 한글의 구조 분석

한글은 자소들이 모여 구성된 조합 문자로서 기본 자모(단자음 14자, 단모음 10자)가 혼합되어 초성, 중성, 종성을 구성한다. 한글의 구성 요소는 표 2와 같다.

본 연구에서는 한글의 특징을 추출하는데 모음의 배치 및 존재 여부에 따라 정의한 이주근의 한글의 6형식[6]을 사용하였다. 한글은 한자와 같은 조합 문자이지만, 한자의 무게 중심이 사각표 내에 골고루 분포 되어 있는 것과는 달리, 한글의 무게 중심이 종모음과 횡모음에 분포되어 있는 것이 특징이다. 그림 4에서는 한글의 6형식 구조를 나타내고 있다.

본 연구의 한글과 한자 구별 대상으로 선정된 글자체는 명조체와 고딕체이며, KS-C5601 표준에서 정의한 2350자의 한글과 4888자의 한자를 대상으로 본 연구의 실험을 하였다.

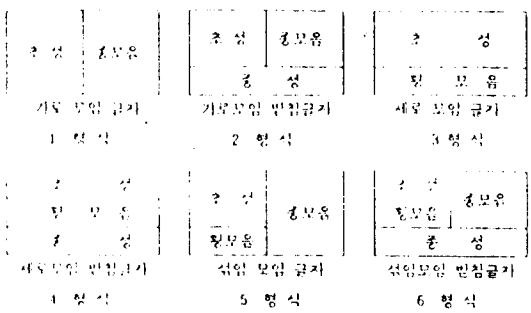


그림 4. 한글의 6 형식
Fig. 4. 6 types of Hangeul

표 2. 한글의 구성 요소
Table 2. The constructural element of Hangeul

종 류	구 성 요 소
한글문자	초성, 중성, 종성
초 성	단자음, 쌍자음
종 성	단자음, 쌍자음, 복합자음
중 성	모음
모 음	단모음, 복합 모음
단 자음	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
쌍 자음	ㄲ, ㄸ, ㅃ, ㅆ, ㅉ
복합 자음	ㄳ, ㄵ, ㄶ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅄ, ㄲ
단 모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ
복합 모음	ㅐ, ㅒ, ㅔ, ㅖ, ㅘ, ㅙ, ㅚ, ㅝ, ㅞ, ㅟ, ㅢ
종 모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅣ, ㅐ, ㅒ, ㅔ, ㅖ
횡 모음	ㅗ, ㅛ, ㅜ, ㅠ, ㅡ

3.3 한글의 구조적 특성

한글은 6형식에 따라 자음과 모음의 위치가 정해지는 글자이며, 매우 과학적인 문자이나 한자와 구조적으로 유사성이 매우 높아 구별하는데 어려움이 많다.

귀에서 언급되었지만 한글은 모음 중심 문자, 즉 종모음과 횡모음의 형태에 따라 글자꼴의 형태가 민감하게 변하기 때문에, 본 연구에서는 다음과 같이 8가지 특징을 이용하여 한글과 한자의 구별을 용이하게 하였다.

(1) 한글에서 각 자소들 중 수평 방향의 횡 모음이 가장 길다.
(2) 3, 4, 5, 6 형식 중 횡 모음이 수평 방향으로 가장 앞쪽에 온다.
(3) 한글에서 각 자소들 중 수직 방향으로 종 모음이 가장 길다.
(4) 1, 2, 5, 6 형식중 종 모음이 가장 윗 쪽에서 시작한다.
(5) 한글의 종성은 횡 모음 보다 앞에 오지 않는다.
(6) 한글의 자음은 대부분 직선 성분이 있다.(예외, ㅇ, ㅅ)
(7) 종모음은 위치 변화가 적지만 종성에 따라 크기변화가 있다.
(8) 횡모음은 위치 변화는 크지만 종성에 따라 크기변화가 없다.

3.4 한자의 요소와 용어

한자의 구성은 그 생성과정에 따라 분류하면 크게 6가지(육서)로 나눈다. 표 3은 6육서에 의한 한자의 자구성을 나타낸다. 본 연구에서 데이타로 삼은 한자는 모필 서체의 의미지를 갖는 명조체와 고딕체인데 오늘날 한자의 명조체는 모필 서체의 이미지 보다는 금속활자의 발달에 기인한 기계적 명쾌함을 가진 서체이다. 그림 5는 명조체를 기준으로한 엘리먼트(element)를 나타낸 것으로 각각의 엘리먼트는 나름대로의 이름과 개성을 가지고 있다[11].

3.5 한자의 구조 및 특성

한자의 구조를 살펴보면, "台"와 같은 상하구조(North-southstructure), "化"와 같은 좌우구조(East-west structure), "國"과 같은 내외구조(Border-Interior), 그리고 이 세 구조의 복합구조등으로 구성되어 있는 경우가 아주 많다. 즉 한자는 상형문자에서 시작되었다고 하지만 문명의 진보와 발달과 함께 문자에 의해

그림 5. 한자의 엘리먼트와 그 명칭
Fig. 5. The element of Hanja and it's name

표 3. 한자의 자구성
Table 3. Character configuration of Hanja

- 상형문자 : 물체의 형태를 모방하여 간략화한 문자로 그 자신이 각각 原義(원뜻)와 音을 지니고 있을 뿐만 아니라 다른 글자의 변, 冠등에 쓰여 새로운 문자를 구성한다.
  (예)艸 → ++ , 𠂆 + 丨 → 竹
- 지사문자 : 象形을 기본으로 하여 의미에 따라 자획을 증감시켜 만든 文字
  (예)朝 → 旦
- 회의문자 : 旣成의 문자를 複合하여 音에 관계없이 뜻만을 취하여 만든 文字
  (예)目 + 手 → 看 , 力 + 田 → 男
- 형성문자 : 두개 이상의 旣成文字를 합하여 만든 점은 회의문자와 동일하나 반은 의미를 나타내고, 반은 音을 표시한 문자
  (예)門 + 口 = 問 , 亡 + 女 = 妄
- 전주문자 : 어떤 자의 본뜻을 확대시켜 새로운 의미 내용을 도출하는 문자
  (예)道 → 道理, 道德 樂 → 音樂, 娛樂
- 가차문자 : 原字의 音을 빌려서 그와 同音의 다른 의미를 가지는 문자
  (예)皮革 → 變革

표현되는 내용이 복잡해지면서 문자의 형도 이미 사용된 문자의 개개의 형이 조합되기도 하고 원래 하나의 형이던 것이 나누어지는 등 복잡한 형으로 발전하였다.

이렇듯 한자는 도형적 계층성을 갖는다는 것을 알 수가 있다. 한자는 자종이 광대하고 구조가 복잡하며, 유사문자가 많이 존재하기 때문에 구별하는 데에도 많은 어려움이 따른다. 그래서 이러한 문제점을 극복하기 위해서, 한자의 구성 및 구조상의 특징들을 살펴본 결과 한자패턴이 부분패턴(부수, radical)을 서로 공통으로 갖는 점에 착안하여, 이들이 문자상에서 위치하는 영역을 중심으로 특징점을 추출하여 구별을 하는 방법을 사용하였다.

본 절에서는 한자의 도형적 계층성에 의거해 그림 6과 같이 분할할 수 없는 구조와 분할할 수 있는 구조로서 좌우 구조, 상하 구조, 받침 구조, 둘러싼 구조 등으로 생각할 수 있지만 이외에도 복잡한 구조를 이루고 있는 문자도 있기 때문에 문자마다 변과 부수등이 위차는 장소에 따라 조금씩 변하는 경우가 있다.

그림 6. 한자의 형태들
Fig. 6. Various type of Hanja

표 4. 한글과 한자의 구조 특성
Table 4. Structural property of Hangeul and Hanja

	한 글	한 자
기둥 시작점	1, 2, 5, 6 형식은 모두 중심의 우편에 위치한다	중심 부분에 위치할 경우가 많다
보 시작점	3, 4 형식은 모두 중심의 아래편에 위치한다	중심 부분에 위치할 경우가 많다
긴 가로선 (긴 보)	3, 4 형식의 보의 길이가 문자의 가로 크기와 같으며 중심의 아래편에 위치한다	문자 중심의 위편에 위치할 경우가 많다
긴 세로선 (긴기둥)	1 형식의 기둥의 길이가 문자의 세로 크기와 같으며 중심의 우편에 위치한다	문자 중심의 좌편에 위치할 경우가 많다

한자의 구조적인 특성을 살펴보면 문자의 중심을 기준으로 쓰여진 글자가 많으며 한글과 달리 시작점이 중심이나 좌측에 위치하는 문자가 많다. 또한 문자를 좌에서 우로 그리고 상에서 하로 스캔할 경우 문자가 여러 개의 블럭으로 나누어지는 문자가 많다. 표 4에는 한글과 한자의 구조적 특성을 나타내었다.

Ⅳ. 한글과 한자의 구별

문서에 사용되는 문자는 한글, 한자, 영문자, 숫자, 특수문자로 구별할 수 있는데 본 연구에서는 실제 문자 크기로 블럭화된 문자를 대상으로 한글과 한자로 구별한다.

본장에서는 한자에 대하여 구조특성을 분석하고 한자의 구조 특성을 가지고 있으며 한자로 구별하고, 한자의 구조특성으로 구별되지 않는 문자중에서, 한글의 구조특성을 갖고 있으면 한글로 구별하고, 나머지는 한자로 구별한다.

4.1 문자의 구별을 위한 한자의 특성

문자의 구별은 구성 성분을 이용한 구별과 구조적 특성을 이용한 구별로 크게 나눌수 있다. 구성 성분을 이용한 구별 방법에는 사선성분의 수, 흑화소의 밀도등과 같이 문자 고유의 특성을 이용하는 것을 말하며, 구조적 특성을 이용한 방법에는 긴 세로선 성분의 위치, 긴 가로선 성분의 위치와 같이 문자의 구조을 이용하여 구별방법을 말한다. 구성성분을 이용한 방법은 한글과 한자의 구별에 있어서 구성 성분이 동일하게 존재하는 경우가 많고, 입력화상의 질에 따라 영향을 많이 받는 단점이 있다. 따라서 본 연구에서는 비교적 잡음의 영향을 작게 받고 다양한 특징을 갖는 구조적 특성을 이용한 방법을 이용하였다.

4.1.1 긴 세로선(긴 기둥)의 위치

각 문자 블럭에서 중심의 우편에 긴 세로선이 존재하면, 한글 구조 특성상 한글 1, 2, 5, 6 형식으로 분류되고 한글의 모음의 위치가 된다. 그러나 한자에서는 긴 세로선이 존재하는 영역은 한글과 달리 블럭의 우편과 좌편에 모두 존재한다. 따라서 그림 7과 같이 긴 세로선이 블럭 중심의 좌편에 존재하면 한자로 구별한다.

4.1.2 긴 가로선(긴 보)의 위치

각 문자 블럭 중심의 하단에 긴 가로선이 존재하면

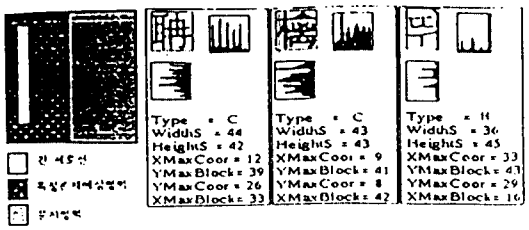


그림 7. 한자의 긴 세로선 위치
Fig. 7. Position of Maximal vertical-line of Hanja

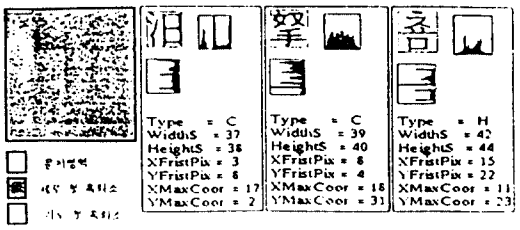


그림 9. 한자의 첫 흑화소 위치
Fig. 9. Position of first black-pixel of Hanja

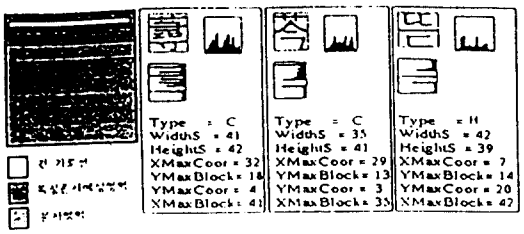


그림 8. 한자의 긴 가로선 위치
Fig. 8. Position of maximal horizontal-line of Hanja

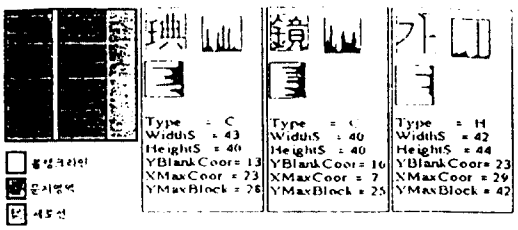


그림 10. 한자의 세로 불연속 블랭크 라인 존재
Fig. 10. Existence of vertical discrete blank line of Hanja

한글의 구조 특성상 한글 3, 4 형식으로 분류된다. 그러나 한자에서 긴 가로선이 존재하는, 영역은 한글과 달리 블럭의 모든 부분에 나타난다. 따라서 그림 8와 같이 긴 가로선이 블럭의 중심 상단에 존재하면 한자로 구별한다.

4.1.3 첫 흑화소의 위치

한자는 중심 위주의 문자인 반면 한글은 모음 중심이기 때문에, 구조 특성상 한글 1, 2, 5, 6 형식 일때 가로의 첫 흑화소의 위치가 중심 오른쪽에 위치하며 모음의 시작점이 된다. 결국 그림 9와 같이 세로 방향의 첫 흑화소 위치가 블럭 중심의 상단에 위치하고, 가로 방향의 첫 흑화소가 블럭 중심의 좌측에 존재하면 한자로 구별한다.

4.1.4 세로의 불연속 블랭크 라인(blank line) 존재 여부

한글은 구조 특성상 세로의 블랭크 라인이 존재하고 세로선(기둥)이 블럭중심의 우편에 존재하는 경우는 한글 1, 2, 5, 6 형식이다. 그림 10과 같이 한자는 세로의 블랭크 라인이 블럭에 존재하고 긴 세로선이 존재하지 않는 것은 한자로 구별한다.

4.1.5 첫 흑화소의 위치와 문자 블럭내의 긴 가로선의 존재 및 위치

가로 방향의 첫번째 흑화소가 블럭의 중심에 위치하며 가로선(보)이 블럭의 상단에 존재하고, 긴 가로선이 블럭의 하단에 존재하는 경우는 한글 3, 4 형식이다. 그림 11과 같이 한자는 첫번째 흑화소가 블럭의 중심에 존재하고, 가로선이 블럭의 상단에 존재할 때 긴 가로선이 블럭의 하단에 존재하지 않는 것이 있으므로 한자로 구별한다.

4.1.6 긴 가로선과 긴 세로선의 위치

한글의 구조 특성상 긴 세로선이 블럭의 우편에 존재하는 경우는 한글 1, 2, 5, 6 형식이고, 긴 가로선이 블럭의 중심 또는 하단에 존재하는 경우는 한글 형식 3, 4 형식이다. 한자는 긴 세로선과 긴 가로선이 블럭에 골고루 분포되어있다. 따라서 그림 12와 같이 긴 세로선과 긴 가로선이 블럭의 좌편과 상단에 존재하면 한자로 구별한다.

4.1.7 긴 세로선과 짧은 세로선의 관계

한글의 구조 특성상 긴 세로선이 블럭 중심의 우편에 존재하고 그 뒤로는 세로선이 나타나지 않는 경우

는 한글 1, 2, 5, 6 형식이며, 블럭 중심의 좌편과 우편에 골고루 존재하며 긴 가로선이 문자폭과 같은 경우는 한글 3, 4 형식이다. 그림 13과 같이 한자는 한글과 달리 세로선중 최대 길이를 갖는 값이 나타나고, 다음에 긴 세로선 보다 작은 짧은 세로선이 위치할 때 긴 가로선 중 최대 길이를 갖는 값이 문자폭보다 작은 것이 있으므로 한자로 구별한다.

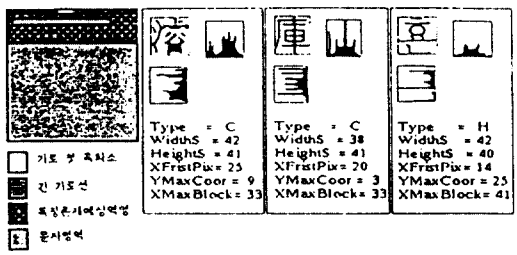


그림 11. 첫 흑화소와 긴 가로선의 존재

Fig. 11. Position of first black-pixel and maximal horizontal of Hanja

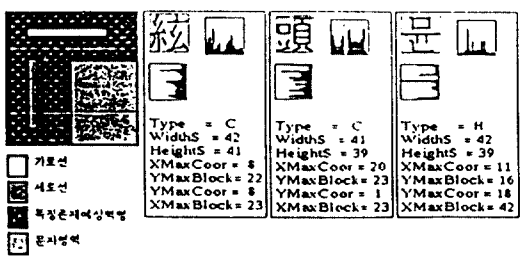


그림 12. 긴 가로선과 긴 세로선 위치

Fig. 12. Position of maximal vertical-line and horizontal-line of Hanja

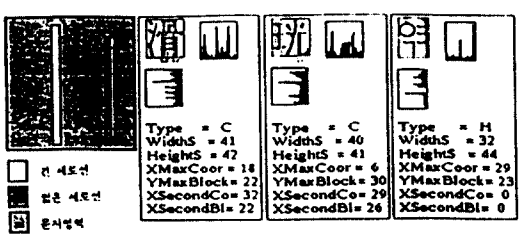


그림 13. 긴 세로선과 짧은 세로선의 관계

Fig. 13. Relation of long and short vertical-line

4.1.8 긴 세로선과 가로 줄기의 관계

한글의 구조 특성상 긴 세로선은 종모음이 되므로 긴 세로선과 가로줄기가 우측이나 좌측으로 나타나고 교차하지 않는다. 그림 14와 같이 한자는 긴 세로선을 기준으로 가로줄기가 교차하므로 한자로 구별한다.

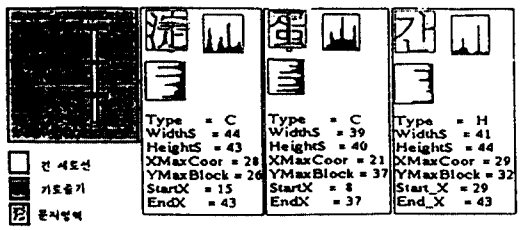


그림 14. 긴 세로선과 교차하는 가로줄기

Fig. 14. Relation of max. vertical-line and horizontal-stroke cross

4.1.9 긴 가로선과 짧은 세로선 및 긴 세로선과의 관계

한글은 구조 특성상 긴 가로선은 횡모음이 되므로 긴 가로선과 기둥의 위쪽이나 밑으로 하나만 존재하고 교차하지 않는다. 그림 15와 같이 한자는 긴 가로선을 기준으로 기둥이 교차하므로 한자로 구별한다.

본 논문에서는 문자의 크기가 비슷하고 복잡도가 비슷한 한글과 한자를 대상으로 하였으며 위에서 논한 9가지의 특징 파라메터로서 구별이 되어진다. 본 논문에서는 한글과 한자의 구별에 우선 순위를 주었으며 특수문자나 영숫자 등은 한자의 종류로 구별하였다. 그림 16에는 본 논문에서 수행한 한글과 한자의 구별에 대한 흐름도를 나타내었다.

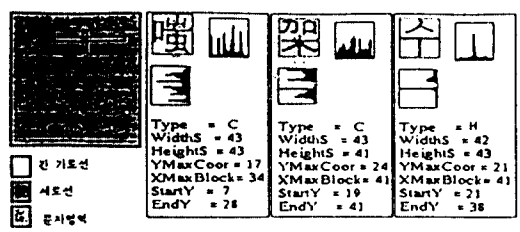


그림 15. 한자의 긴 가로선과 교차하는 세로선의 관계

Fig. 15. Relation of max. horizontal and vertical-line cross of Hanja

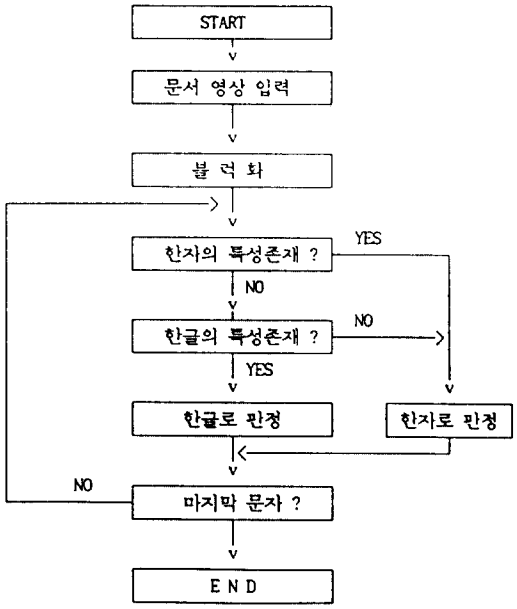


그림 16. 한글과 한자 구별 흐름도
Fig. 16. Classification flow chart of Hangeul and Hanja

4.2 블럭화

본 연구에서는 문자 영역에서 대하여 문자 부분의 각각의 위치를 알기 위하여 블럭화를 행하였다. 블럭화 알고리즘은 오인권의 「Down-up」[5]을 이용하였고 가로쓰기 형태의 문서에서는 블럭화가 우수하게 이루어졌고 세로쓰기 문서에서는 문자사이의 간격이 좁아 블럭화가 약간 불안한 면이 있기는 하지만, 대상문서가 주로 가로쓰기 형태의 문서이므로 별무리 없이 잘 적용되었으며, 그림 17는 블럭화된 데이타의 예이다.

4.3 입력된 문자에 대한 정의

문자는 다양한 크기로 입력되므로, 그 크기에 관계없이 특징점을 적용하기 위하여 문자의 기준을 정해야 한다. 그래서 모든 입력 문자는 그림 18와 같이 횡폭과 종폭은 각각 최대 흑화소 크기를 100%로 정하도록 한다.

또 각 문자의 구별을 위해서는 특징점을 찾기위해 흑화소의 정보가 필요하므로 알고리즘에 정보테이블과 변수를 작성하여 이용할 수 있도록 하였다. 그림 19에 정보 테이블을 나타내었고 표 5는 변수를 나타내었다.

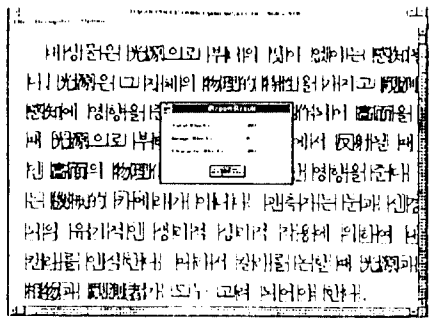


그림 17. 블럭화된 데이타의 예
Fig. 17. Example of blocked data

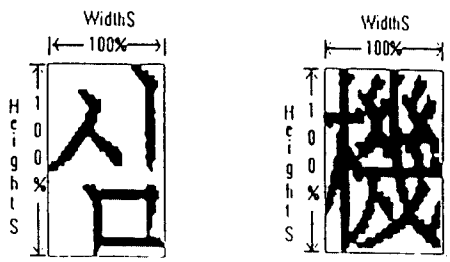


그림 18. 문자에 대한 정의
Fig. 18. Definition of input character

표 5. 변수 테이블
Table 5. Variable table

MaxBlock	: 세로선(기둥)값. x, y는 블럭의 좌표 예)X_TABLE[x]. End_y[X_TABLE[x]. Block_line_num] - X_TABLE[x]. Start_y[X_TABLE[x]. Block_line_num] Y_TABLE[y]. End_x[Y_TABLE[y]. Block_line_num] - Y_TABLE[y]. Start_x[Y_TABLE[y]. Block_line_num]
YMaxBlock	: 세로선(기둥)의 중에서 가장 큰성분
XMaxCoor	: 세로선(기둥)값이 위치하는 X좌표
YFirstPoint	: 세로선(기둥)의 첫번째 흑화소의 Y좌표
BlankLineFlag	: 연속되지 않는 세로선
MFlag	: 세로선(기둥)이 블럭의 중심의 우편에 존재하는지 체크 존재하면 ON, 존재하지 않으면 OFF
TopY	: YMaxBlock 상단의 Y의 좌표
BottomY	: YMaxBlock 하단의 Y의 좌표
XMaxBlock	: 가로선(보)의 중에서 가장 큰성분

YMaxCoor	: 가로선(보)값이 위치하는 Y좌표
MFirstPoint	: 가로선(보)의 첫번째 흑화소의 X좌표
YFlag	: 가로선(보)이 블럭의 중심의 상단에 존재하는 체크 존재하면 ON, 존재하지 않으면 OFF
LeftX	: XMaxBlock 좌편의 X의 좌표
RightX	: XMaxBlock 우편의 X의 좌표
CenterFlag	: 첫번째 흑화소가 블럭의 중심에 위치하는지 체크 위치하면 ON, 위치하지 않으면 OFF

● X 방향에 대한 정보테이블

```
struct X {
    int Block_dot_num:   : 라인의 흑화소 수
    int Block_line_num:  : 라인의 시작점,끝점 갯수
    int Start_y[10]:     : 라인 시작
    int End_y[10]:       : 라인 끝
} X_TABLE[1000]
```

● Y 방향에 대한 정보테이블

```
struct Y {
    int Block_dot_num:   : 라인의 흑화소 수
    int Block_line_num:  : 라인의 시작점,끝점 갯수
    int Start_x[10]:     : 라인 시작
    int End_x[10]:       : 라인 끝
} Y_TABLE[1000]
```

그림 19. 테이블 구성 및 예
Fig. 19. Table configuration and example

4.4 한글·한자 구별 알고리즘

위에서의 조사 결과, 한자의 구별 특징이 명확하므로, 한자의 특성을 먼저 조사하여 한자의 특성을 가진 문자를 한자로 구별하고, 제외된 모든 문자에 대하여 한글의 특성을 조사한다. 이때 한글의 특성을 가지고 있는 것은 한글로 구별하고 한글의 특성을 가지고 있지 않으면 한자로 구별하였다. 아래 알고리즘에서 %는 구별의 정도를 높여주는 것으로 본 연구에서 제시한 경험적 수치이다.

단계 1. 문서를 세로 방향으로 스캔하면서 X 방향의 정보테이블을 작성하면서 MaxBlock 값이 Y_length의 75% 보다 크고 블럭의 중심의 좌편에 존재하면 한자로 구별한후 스캔을 정지한다.
한자로 구별하지 못하면 스캔을 계속 하면서 X 방향의 정보 테이블을 작성하고, YMaxBlock, XMaxCoor, BlankLineFlag, XFirstPoint, TopY, BottomY 값을 구한다.

단계 2. 단계 1에서 구한 BlankLineFlag가 ON이고 XMaxCoor이 블럭의 우측에 존재하며 YMaxBlock의 값이 Y_length의 57% 작으면 한자로 구별한다.

단계 3. 문서를 가로 방향으로 스캔하면서 Y 방향의 정보테이블을 작성하면서 MaxBlock 값이 X_length의 89% 보다 이상이고, Y_length의 27% 위쪽에 존재하면 한자로 구별한후 스캔을 정지한다.
한자로 구별하지 못하면 스캔을 계속 하면서 Y 방향의 정보 테이블을 작성하고, XMaxBlock, YMaxCoor, XMaxBlock, YmaxCoor, YFirstPoint, YFlag, LeftX, RightX, CenterFlag값을 구한다.

단계 4. 단계 3에서 구한 CenterFlag가 ON이고 MaxBlock이 X_length이 70% 이상이고, 블럭의 30% 위쪽에 존재하면 한자로 구별한다.

단계 5. 단계 1에서 구한 YFirstPoint 좌표가 Y_length의 30% 앞쪽에 있고, 단계 3에서 구한 XFirstPoint 좌표가 X_length의 30% 앞쪽에 존재하면 한자로 분리한다.

단계 6. 단계 1에서 구한 MFlag가 OFF이고 YMaxBlock이 Y_length의 55% 보다 이상이면, 단계 3에서 구한 YFlag가 OFF이고 XMaxBlock이 X_length의 80% 이상이면 한자로 구별한다.

단계 7. 단계 1에서 구한 XMaxCoor 존재하고 다음으로 XMaxCoor 보다 작은 세로선이 존재하며 단계 3에서 구한 XMaxBlock이 X_length의 95% 작으면 한자로 분리한다.

단계 8. 단계 1에서 구한 YMaxBlock의 Y좌표 TopY와 BottomY 사이에 XMaxCoor을 기준으로 단계 2에서 구한 MaxBlock의 시작점과 끝점이 존재하면 한자로 구별한다.

단계 9. 단계 2에서 구한 XMaxBlock의 X좌표 LeftX와 RightY 사이에 YMaxCoor을 기준으로 단계 1에서 구한 MaxBlock의 시작점과 끝점이 존재하면 한자로 구별한다.

단계 10. 앞의 단계에서 한자로 구별되지 못한 문자에 대하여 단계 1에서 구한 XMaxCoor이 블

럭이 우편에 존재하고, YMaxBlock의 크기가 Y_length의 60% 이상 존재하면 한글의 종모음이고, 단계 3에서 구한 YMaxCoor의 블럭의 40% 아래에 존재하며 XMaxBlock이 크기가 X_length의 90% 이상 존재 하면 한글의 횡모음이다. 그러므로 위의 특성을 가지고 있는 문자는 한글로 분리하고 이에 속하지 않으면 한자로 구별한다. 마지막 문자가 될때 까지 단계 1부터 단계 10을 반복한다.

V. 실험 및 고찰

5.1 실험 시스템

본 연구에서 실험에 사용된 문서는 전자공학회지와 일반문서를 대상으로 하였으며 시스템은 Intel-80387 수치 연산 보조 프로세서를 정착한 IBM 386-PC 호환 컴퓨터에서 사용하였고, 실험환경은 MS-WINDOW 3.1의 SDK(SoftWare Development Kit)를 사용하여 구현한다.

문서 영상은 Hewlett Packard의 이미지 스캐너(Image scanner)를 이용하여 인치당 300 화소의 해상도로 입력받았다. 그림 20에 사용한 실험 시스템의 구성도를 보였다.

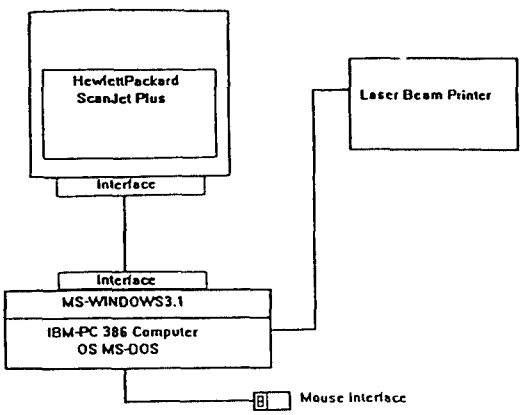
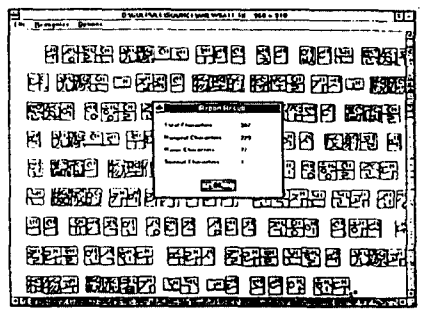


그림 20. 실험 시스템의 구성도
Fig. 20. Block diagram of experimental system

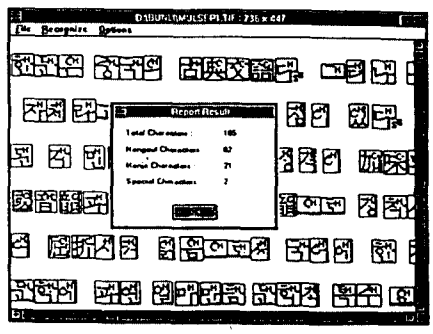
5.2 실험 결과

한글과 한자를 구별하기 위하여 먼저 한글과 한자, 그리고 명조체와 고딕체 각각에 대하여 실험을 행한 후에, 한글과 한자가 섞여 있는 입력 영상에 대하여 구별 실험을 행하였다. 대상 문자는 KS C 5601에 있는 2,350자의 한글과 4,888자의 한자에 대하여 실험을 행하였다. 각 문자에 대하여 실제 문자 크기로 블럭화하여 한글은 'H', 특수 문자는 'S', 한자는 'C'로 표시하였다.

한글 명조체에 대해서는 99.4%, 고딕체에 대해서는 98.8%의 구별율을 얻었고, 한자 명조체에 대해서는 91.9%, 고딕체에 대해서는 94.4%의 구별율을 얻었으며, 문자 자당의 처리 속도는 1초에 16자 정도를 보였다. 그림 21에 최종 구별된 결과를 보였으며, 표 6에 구별되지 않은 문자의 예를 보였다.



(a) 공학회지
(a) Academic society's paper



(b) 일반문서
(b) Document

그림 21. 처리 결과
Fig. 21. Result of processing

표 6. 오분류된 데이타
Table 6. Data of ill-classification

한글		한자	
명조	고딕	명조	고딕
[illegible]	[illegible]	[illegible]	[illegible]
합계 12	27	400	273

5.3 고 찰

본 연구는 인쇄체 문서를 자동 인식하기 위한 전처리로서 표, 그림 등의 그림영역(graphic)을 제외한 문자영역(text)에서 문자만을 입력받아, 문자 자체가 가지는 정보를 이용하여 한글과 한자를 구별하는 연구를 하였다. 입력 화상에서 문자의 블럭화는 매우 충실하게 되었으며, 문자구별은 이 블럭의 정보를 이용하여 행하였고, 제안한 한자 및 한글의 특징을 적용하면 문자의 크기에 제한을 받지 않고 다양한 크기를 문자에 적용할 수 있다.

오분류가 발생하는 주된 원인을 살펴보면 입력된 문자의 품질이 나쁜 경우와 문자의 구성 요소가 완벽히 구별되지 못하거나 또는 한글과 한자의 구조적인 특성을 공통적으로 가지고 있는 문자가 존재하는데 있다. 특징점을 추출하여 한글과 한자를 구별하는데 있어서 글자체에 따라 구별이 약간의 차이를 나타내었다. 그림 22와 같이 같은 글자인경우 글자체에 따라 문자의 두께, 위치에 영향에 의하여 특징점이 달라져 오분류 되는 것을 알 수 있었다. 본 연구에서 실험 대상으로 일반문서와 신문문서(가로쓰기)를 사용하였으며, 일반문서에서는 문자와 문자 사이가 일정한 간격으로 띠어져 있으므로 잡음의 영향이 적어 구별률이 높았다. 그러나 신문 문자에서는 적은 공간에 많은 내용을 전달하여야 하므로 문자와 문자 사이가 좁고, 재질에 있어서 일반 문서와 달리 입력 화상에 잡음의 영향을 많이 받아 구별률이 약간 저조하였다.

따라서 명조체와 고딕체를 제외한 다양한 글자체에 대해서도 적용할 수 있는 특징에관한 연구와 오분류된 문자에 대하여 더 개선된 구별 연구가 필요할 것으로 본다.

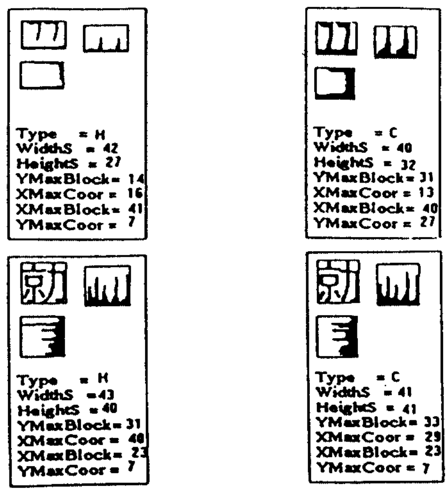


그림 22. 글자체에 따른 특징점의 차이
Fig. 22. Difference of feature-point for font

Ⅵ. 결 론

본 논문에서는 한글과 한자의 구조적 특성을 이용하여 한글과 한자를 구별하는 연구를 하였다. 한글과 한자의 구별에 있어서 한자와 한글의 특성을 반영하여 실험을 행한 결과 구별률과 처리속도에서 우수한 결과를 얻었다. 문자의 구별에 관하여 한글과 한자의 특성을 충분히 이용하여 신문(가로쓰기)를 대상으로 92%, 잡지와 교과서는 96%, 학회지를 대상으로 98%, 아래아한글 문서를 대상으로 98% 문자 구별률을 얻었다. 본 논문은 문서 자동 입력 시스템을 위한 전처리 단계로서 한글과 한자 글자체의 기본이 되고 고딕체, 명조체를 대상으로 하였고, 다양한 크기의 문자에 대하여 좋은 결과와 가능성을 발견하였다.

참 고 문 헌

1. 김형훈, 이성환, 김진형, "한국 신문 영상의 구조 분석을 통한 기사의 추출," 정보과학회논문지, 제15권, 제5호, pp.392-404, 1988. 10.
2. Dacheng Wang and Sargur N. Srihari, "Classification of Newspaper Image Blocks Using Texture Analysis," Computer Vision, Graphics and Image Processing 47, pp.327-352. 1989.
3. 남궁재찬, 유황빈, 남궁연, "한국어 문서로 부터 문자 분리 및 도형추출에 관한 연구," 대한 전자공학회 논문지, Vol.25, No.9, pp.73-83, 1988.
4. 신현관, "문서의 영역 분리와 레이아웃 정보 추출에 관한 연구," 광운대학교 대학원 석사학위 논문, 1992.
5. 오인권, "영문과 혼합된 한글 문서에서의 문자 및 특수 문자추출에 관한연구," 광운대학교 대학원 석사학위 논문, 1988.
6. J. K. Lee, "Korean Character Display Variable Combination and its Recogition by Decomposition Method," Ph. D. dissertation in Keio Univ., Japan, 1972.
7. 남궁재찬, "Index-window 알고리즘에 의한 tern의 부분 분리와 인식에 관한 연구," 인하교 박사 학위 논문, 1982.
8. 이주근, 남궁재찬, 김영건, "한글 pattern에서 bpattern 분리와 인식에 관한 연구," 대한 전학회논문지, Vol.18, No.3, pp.1-8, 1981.
9. 남궁재찬, "Font 개발을 위한 한글특성 분석어 한 연구," 광운대학교 논문집, Vol.18, pp.149-1989.
10. Yasuaki Nakano and Kazuo Nakata, "Rec nition of Chinese Characters using Periph Distributions and their Amplitude Spect 日本전자통신학회지 논문집, Vol.56-D No.3, 47-59, 1973.
11. 김학성, 레테링 디자인, 조형사, pp.60, 1988.
12. 이승형, "문서인식을 위한 한글과 한자의 구별 한글의 형식 분류에 관한 연구," 광운대학교 대원 석사학위 논문, 1990.
13. 김진평, 한글의 글자 표현, 미진사, 1989.

沈 相 完(Sang Ouan Sim) 정회원

1990년 : 광운대학교 전자계산기공학과 졸업(공학사)
1993년 : 광운대학교 산업정보대학원 전산기과 졸업(공학석사)
1992년~현재 : 유니텍 부설연구소

南宮在贊(Jae Chan NamKung) 정회원

1970년 : 인하대학교 전기공학과 졸업(공학사)
1976년 : 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 : 일본 TOHOKU대학 객원 교수
1979년~현재 : 광운대학교 전자계산기공학과 교수
※관심분야 : 패턴인식, 컴퓨터 비젼, 인공지능

李 成 範(Sung Bum Lee) 정회원

1973년 : 한양대학교 전기과공학과 졸업(공학사)
1981년 : 동국대학교 대학원 전기공학과 졸업(공학석사)
1989년 : 광운대학교 대학원 전자계산기공학과 재학
1981년~현재 : 대유공업전문대학 전기과 부교수
※관심분야 : 패턴인식, 인공지능, 컴퓨터 비젼