DBpia

論 文

한글 Shape 문자 Pattern에서의 구조적 정보를 이용한 형식분류와 인식에 관한 연구

正會員 田 鍾 益* 正會員 趙 鎔 周** 正會員 南宮 在 贊**

A Study on Type Classification and Recognition Using Structural Information in Character Pattern of HANGEUL Shape

Jong Ik CHEON,* Yong Joo CHO,** Jae Chan NAMKUNG** *Regular Members*

要 約 본 논문은 한글 문자 패턴을 인식하기 위하여 쉐이프 패턴을 대상으로 한글의 구조적 정보를 이용한 새로운 인식 방법에 대해서 연구하였다.

먼저, 입력된 데이타에 대하여 문자의 위치를 알기 위하여 문자 영역의 블럭화를 행하였고, 둘째로, 한글의 무게 중심을 이용한 중모음의 유무를 결정하고, 정보음과 짧은 기둥의 위치와 길이 조사 및 종성의 존재 여부를 조사하여 한글이 가지는 6 가지의 기본 구조로 형식 분류를 하였으며, 마지막으로 분류된 형식에 따라 각각의 자소에 대해 유클리드 거리(Uclid' distance)에 의한 템플레이트 매칭(template matching)을 실행하여 인식을 행하였다.

본 연구는 실험을 통하여 총 2350자의 문자에 대하여 98.3%의 형식 분류율과 95.2%의 한글 인식율을 얻었다.

ABSTRACT In this paper, we studied on new method of recognition using structural information to recognize character pattern in original shape of Hangeul.

First, for the purpose of knowing location of character in input image, it processed Making block. Second, after we investigated, whether vertical vowel existed or not in character image accordingly the center of gravity of Hangeul, each character was classified into 6 Type of Hangeul by searching location and length for horizontal vowel and short pole.

Last, we processed it by means of template matching which calculate Uclid's distance on each Jaso in accordance to type classified.

This paper made an experiment on 2350 characters and obtained 98.3% classifing rate and 95.2% recognizing rate.

I. 서 론

최근 정보 처리 환경은 사무 자동화와 사용자 인터페이스의 급격한 요구로 현대의 정보 사회는 많은 분야에서 다양한 연구를 행하고 있다. 패턴 인식 분야는 문자 인식, 음성 인식등 다양한 매체에서 많은 연구가 행해지고 있으며, 특히 패턴인식에 관한 연구 분야 중에서 문자 인식분야는 그 나라의 사회와 문화를 고려하여 많은 연구를 해오고 있다.

본 논문은 우리 고유의 한글 문자 패턴을 인식하기 위한 연구로써 기존의 방법과는 달리, 쉐이프(shape) 자체를 대상으로 한글 고유의 6가지 형식을 자동 분류하고, 분류된 형식 문자에 대해 정확한 인식 방법을 구현하는데 목적을 둔다.

한글 인식에 관한 연구는 1960년대를 시발로 하여 현재까지 꾸준히 연구된 분야 중의 하나이다. 한글은 한글 문자 패턴이 가지는 고유의 조합 법칙을 고려한 심용적인 문자로서 6가지 형식을 가지고 있다.

한글의 인식 방법은 크게 두가지 분류로 연구되어 왔다.

첫 번째는 형식 분류를 한 후 각 자소에 대한 인식을 하는 것이고, 두 번째로 한 문자에 대해

* (株)韓國通信
Korea Telecom
** 光云大學校 電子計算學科
Dept. of Computer Science, Kwangwoon University.
論文番號 : 91-17(接受1990. 11. 23)

www.dbpia.co.kr

곧 바로 인식하는 방법이 있는데, 본 연구에서는 첫 번째 방법을 사용하였다.

보통 문자 인식에 있어서 접근 방식으로는 결정론적 방법, 구문론적 방법, 구조적 방법등이 있다.

결정론적 방법(decision theoritical method)은 문자 패턴을 어떤 특징량 또는 벡터량으로 변환한 후 특징 공간으로 나누어 벡터 정보의 거리값에 의해서 인식을 하는 방법으로 보통 자종이 많은 필기체 문자 인식에서 많이 사용된다. 다음으로 구문론적 방법(syntatic method)은 인식하고자 하는 문자 패턴에서 문자의 특징점과 스트로크(stroke)등을 프리미티브(primitive)로 변환한 후 여러가지의 문법(Grammar)을 이용하는 방법으로서 획 성분의 구분이 명확한 문자 인식 등에서 많이 사용된다. 마지막으로 구조적인 방법(structural method)은 문자의 구조적 정보를 이용하여 인식하는 방법으로서 현재의 문자 인식 방법에 많이 사용될 뿐만 아니라 다양한 패턴 정보의 인식에서도 많이 사용되는 방법이다. 본 연구에서는 구조적인 방법을 이용하여 한글 문자 패턴을 인식하였다.

Ⅱ. 한글의 구조분석 및 전처리

2.1 한글의 구조 분석 및 특성

문자 인식 시스템을 설계하기 위해서는 각 문자의 구조적 특성을 분석하고 각각의 특징에 따라 각 문자의 인식 시스템을 설계해야 한다. 따라서 본 장에서는 먼저 한글의 용어에 대해서 알아본 다음, 한글의 기본적 구조를 분석하고 구조적 특성에 대해 알아본다.

2.1.1 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. 인식에 있어서는 글자의 꼴이 어떤 형태인가, 즉 서체가 어느 것인가가 매우 중요하다.

본 연구에서는 현존하는 글자체중 가장 많이 이용되고 있는 명조체 글자체를 대상으로 하였고, 1979년 한글 시각 문제 연구 단체인 글꼴모임[8]에서 정의한 기본 줄기 용어를 사용하였다. 그림 1은 글꼴 모임에서 정의한 기본 줄기와 용어를 보였다.

그림 1. 한글의 기본 줄기와 용어
Fig. 1. Primary stem and vocabulary of Hangeul

2.1.2 한글의 구조 분석

한글의 자소들이 모여 구성된 조합 문자로서 기본 자모(닥자음 14자, 닥모음 10자)가 혼합되어 초성, 중성, 종성을 구성한다. 한글의 구성 요소는 표 1과 같다.

한글의 6가지 형식은 이주논[9]이 제안한 정의로서, 초성, 중성, 종성의 배치에 따라 6가지 형식으로 구분이 되어져있다. 본 연구에서는 6형식 정의를 도입하여 인식에 이용하였다. 한글은 한자와 같이 조합문자 이지만, 한자의 무게 중심이 사각형 내에 골고루 분포되어 있는것과는 달리 한글의 무게 중심은 모음, 즉 종모음과 횡모음에 분포되어 있는것이 특징이다. 그림 2에서는 한글의 6형식 구조를 나타내고 있다.

www.dbpia.co.kr

표1. 한글의 구성 요소
Table 1. The constructural element of Hangeul

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

본 연구에서 실험 대상으로 삼은 문자는 삼보 레이저 빔 프린터(LBP) 명조체 폰트(font)이며, KS C5601표준에서 정의한 2350자를 대상으로 본 연구의 실험을 하였다. 위의 2350자를 그림 2에 따라 형식별로 분류한 예를 표 2에 보인다. 표 2에서 알 수 있듯이 2형식인 가로모임 받침글자가 거의 50%에 가까운 점유율로 많이 쓰이고 3형식인 세로모임 글자는 88자에 지나지 않음을 알 수 있다.

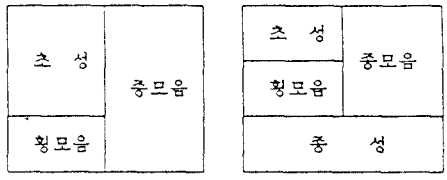
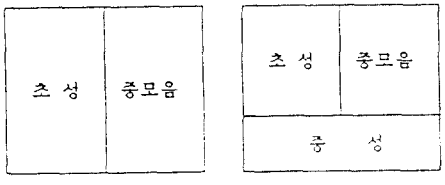
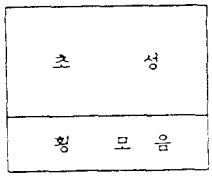


그림 2. 한글의 6형식
Fig. 2. 6 Types of Hangeul





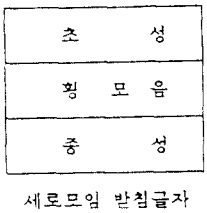


표 2. KS C5601의 형식별 분포
Table 2. Distribution for each type in KS C5601

형식	1	2	3	4	5	6	총
개수	150	1083	88	577	105	347	2350

2.1.3 한글의 구조적 특성

한글은 6 형식에 따라 자음과 모음의 위치가 정해진 문자이며, 매우 과학적인 문자이나 구조적으로 유사성이 매우 높아 인식을 하는데는 어려움이 많다.

본 연구에서는 문자를 인식하기 위해 구조적인 방법을 이용한다. 위에서 언급 되었지만 한글은 모음 중심 문자 즉 종모음과 횡모음의 형태에 따라 글자꼴의 형태가 민감하게 변하기 때문에, 본 논문에서는 표 3과 같이 6가지 특징을 이용하

www.dbpia.co.kr

표 3. 한글의 특징
Table 3. Characteristics of Hangeul

	특 징
1	한글에서 각 자소들 중 수평 방향으로 흰 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 따른 크기변화가 없다.

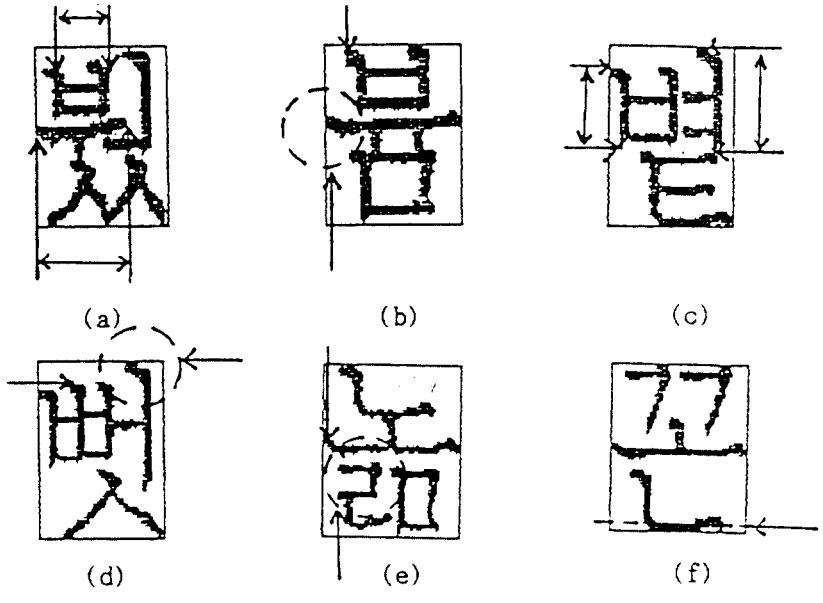


그림 3. 한글의 구조 특징 예
Fig. 3. Example of Hangeul structure charatenstics

여 인식을 용이하게 하였다. 그림 3은 한글의 특징에 의한 예이다.

2.1.4 한글 모음의 결합 특성

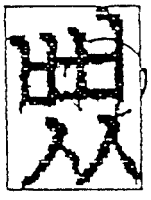
한글의 모음은 크게 횡모음과 종모음으로 나눌 수 있는데, 이들은 기둥, 곁줄기, 보 그리고 짧은 기둥으로 구성되어 있으며, 이것은 곁줄기와 짧은 기둥에 따라 모음의 성분들이 결합된다.

가. 종모음의 결합
- 수직 성분(기둥) : 1 혹은 2개
- 수평 성분(곁줄기) : 2개 이하
- 횡모음 존재 시 : 수평 성분(곁줄기)은 1개 이하

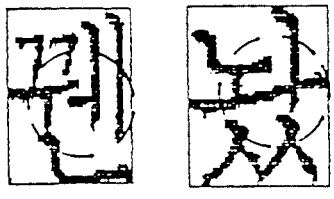
나. 횡모음의 결합
- 수평 성분(보) : 1개
- 수직 성분(짧은 기둥) : 2개 이하

www.dbpia.co.kr

— 종모음 존재 시
: 수직 성분(짧은 세로)은 1개 이하
: "ㅗ" 존재시 "ㅏ"혹은 "ㅣ"의 결합된 형태가 온다.
: "ㅜ" 존재시 "ㅓ"혹은 "ㅣ"의 결합된 형태가 온다.
— 그림 4에 예를 보였다.



(a) 종모음 결합 예



(b) 종모음 결합 예

그림 4. 모음 어절 결합 예
Fig. 4. Cohesion of vowel

2.2 전처리

전처리 단계는 입력된 문자 영상을 문자 단위의 영상으로 분리하는 단계로서 한 문자씩 인식하기 위하여 분리하는 단계이다.

전처리에는 양자화, 평활화, 분리화, 정규화 등 여러가지가 있으나 본 연구에서는 인쇄체 문자의 shape을 대상으로 입력시 세선을 노이즈로 정규화는 행하지 않았다.

2.2.1 양자화

문자 도형이나 도면의 화상을 image scanner 등 통하여 입력할때, 주로 3차 이상의 농담 화상으로 변화된다. 이러한 다치 정보를 2치 정보로 변환하는 처리를 양자화라고 한다. 이러한 처리는 레지값(threshold)에 의하여 행해진다.

$$g_{ij} = \begin{cases} f_{ij} \le T & f_{ij} : \text{입력 값}(0,1,\cdots 255) \\ & g_{ij} : \text{출력 값}(0,1) \\ f_{ij} > T & T : \text{레지 값}(\text{Threshold value}) \end{cases}$$

본 연구에서는 문자입력 장치인 image scanner 에서 직접 양자화된 데이타를 사용하였다. 그림 5는 이미지 스캐너에서 입력한 양자화된 데이타로 640×400의 픽셀밀도에 해당한 예이다.

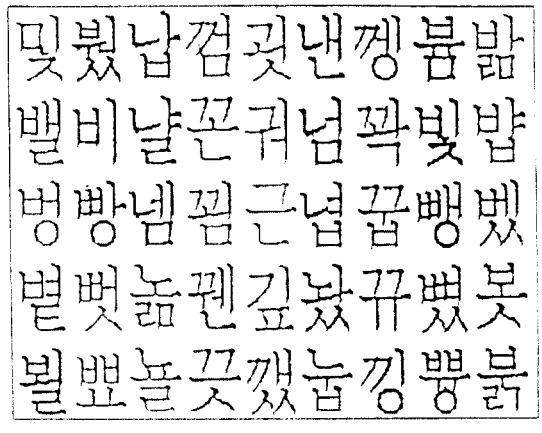


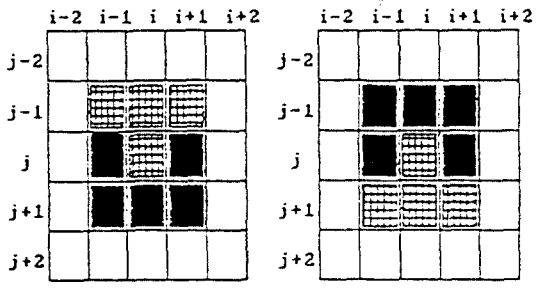
그림 5. 디지타이저된 데이타의 예
Fig. 5. Example of digitized data

2.2.2 평활화

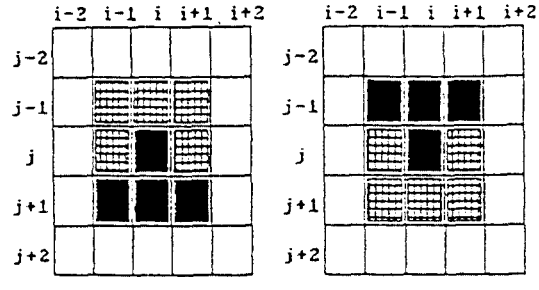
고려한 다치된 영상에 있어서 문자 인식에 바람직하지 않은 잡음을 제거하기 위하여 평활화 처리를 한다. 본 연구에서는 노이즈제거를 위하여 5×5 마스크(그림 6(c))를 사용하였으며 평활화 처리를 위한 마스크는 그림 2(a), 그리고 2(b)는 x축과 y축에 대해 실행하였다. 노이즈제거 마스크에서는 unfill 조건을 만족하는 경우 don't care 영역을 총 1인 화소들은(최대 9개) 제거하고, 평활화 마스크에서는 마스크의 조건이 모두 만족할 때 g(i,j)의 화소가 보정되거나(그림

www.dbpia.co.kr

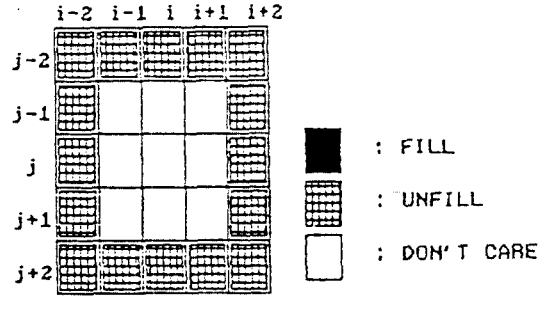
6(a)) 제거된다(그림 6(b)).



(a) 보정



(b) 제거



(c) 고립점 제거

그림 6. 전처리 마스크
Fig. 6. Masks of Preprocessing

2.2.3 블럭화

문서를 이루는 다양한 구성 요소로는 영상(Image), 문자(Text) 그리고 그래픽(Graphic)으로 구성된다. 본 연구에서는 이러한 영역중 문자 부분에 대하여 각각의 문자 위치를 알기 위하여 문자 영역에 대해 블럭화를 행하였다. 블럭화의 알고리즘은 오인귀의 Down up 알고리즘[9]을 개선하여 사용하였다. 그림 7은 블럭화된 데이타의 예이다.

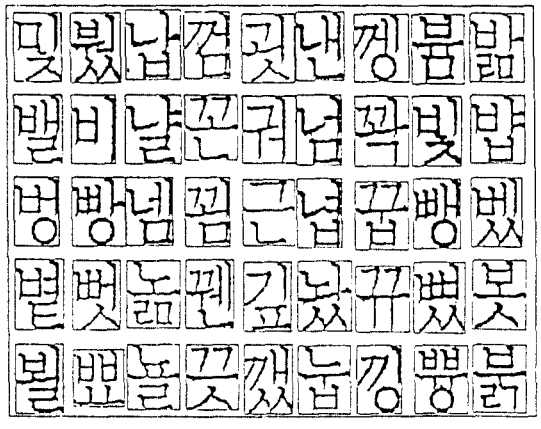


그림 7. 블럭화된 데이타
Fig. 7. Data of Making block

Ⅲ. 한글의 형식 분류

한글의 인식 방법에 있어서 문자에 대한 인식 방법과 한 문자를 자소별로 분리하여 각각의 자소에 대해 인식하는 방법이 있다.

본 연구에서는 자소별로 인식하기 위해 먼저 형식을 결정하는데 한글은 모음을 중심으로 종성과 함께 6가지 형식으로 분류할 수 있다.

본 장에서는 이에 착안하여 횡모음과 종모음을 추출한 다음, 제2장의 모음 결합 특성에 의하여 길쭉기와 짧은 기능을, 다음에 종성을 조사함으로서, 인식을 위한 소분류를 하는 모음 판정 방법과 이에 따른 6형식 식별 알고리즘을 제안한다.

3.1 모음 판정 방법

모음의 존재 여부를 알아내기 위해서 먼저

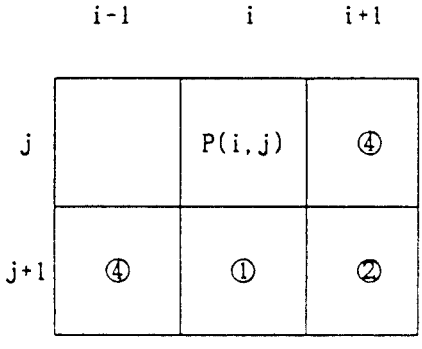
www.dbpia.co.kr

횡모음의 보와 종모음의 기능을 찾았으며, 이때 짧은 기능과 길줄기를 조사하여 모음에 대한 소분류를 행하였다.

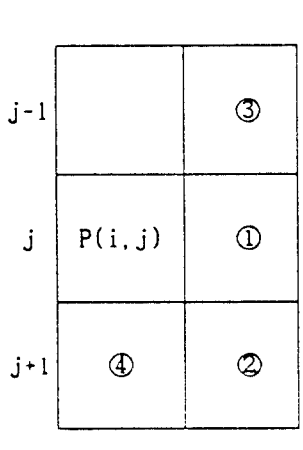
먼저, 보와 기능의 존재 여부를 알아내기 위하여 2×3의 마스크를 적용하였으며 곡률을 조사함으로서 사선이 추출되지 않도록 하였으며, 사선이 아닌 직선 성분을 가진 자음이 추출되지 않도록 하기 위하여 단점을 조사하였다.

3.1.1 2×3 마스크

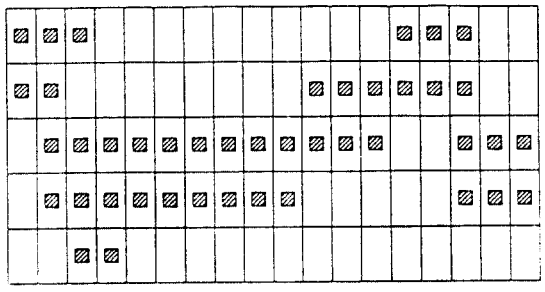
횡모음과 종모음이 존재하는지를 알기 위한



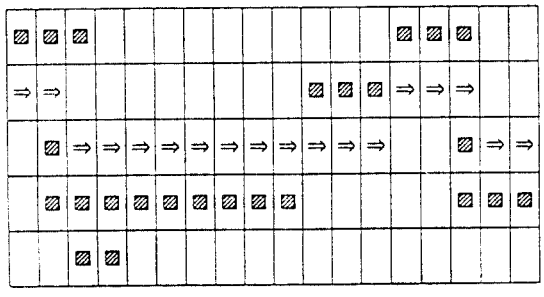
(a) 종모음 추출 시



(b) 횡모음 추출 시



(진행 전)



(진행 후)

(c) 마스크의 진행 과정

그림 8. 2×3 마스크의 진행 예
Fig. 8. Example of 2×3 mask and processing mask

보와 기능을 찾기 위하여 우선 순위에 따라 2×3 마스크의 8방향 중 네개의 방향으로만 진행에 나간다.

그림 8에는 2×3 마스크를 정의 하였는데, 그림 8(a)는 종모음 추출 2×3 마스크와 우선 순위를 보였고, 그림 8(b)는 횡모음 추출 2×3 마스크와 우선 순위를 나타내었다. 그림 8(c)에서는 2×3 마스크를 이용하여 횡모음을 찾아가는 예로서 진행 전과 진행 후를 보였다.

3.1.2 곡률 조사

횡모음중 경사진 횡모음이 있는 것은 5,6 형식에서 대부분 나타난다. 3.1.1의 2×3 마스크를 사용하여 진행해 가면 밋침(ㅅ,ㅈ,ㅊ등), 곡직성분(ㅇ,ㅁ), 횡모음 모두 횡모음의 보로 추출되어 진행이 계속되는데 밋침과 곡직 성분이 횡모음으로 추출되지 않도록 하기위해 (식 3.1)을

www.dbpia.co.kr

이용하여 주기적으로(5 pixel 마다) 곡률을 계산한다. 그림 9에서 현재 진행 점을 c라 하고, a와 b점간의 사이 각을 이전 각, b(x1,y1)와 c(x2,y2)사이 각을 현재 각이라 할 때, 이전 각과 현재 각의 차이각이나 현재 각이 임계각(25도)보다 작으면 진행이 계속되어 보를 추출한다.

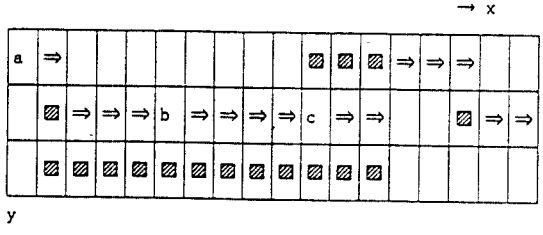


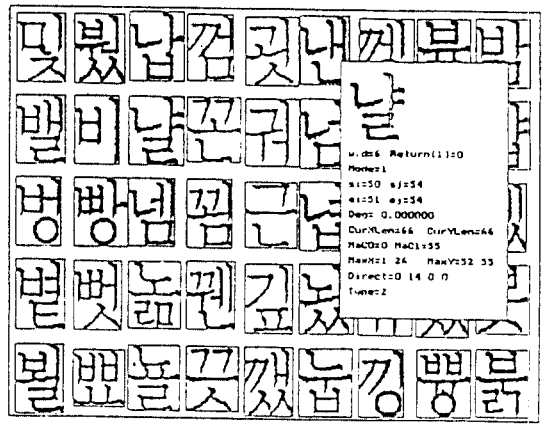
그림 9. b와 c사이의 곡률
Fig. 9. Curvature between b and c

L0=x2−x1, L2=y2−y1, L1=(L0²+L2²)^1/2
A=2tan⁻¹*((L−L0)*(L−L1)/(L*(L−L2)))^1/2−(식 3-1)
단, L0, L1, L2 : 3변의 길이
L=(L0+L1+L2)/2
A : 라디안(radian)

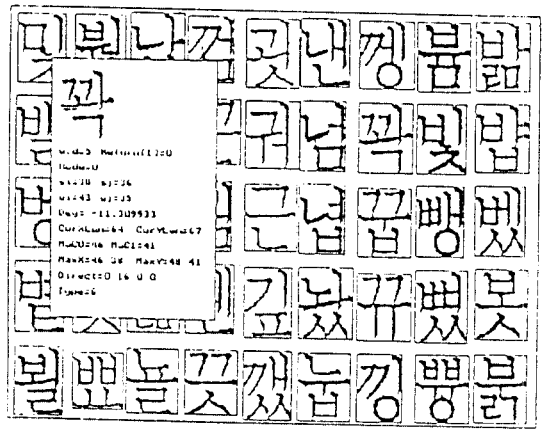
3.1.3. 단점 조사

횡모음의 시작 부분은 단점이다.

단점을 조사하지 않았을 때 "ㄹ", "ㅁ", "ㅂ"과 같은 초성이 횡모음으로 추출되는데 이를 방지하기 위해서 횡모음의 보를 추출해 가면서 평균 문자두께(ChWd)×2 (3.2 알고리즘 참조) 이상의 화소가 y축을 따라(x/5구간 동안) 존재하지 않으면 단점이 존재하는 것으로 하였다. 그림 10에는 위의 특성에 의하여 추출된 종모음과 횡모음의 추출 예를 보였다.



(a) 종모음 추출



(b) 횡모음 추출

그림 10. 종모음과 횡모음의 추출 예
Fig. 10. Extraction of horizontal and vertical vowels

3.2. 형식 분류 알고리즘

한글은 모음을 중심으로 6가지의 형식을 가지고 있으므로, 본 연구에서는 모음을 중심으로 한글의 구조적 특성을 이용하여 형식을 분류하였다.

종모음은 위치 변화가 적으나 종성에 따른 변화가 크고, 횡모음은 위치변화는 많으나 종성에 따른 크기 변화가 없으므로 횡모음의 긴 보부터 조사했다.

자소들 중에서 긴 보가 있으며 3형식이나 4형식이고, 그렇지 않으면 1, 2, 5, 6형식이다. 그러나 긴 보로 추출되는 것 중에서, 긴 보는 짧은 보와 종모음의 긴 줄기가 붙어서(뵈, 왕 등의 경우) 긴 보로 추출되는 경우가 있으므로 긴 기둥(MaxCnt[Y])이 존재하면 5형식이나 6형식이고, 그렇지 않은 것은 3 혹은 4형식이며, 이들에 대하여 종성 성분을 조사하여 6형식과 4형식을 분류한다.

www.dbpia.co.kr

긴 기둥이 있는 것은 1, 2, 5, 6형식이다. 긴 기둥이 블럭의 y축 보다 작으면 2, 6형식이고, 그렇지 않으면 1, 5형식과 2, 6형식(긴 기둥과 종성이 붙는 경우－걱, 밥 등 종성이 "ㄱ", "ㅁ", "ㅂ", "ㅋ"의 경우)이다. 이것을 대상으로 짧은 보(짧은 횡모음)가 존재하는지를 조사하여, 존재하면 5, 6 형식이고, 그렇지 않으면 1, 2형식이다. 이들 중에서 종성을 조사하여, 2형식과 6형식을 분류한다.

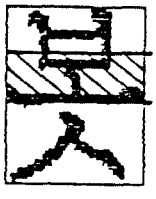
종성 성분을 조사하는 것은 긴 기둥이 블럭의 y축 길이 보다 짧으면 종성이 있는 것으로 판단하며, 그렇지 않은 경우에는 특성 지역의 하변에 연속된 수평 성분이 있으면 종성이 있는 것으로 판단한다. 이와 같이 문자의 특성을 이용하여 6가지 형식으로 분류하였다. 표 4에 알고리즘을, 그림 11에 그에 따른 영역을 보였다. 그림 12에 한글 6형식을 분류하는 알고리즘 흐름도를 보였으며 이 알고리즘에 의해 분류된 결과는 그림 13에 보인다.

표 4. 알고리즘
Table 4. Algorithm

내 용	영역
BLOCK_0] : 블럭의 가장 왼쪽 좌표(x1)	
BLOCK_1] : 블럭의 가장 위쪽 좌표(y1)	
BLOCK[2] : 블럭의 가장 오른쪽 좌표(x2)	
BLOCK[3] : 블럭의 가장 아래쪽 좌표(y2)	
x : BLOCK_2] BLOCK_0]+1 y=BLOCK_3] BLOCK_1]+1	
MaxC_0] : 횡모음의 보를 추적한 길이	
MaxC_1] : 종모음의 기둥을 추적한 길이	
Stem[0] : "ㅏ"의 검출기	
Stem_1] : "ㅓ"의 검출기	
Stem_2] : "ㅗ"의 짧은 기둥	
Stem_3] : "ㅜ"의 짧은 기둥	
ChWd : 영문문자폭에=q : (q a) ^dp q : 정하고 싶 a : 2×2 마스크에서 어문화로 보기 위한 상한값 수	
MaxL[0] : 보 추적시 y축 좌표값이 가장 큰 좌표	
MaxL_1] : 기둥 추적시 x축 좌표값이 가장 큰 좌표	
MinL_0] : 보 추적시 y축 좌표값이 가장 작은 좌표	
MinL[1] : 기둥 추적시 x축 좌표값이 가장 작은 좌표	
PoleCo_0]_100] : 보가 추적되는 y축 좌표	
PoleCo_1]_100] : 기둥이 추적되는 x축 좌표	
1 if(x축 추화소 더함=x) goto 2 else goto 11	
2 (3, 4, 5, 6 형식의 경우) 횡모음 조사	
3 종모음 조사	
4 Stem_2]와 Stem[3] 조사 (223 의거)	(a),(b)
5 if(Stem_2] !=0) Stem_1] 조사	(c)
6 else if(Stem_3] !=0) Stem[0] 조사	(d)
7 if(MaxL_1] Y V C) type=3;	

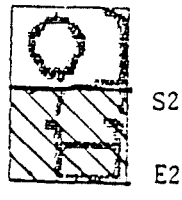
www.dbpia.co.kr

```
    else type=5;
 8  if(type=5)
        if((c=MaxkL[0]+Stem[3])>=MaxC[1])
            starty=MaxC[1];
        else starty=c;
    else
        starty=MaxL[0]+ChWd;
 9  if(종성(starty-y2 주위)존재 ?) type=type+1;
10  goto 14
11  (1,2,5,6형식의 경우)
    중모음 조사
12  횡모음 조사
13  if(횡모음 존재 ?) type=5;Stem[2], Stem[3] : goto 15                    (e),(f)
                  else type=1; goto 15
14  end
15  (종성 조사)                                                            (c),(d)
    Stem[0], Stem[1] 조사
16  if(type=5)
      if((c=MaxL[1]+Stem[3])>=MaxC[1])
        starty=MaxC[1];
      else starty=c;
    else
        starty=y/4*3;
17  if(종성(starty y2주위) 존재 ?) type=type+1;
18  return;
```



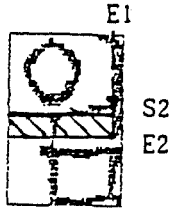
S1: 0
E1: x
S2: y/3
E2: MaxL[0]-1

(a)



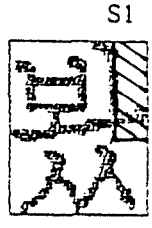
0
x
MaxL[0]+1
y

(b)



0
MaxL[1]-1
MinL[0]+1
MaxC[1]

(c)



MaxL[1]+1
x
MaxC[1]/5
MaxC[1]

(d)

www.dbpia.co.kr

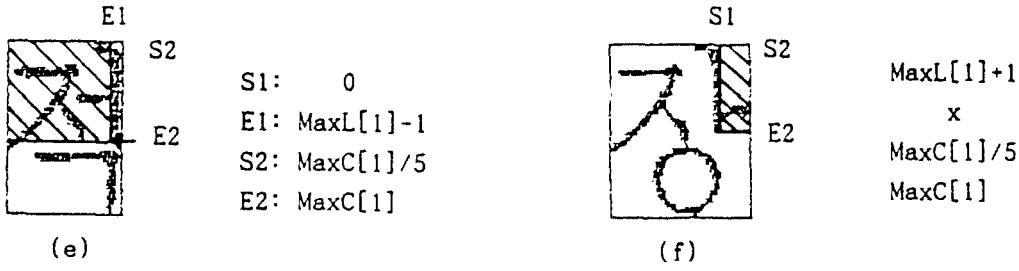


그림 11. 잔 줄기와 짧은 기둥 추출 영역
Fig. 11. Tracing Area of side stem and short pole

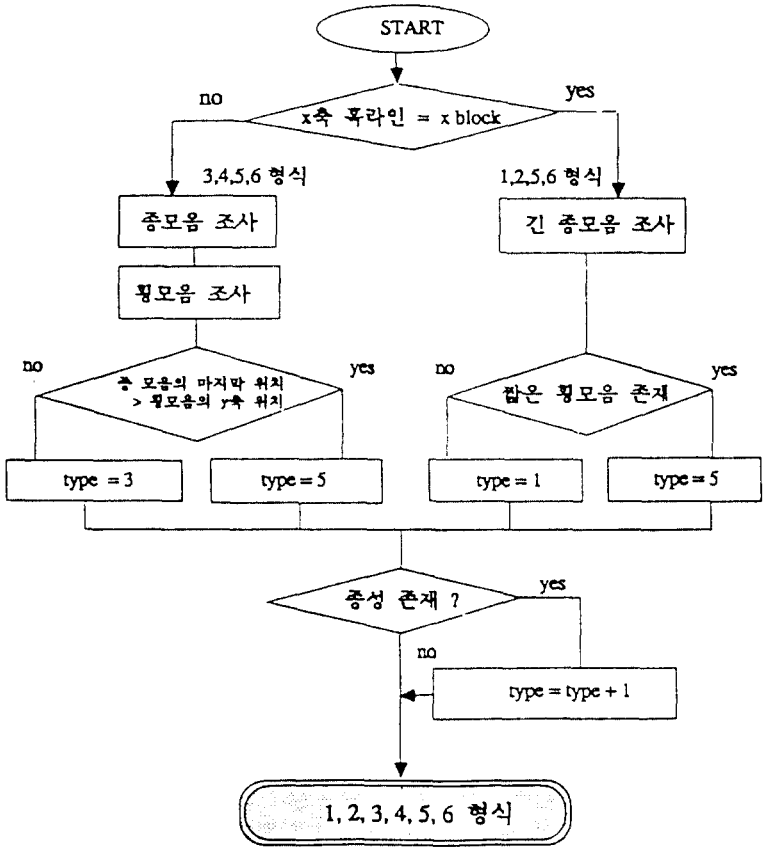


그림 12. 형식분류 알고리즘 흐름도
Fig. 12. Flow Diagram of Classification algorithm of Type

www.dbpia.co.kr

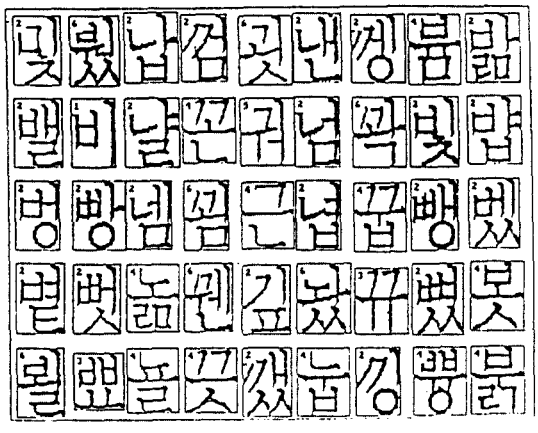


그림 13. 형식분류 결과
Fig. 13. Result of type classification

Ⅳ. 인식 시스템

이 장에서는 6가지 형식으로 분류된 글자를 초성, 횡모음, 종모음 그리고 종성 각각에 대해 템플레이트 매칭 방법에 의하여 인식을 하였으며 템플레이트 매칭시 한글의 유사성 때문에 자소들의 특성에 따라 우선 순위를 주었다.

4.1. 사전 작성

템플레이트 매칭에 필요한 사전은 각 형식별로 초성, 횡모음, 종모음 그리고 종성의 자소들을 가지고 있다. 각 형식에 따른 자소들을 최대 3개씩 구성하였다. 템플레이트 매칭시 한글 형태의 유사성으로 인해 다른 자소인데도 그 유사도가 같은 경우가 있으므로 사전에 표 6과 같은 우선 순위를 주어 복잡한 자소부터 매칭하여 유사도에 따른 오인식을 피할 수 있었다. 또한 위상적인 (topological) 성질로 인하여 1, 2형식에서 초성 "ㅍ"의 이음 줄기 부분이 횡모음의 ㅗ로 인식하는 것을 피할 수 없기 때문에, 본 연구에서는 5, 6 형식에 특별한 자소 "ㅠ"을 추가하는 방법을 제안한다. 형식별 사전 구성 자소를 표 5에, 매칭의 우선 순위를 표 6에 보였다.

표 5. 형식별 사전의 자소 구성
Table 5. JASO composition of dictionary about each type

형 식	1	2	3	4	5	6
자소수	48	71	36	40	65	61

표 6. 매칭 우선 순위
Table 6. Priority of matching

	우 선 순 위
초 성	ㅃ, ㅍ, ㅂ, ㅁ, ㅋ, ㅌ, ㄹ, ㄷ, ㅈ, ㅊ, ㄱ, ㄲ, ㅆ, ㅎ, ㅉ, ㄸ, ㅅ, ㅇ
횡모음	ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, ㅢ, ㅘ, ㅙ
종모음	ㅛ, ㅗ, ㅠ, ㅜ, ㅡ
종 성	ㅀ, ㅄ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅁ, ㅍ, ㅋ, ㅌ, ㄹ, ㄷ, ㅈ, ㅊ, ㄱ, ㄲ, ㅆ, ㅎ, ㅉ, ㄸ, ㅅ, ㅇ, ㅂ, ㅃ, ㅍ, ㄵ, ㄶ

4.2. 템플레이트 매칭

검출하고자 하는 패턴을 나타내는 부분 패턴을 대상 패턴의 시작 부분부터 이동해가면서 대상 패턴 가운데 있는 각 부분 패턴과의 유사도 (상관관계)를 조사하는 것인데 이 방법은 입출력시의 왜곡 현상 때문에 템플레이트와 정확히

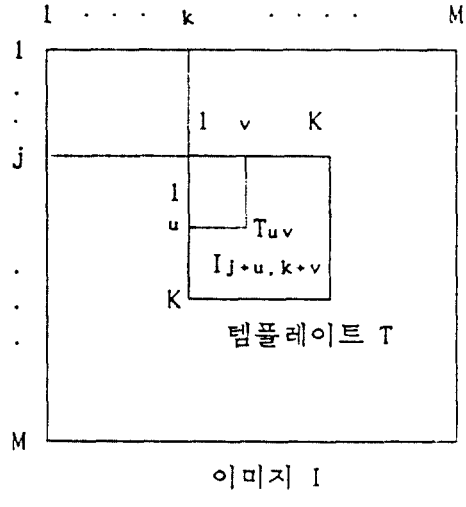


그림 14. 매칭 과정
Fig. 14. Matching procedure

www.dbpia.co.kr

같은 이미지를 기대할 수는 없으므로 이 중 유사도가 가장 큰 값을 검출한다.

이미지 I 내에서 어느 한 부분을 다른 이미지(템플레이트) T와 비교하는 경우에 사용한다. 그림 14는 그 거리를 구하는 방법이다.

위 그림에서 템플레이트 T와 I(j, k)에서 이미지 패턴 I와의 유사도는

$$\text{거리}=\sum_{u=1}^{M_t}\sum_{v=1}^{M_t} I_{u+j,\,k+v}-T_{uv} \qquad (4.1)$$

또는

$$\text{거리}=\sum_{u=1}^{M_t}\sum_{v=1}^{M_t} (I_{u+j,\,k+v}-T_{uv})^2 \qquad (4.2)$$

이다.

여기에서 유사도는 거리와 반비례하는 것으로서 유사도가 가장 크기 위해서는 거리의 값이 적어야 한다. 위치 (j, k)에서 완전히 매칭이 되는 경우에 거리의 값은 0이 된다.

한글의 경우, 거리를 구하는데 있어서 유사성이 매우 심하여 매칭할 때 유사성을 고려해야 한다. 한글은 초성, 횡모음, 종모음, 종성이 조합되어 만들어지는 문자이므로 각 형식에 따라 각각의 위치들이 달라지기 때문에, 같은 자소이더라도 그 모양과 크기가 달라지므로 한글에 있어서 기존의 템플레이트 매칭은 한글의 유사성으로 인해 그 거리의 값이 정확하지 못하였다.

그림 15(a)의 경우 초성을 찾기 위해서 "ㄴ"에 매칭하는 경우 종모음 "ㅔ"의 같은기가 비교 영역에 들어와 그 유사도가 떨어지며, 그림 (b)의 경우 횡모음을 찾기 위해서 "ㅞ"에 매칭하는 경우 종성 "ㅇ"에 의해 거리 값이 커져 유사성이 떨어지게 된다. 그림 15는 매칭시 타 자소에 의해 그 유사도를 떨어뜨리는 예이다.

모음의 일부분이 자음의 영역에 들어가 그 거리를 구할 때 매우 커지는 경우가 많다.

$$\text{거리}=\sum_{u=1}^{M_t}\sum_{v=1}^{M_t} |I_{u+j,\,k+v}-T_{uv}|$$

단, $T_{u+j,\,k+v}=1$일 때 (4.3)

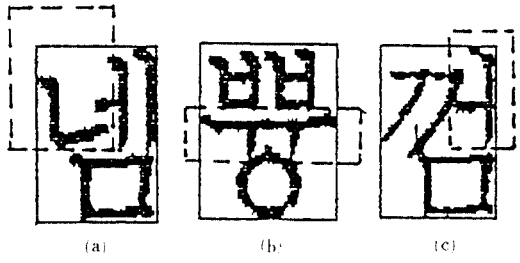


그림 15. 타 자소로 인한 유사도 감소의 예
Fig. 15. Example of similarity reduction owe to another JASO

이러한 두 패턴 간의 차의 증가로 인한 오분류를 없애기 위하여 본 연구에서는 템플레이트의 값이 1일 때만(식 4.3) 원 이미지와의 차를 구함으로서 타 자소가 비교 영역에 침범해 있어도 그 부분에는 영향을 받지 않도록 하였다.

V. 실험 및 고찰

5.1. 실험 시스템

실험에 사용된 문자 패턴을 입력 레이저 빔 프린터(LBP-408)의 출력된 문자 2350자를 대상으로 하였으며 System Quality 사의 IS-300 Image Scanner로 부터 240 DPI로 보낸 데이타를 NEC 9801에서 640×400의 크기로 받아 IBM AT 시스템에서 C 언어(compiler Turbo C)를 사용하여 처리하였다. 그림 16에는 본 연구에서 사용한 실험 시스템의 구성도를 보였다.

5.2. 실험 결과

2350자의 문자를 6형식에 의하여 분류한 결과 98.3%의 형식 분류율을 얻었으며, 형식이 분류된 것을 대상으로 95.2%의 인식율을 얻었다. 표 7에 오분류된 데이타와 오인식된 데이타를 보였다.

5.3. 고 찰

본 연구에서는 세선화를 하지 않고 제이로

www.dbpia.co.kr

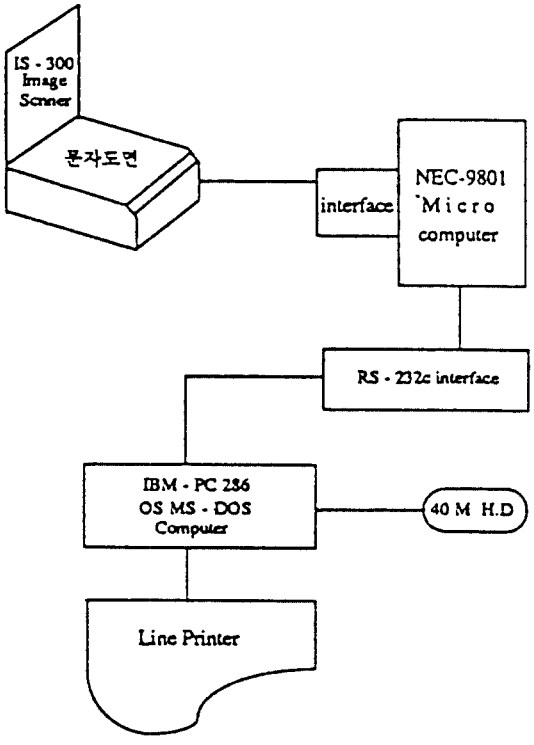


그림 16. 실험 시스템 구성도
Fig. 16. Block diagram of experiment system

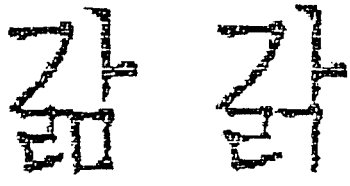
표 7. 오분류된 데이타와 오인식된 데이타
Table 7. mis-classified and mis-recognized data

자	오분류된 데이타		오인식된 데이타	
	분류율	98.3%	인식율	95.2%
1	러, 커		더, 나	
2	갉, 덤, 젊		멱, 발, 점, 닮	
3	요, 뵤			
4			군, 곱, 춘	
5	쇠		와	
6	괆, 뢉		왛, 쒾, 쾠, 잰	

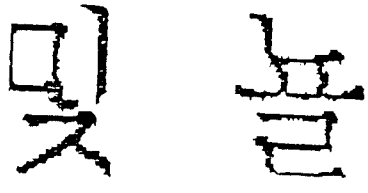
그 자체를 가지고 처리를 함으로써, 세선화시에 필요한 시간을 절약하였으며, type 분류시에 오분류는 입력시에 왜곡된 문자경우, shape 패턴의 사용으로 잡음이 과도하게 들어갔거나 모음 부분이 많이 손상된 경우에 오분류가 일어났다.

또한 초성 자음이 문자의 x축 가장 왼쪽에 붙으며, 그 자음 바로 밑에 종성이 붙은 경우 같은 모로 인식하여 오분류가 발생하였으며 (갊, 겱), 초성 "ㅍ"의 경우 기존의 방법들에서는 횡모음의 짧은 모로 추출되어 오분류가 발생하였으나 본 연구에서는 4.2 절에서 제안한 방법에 의해 좋은 효율을 얻을 수 있다. 이승형[7]은 한글 형식분류율이 86.2%이었으나 본 연구에서는 98.3%를 보였다. 인식시에는 패턴의 일부가 잘라져 입력되거나 문자의 부분이 심하게 왜곡된 경우에 오인식되었으며 95.2%의 인식 결과가 나왔으나 패턴이 잘라져 입력된 데이타를 제외하면 약 4% 이상 향상할 수 있다.

사전을 구성함에 있어서 사전(dictionary)에 우선 순위를 부여함으로서 거리의 값이 같아 (식 4.3)에 의해 초성 "ㄷ"과 일치하는 것은 "ㄴ"과 "ㄷ"이지만 우선 순위(표 6)에 의해 "ㄷ"이 인식된다.



(a) 오분류된 데이타



(b) 오인식된 데이타

그림 17. 오분류와 오인식의 실제 데이타
Fig. 17. Original data of mis classified and mis roginized data

Ⅵ. 결 론

본 연구에서는 문서 인식과 한글 인식을 위하

www.dbpia.co.kr

여 한글의 형식(type) 분류와 한글 인식에 관한 연구를 하였다.

한글이 모음을 중심으로 이루어진 조합 문자라는 한글의 고유 특성을 고려한 구조적 방법을 사용한 본 연구에서는, 기존의 방법들이 사용한 세선화를 하지 않고 마스크에 의해 모음을 검출함으로서 막대한 시간을 줄일 수 있었으며, 형식 분류에 있어서도 좋은 결과(형식 분류율 98.3%)가 나왔다. 그러나, 인식 단계에 있어서는 템플레이트 매칭 방법의 한계성과 인식 대상의 자소가 많아 그리 빠르지 못하였다.

기존의 세선화를 사용한 인식 방법은 세선화의 효율성에 따라 인식율이 좌우하는 세선화에 대한 종속적 비용이 들었으나, 본 연구에서는 그러한 비용을 완전히 줄일 수 있었다.

앞으로의 한글 인식은 효율과 성능면에서, 자소를 추출하고 추출된 자소를 인식하여, 인식의 대상을 줄임으로서 좀 더 빠르고 높은 인식율을 가져올 수 있다.

또한 실시간 인식(real-time recognition)을 위해서 분류된 자소에 대해 템플레이트의 비교 방법이 개선되어야겠으며, 좀 더 다양한 서체와 폰트에서 이러한 연구가 더 확장되어야겠다.

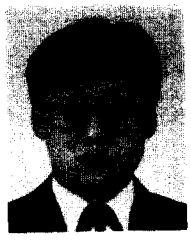
參 考 文 獻

1. J.K. Lee, "Korean Character Display by Variable Combination and its Recognition by Decomposition Methods", Ph. D. dissertation in Keio University, Japan, 1972.
2. 남궁재찬, "Index Window 알고리즘에 의한 한글 Pattern의 분류 및 인식에 관한 연구", 인하대학교 박사학위 논문, 1982.
3. 김태균, "구조 해석에 의한 한글 인식", 한국 정보과학회 인공지능 연구회 영상 처리 및 이해에 관한 워크샵 발표 논문집, pp. 110~116, 1989.
4. 손영우, "속성에 의한 한글 패턴의 온라인 인식에 관한 연구", 광운대학교 석사학위논문, 1982.
5. 조성배, 김진형, "한글 문자 인식을 위한 신경망 기법의 해석에 관한 연구", 제2회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp. 56~62, 1990.
6. 장희동, "필기체 한글 패턴의 실시간 인식에 관한 연구", 광운대학교 석사학위논문, 1986.
7. 이승형, "문서 인식을 위한 한글과 한자의 구별과 한글의 형식분류에 관한 연구", 광운대학교 석사학위논문, 1990.
8. 김진형, "한글의 문자 공학", 마진사, 1989.
9. 우영길, "영문이 혼합된 한글 문서에서의 문자 및 특수 문자 추출에 관한 연구", 서울대학교 대학원 석사학위 논문, 1988.
10. YAMASHITA, HIGUCHI등, "Classification of handprinted Kanji characters by the structured segment matching method", Pattern Recognition Letters Vol. 1, No. 5, 6, 1983.
11. 이주근, 박종욱, "Shape Pattern에 의한 완성형의 한글 인식", 대한 전자공학회지, 제22권 5호, pp. 1-9, 1985.
12. H. Niemann, "Pattern Analysis", Springer Series in Information Sciences, Springer Series, pp. 98~100.
13. 이주근, 남궁재찬, 김영길, "한글 pattern에서 subpattern 분리와 인식에 관한 연구", 대한전자공학회지, Vol. 18, No. 3, 1983.
14. 長谷川純一, 輿水大和, 中山晶, 榎井茂樹 共著, "畫像處理の基本技法", 加勝文明社, 1986.

www.dbpia.co.kr

田 鍾 益(Jong Ik CHEON) 正會員
1963년 4월 1일생
1989년 2월 : 원광대학교 전자계산공학과 졸업(공학사)
1991년 2월 : 광운대학교 대학원 전산기공학과 졸업(공학석사)
1991년 현재 : (주)한국통신 연구개발단 근무중
관심분야 : 패턴인식

南宮 在贊(Jae Chan NAMKUNG) 正會員
1947년 6월 13일생
1970년 : 인하대학교 전기공학과 졸업(공학사)
1976년 8월 : 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 2월 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 : 일본 Tohoku 대학 객원교수
1979년~현 재 : 광운대학교 전자계산기공학과 교수
관심분야 : 패턴인식, 컴퓨터 비전, 인공지능



趙 鎔 周(Yong Joo CHO) 正會員
1964년 3월 15일생
1986년 2월 : 원광대학교 전자계산공학과 졸업(공학사)
1988년 8월 : 광운대학교 전산기공학과 졸업(공학석사)
1991년 현재 : 광운대학교 대학원 전산기공학과 박사과정 재학중.
관심분야 : 패턴인식, 컴퓨터 비전.

www.dbpia.co.kr