

한글 Shape 문자 Pattern에서의 구조적 정보를  
이용한 형식 분류와 인식에 관한 연구

\*  
전종익, 조용주, 이성범, 남궁재찬  
광운 대학교 전자계산기공학과

- A Study On Type Classification And Recognition  
Using Structural Information in Character  
Pattern Of HANGEUL Shape -

CHEON, J I, CHO, Y J, LEE, S B, NAMKUNG, J C  
THE DEPARTMENT OF COMPUTER ENGINEERING  
IN KWANGWOON UNIVERSITY

Abstract

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

First for the purpose of knowing location of character in input image, it processed Making block. Second, after it was 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 according to type classified.

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

제 1 장 서 론

최근의 정보처리(information processing) 환경은 사무 자동화(office automation)와 사용자 인터페이스(user interface)의 급격한 요구로 현대의 정보 사회는 많은 분야에서 다양한 연구를 행하고 있다. 특히 패턴 인식 분야는 문자 인식, 음성 인식등 다양한 매체에서 많은 연구가 행해지고 있으며 그동안의 꾸준한 연구 결과와 컴퓨터 하드웨어 및 소프트웨어 기술의 발전으로 인하여 실용화의 시기에 접어들었고, 부분적으로는 실용화된 상품도 발표되고 있다. 또한 연구의 범위가 단편적인 측면을 벗어나, 인간 생활에 유용한 거의 모든 분야에서 상당량의 연구가 진행되고 있는 실정이다.

이러한 패턴 인식에 관한 연구 분야 중에서, 가장 역사가 깊고 정통성 있게 연구된 분야가 문자 인식 분야이다. 각 나라에서는 그 나라의 문화에 맞는 글자를 가지고 있어 그 나라의 문자를 인식하기 위한 노력이 진행되고 있으며, 우리나라는 한글이란 자모가 조합된 문자 형태로 이루어진 문자를 가지고 있어 1960대 이후 이 분야에서 꾸준한 연구를 해 오고 있다.

본 연구는 우리 고유의 한글 문자 pattern을 인식하기 위하여 종전의 방법과는 다른, shape 자체에서 한글 고유의 6가지 형식을 자동적으로 분류하고, 분류된 형식 문자에 대하여 빠른 인식 방법 구현에 목적을 둔다.

한글 인식에 관한 연구는 1960년대를 시발로 하여 현재까지 꾸준히 연구된 분야 중의 하나이다. 하지만 기존의 연구는 시설 및 연구 인력의 미비로 인하여 연구의 단계가 이론적인 측면에서 연구되어져 왔으나, 현재는 정보 산업 기술의 발달로 인하여 응용적인 측면에서 연구가 되어져 실생활에 응용될 수 있는 연구가 되어져야 한다. 이러한 방향으로 연구

가 진행되기 위해서는 좀더 높은 인식율과 빠른 속도등 여러 요소가 검토되어야 한다. 그러므로 한글을 인식하기 위해서 한글 문자 패턴이 가지는 고유 의 조합 법칙을 고려한 실용적인 응용 연구가 결실 히 요구된다.

한글 패턴 인식에 관한 연구는 크게 결정론적 방법, 구분론적 방법, 구조적 방법 3가지로 분류할 수 있는데 본 연구에서는 구조적 정보를 이용하여 한글의 문자 패턴을 인식하였다.

기존의 한글 패턴 인식에 관한 연구는 대부분 세 선화등의 전처리를 행한 후 인식을 하였는데, 세 선화를 행한 한글 인식은 세 선화의 효율성에 따라 인식율이 결정되었으므로, 본 연구에서는 이러한 문제점을 해결하기 위해 세 선화를 행하지 않고, shape 자체에서 한글의 형식 분류 및 인식을 행하였다.

본 연구는 구조적 방법으로서, 한글의 구조적인 특성을 이용하여 6 형식으로 분류한 후, 각 형식에 따른 각각의 자소에 대해 템플레이트 매칭을 행하여 인식하였다.

제 2 장 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. 본 장에서는 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. 본 연구에서는 현존하는 글자체중 가장 많이 이용되고 있는 명조체 문자체를 대상으로 하였고, 1979년 한글 시각 문제 연구 단체인 글꼴모임(7)에서 정의한 기본줄기 용어를 사용하였다. 그림 1은 글꼴모임에서 정의한 기본줄기와 용어를 보였다.



그림 1. 한글의 기본 줄기와 용어

2.1.2 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 1과 같다.

표 1. 한글의 구성 요소

종 류	구 성 요 소
한글 문자	초성, 중성, 종성
초 성	단자음, 받침음
중 성	단자음, 받침음, 복합자음
종 성	모음
모 음	단모음, 복합모음
단 자음	가, 나, 다, 라, 마, 바, 사, 아, 자, 차, 카, 트, 프, 호
받 침음	가, 마, 바, 프, 썩
복합 자음	프, 나, 마, 라, 마, 라, 마, 라, 마, 바, 마, 가
단 모음	아, 이, 우, 유, 에, 오, 오, 이
복합 모음	애, 윌, 윌, 아, 아, 이, 아, 아, 아
중 모음	아, 이, 이, 이, 이, 이, 이, 이
결 모음	ㅡ, ㅜ, ㅠ, ㅡ

2.1.3 한글의 구조적 특성

위에서 언급되었지만 한글은 모음 중심 문자 즉 종모음과 횡모음의 형태에 따라 글자꼴의 형태가 민감하게 변하기 때문에, 본 연구에서는 표 3과 같이 6가지 특징을 이용하여 인식을 용이하게 하였다.

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)는 1개 이하

나. 횡모음의 결합

- 수평 성분(보) : 1개
- 수직 성분(짧은 기둥) : 2개 이하
- 종모음 존재 시 : 수직 성분(짧은 기둥)은 1개 이하
- "ㅜ" 존재 시 - "ㅣ" 혹은 "ㅣ"의 결합된 형태가 온다.
- "ㅠ" 존재 시 - "ㅣ" 혹은 "ㅣ"의 결합된 형태가 온다.

표 3. 한글의 특징

Table 3. Characteristics of Hangeul

특	징
1	한글에서 각 자소들 중 수평 방향으로 횡 모음이 가장 길다.
2	3, 4, 5, 6 형식 중 횡 모음이 수평 방향으로 가장 길쪽에 온다.
3	한글에서 각 자소들 중 수직 방향으로 종 모음이 가장 길다.
4	1, 2, 5, 6 형식 중 종 모음이 가장 긴 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.

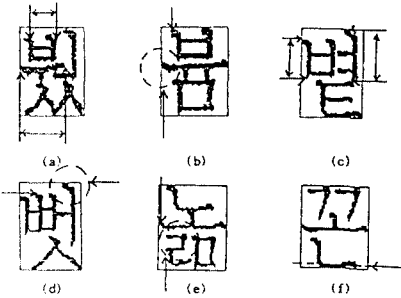


그림 3. 한글의 구조 특성에

2.2 전처리

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

전처리에는 양자화, 평활화, 블러화, 정규화등 여러가지가 있으나 본 연구에서는 인쇄체 문자의 shape를 대상으로 입력시 제한을 두었으므로 정규화는 행하지 않았다.

2.2.1 양자화

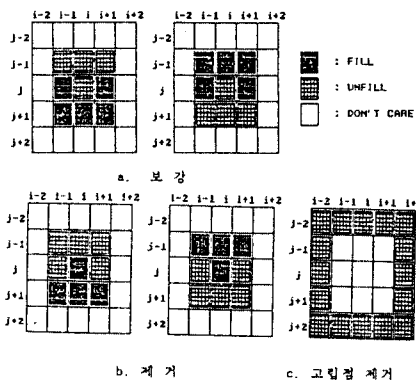
본 연구에서는 문자 입력 장치인 image scanner에서 직접 양자화된 데이터를 사용하였다. 이미지 스캐너에서 입력한 640 X 400의 크기인 양자화된 데이터를 사용하였다.

2.2.2 평활화

2차원 디지털 영상에 있어서 문자 인식에 바람직하지 않은 잡음을 제거하기 위하여 평활화 처리를 한다. 본 연구에서는 고립점을 제거하기 위하여 5 x 5 마스크(그림 5(c))를 사용하였으며 평활화 처리를 위한 마스크는 그림 5(a), 그림 5(b)를 x축과 y축에 대해 실행하였다. 고립점 제거 마스크에서는 unfill 조건을 만족하는 경우 don't care 화소 (fille)들중 1인 화소들을(최대 9개) 제거하고, 평활화 마스크에서는 마스크의 조건이 만족될 때 g(i, j)의 화소가 보장되거나(그림 5(a)) 제거된다(그림 5(b)).

2.2.3 블러화

입력 영상에서 문자 위치를 알기 위한 문자 영역의 블러화를 행하였다.



계 3 장 한글의 형식 분류

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, 제 2장의 모음의 결합 특성에 의하여 줄기와 짧은 기둥을 조사함으로써 인식을 위한 소분류를 하기 위하여 모음 판정 방법과 이에 따른 6 형식 식별 알고리즘을 제안한다.

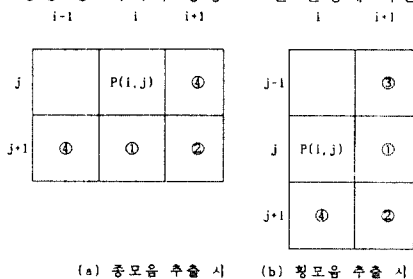
3.1 모음 판정 방법

모음의 존재 여부를 알아내기 위해서 먼저 횡모음의 보와 종모음의 기둥을 찾았으며, 이 때 짧은 기둥과 결합기를 조사하여 모음에 대한 소분류를 행하였다.

먼저, 보와 기둥의 존재 여부를 알아내기 위하여 2 X 3의 마스크를 적용하였고, 곡률을 조사함으로써 사선이 추출되지 않도록 하였으며 또한 사선이 아닌 직선 성분을 가진 자음이 추출되지 않도록 하기 위하여 단점을 조사하였다.

3.1.1 2 x 3 마스크

횡모음과 종모음이 존재하는지를 알기 위한 보와 기둥을 찾기 위하여 우선 순위에 따라 2 x 3 마스크의 8방향 중 네개의 방향으로만 진행해 나간다.



(a) 종모음 추출시 (b) 횡모음 추출시

그림 8. 2 X 3 마스크와 진행 예

3.1.2 곡률 조사

횡모음중 경사진 횡모음이 있는 것은 5, 6 형식에서 대부분 나타나는데 3.1.1의 2 x 3 마스크를 사용하여 진행해가면 주기적으로 곡률을 계산하여 임계값에 따라 빗침과 굴곡 성분이 보나 기둥으로 추출되지 않도록 한다.

예를들어, a(x1, y1)점과 b(x2, y2)점의 곡률을 조사하면 다음과 같다.

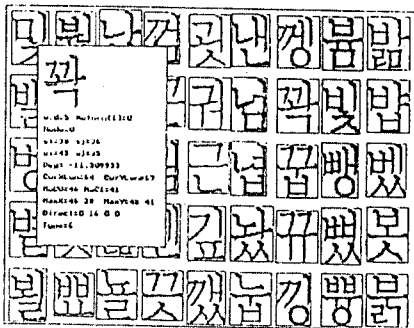
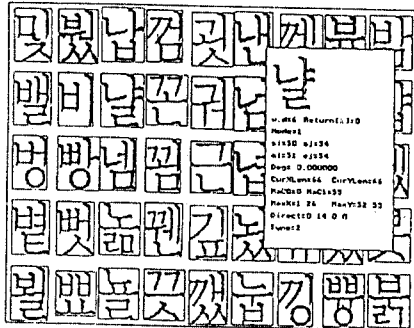
$$L0 = x2-x1, L2 = y2-y1, L1 = (L0^2+L2^2)^{1/2}$$

$$A = 2 \tan^{-1} * ((L-L0) * (L-L1) / (L * (L-L2)))^{1/2}$$

$L0, L1, L2$ : 3변의 길이  
 $L = (L0+L1+L2) / 2$   
 $A$ : 라디안(radian)

3.1.3 단점 조사

횡모음의 시작 부분은 단점이다. 따라서 단점을 조사하지 않았을 때 "ㄹ", "ㄴ", "ㄷ"과 같은 초성이 횡모음으로 추출되는데 이를 방지하기 위해서 횡모음의 보를 추출해 가면서 ChWd x 2 (3.2의 알고리즘 참조) 이상의 화소가 y축을 따라(x/5구간 동안) 존재하면 단점이 존재하는 것으로 하였다. 그림 8에는 위의 특성에 의하여 추출된 종모음과 횡모음의 추출 예를 보였다.



3.2 형식 분류 알고리즘

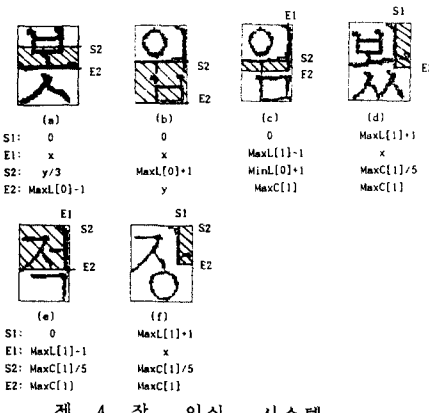
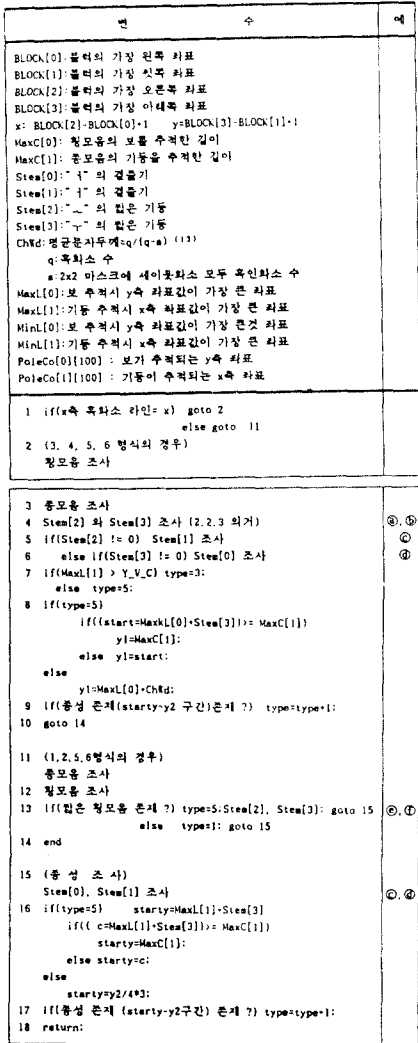
한글은 모음을 중심으로 6 가지의 형식을 가지고 있으므로, 본 연구에서는 모음을 중심으로 한글의 구조적 특성을 이용하여 형식을 분류하였다.

종모음은 위치 변화가 적으나 종성에 따른 변화가 크지만, 횡모음은 위치변화가 많으나 종성에 따른 크기 변화가 없으므로 횡모음의 긴 보 부터 조사했다.

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

긴 기둥이 있는 1, 2, 5, 6 형식 중에서 긴 기둥이 불려의 y축 길이 보다 작으면 2형식 혹은 6 형식이고, 그렇지 않은 경우는 1 혹은 5 형식이며, 대상으로 짧은 보(짧은 횡모음)가 존재하는 지를 조사하여, 존재하면 5 형식 혹은 6 형식 이고, 그렇지 않으면 1 형식 혹은 2 형식이다. 이들 중에서 종성 성분을 조사하여, 2형식과 6 형식을 분류한다. 종성 성분을 조사하는 것은 긴 기둥이 불려의 y축 길이 보다 짧으면 종성이 있는 것으로 판단하며, 그렇지 않은 경우에는 특정 지역의 하변에 연속된 수

평 성분이 있으면 중성이 있는 것으로 판단한다. 이와 같이 문자의 특성을 이용하여 6 가지 형식으로 분류하였다. 그림 10에 한글 6 형식을 분류하는 알고리즘 흐름도를 보인다.



제 4 장 인식 시스템

이 장에서는 6 가지 형식으로 분류된 글자를 초성, 황모음, 종모음 그리고 중성 각각에 대해 템플레이트 매칭 방법에 의하여 인식을 하였으며 템플레

이트 매칭 시 한글의 유사성 때문에 자소들의 특성 에 따라 우선 순위를 주었다.

4.1 사전 작성

템플레이트 매칭에 필요한 사전은 각 형식 별로 초성, 황모음, 종모음 그리고 중성의 자소를 가지고 있다. 각 형식에 따른 자소들을 최대 3개씩 구성하였다. 템플레이트 매칭시 한글 형태의 유사성으로 인해 다른 자소인데도 그 유사도 같은 경우가 있으므로 사전에 표 5와 같은 우선 순위를 주어 복잡한 자소부터 매칭하여 유사도에 따른 오인식을 피할 수 있었다. 또한 위상적인 (topological) 성질로 인하여 1, 2 형식에서 초성 "ㅍ"의 이름 값기 부분이 황모음의 보로 인식하는 것을 피할 수 있기 때문에, 본 연구에서는 5,6 형식에 특별한 자소 " " 을 추가하는 방법을 제안한다. 매칭의 우선 순위를 표 4.

4.2 템플레이트 매칭

검출하고자 하는 패턴을 나타내는 부분 패턴을 대상 패턴의 시작 부분 부터 이동해 가면서 대상 패턴 가운데에 있는 각 부분 패턴과의 유사도 (상관관계)를 조사하는 것인데 이 방법은 잡음과 임력시의 왜곡 현상 때문에 템플레이트와 정확히 같은 이미지를 기대할 수는 없으므로 이 중 유사도가 가장 큰것을 검출한다.

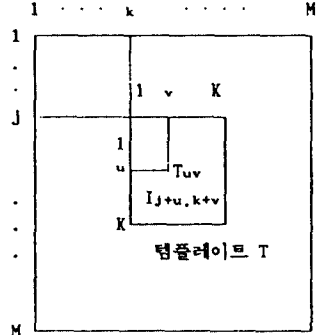
한글의 경우, 거리를 구하는데 있어서 유사성이 매우 심하여 매칭할 때 유사성을 고려해야한다. 한글은 초성, 황모음, 종모음, 중성이 조합되어 만들어진 문자이므로 각 형식에 따라 각각의 위치들이 달라지기 때문에, 같은 자소이더라도 그 모양과 크기가 달라지므로 한글에 있어서 기존의 템플레이트 매칭은 한글의 유사성으로 인해 그 거리의 값이 정확하지 못하였다.

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

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

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

매칭 우선 순위	
	우 선 순 위
초 성	ㅍ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅞ, ㅝ, ㅚ, ㅝ, ㅜ, ㅞ, ㅝ, ㅚ, ㅝ, ㅞ, ㅝ
황 모 음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ
종 모 음	ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ
중 성	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅞ, ㅝ, ㅚ, ㅝ, ㅞ, ㅝ, ㅚ, ㅝ, ㅞ, ㅝ



이미지 I

제 5 장 실험 및 고찰

5.1 실험 시스템

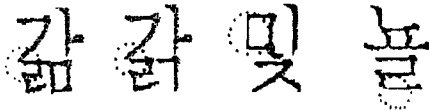
실험에 사용된 문자 패턴은 삼보 LBP 4081 폰트 2350자를 대상으로 하였으며 system quality 사의 IS-300 Image Scanner로 부터 240 DPI로 NEC-9801에서 640 X 400 의 크기로 받아 IBM-AT 시스템에서 C 언어를(compiler - Turbo C) 사용하여 행하였다.

5.2 실험 결과

2350자의 문자를 6 형식에 의하여 분류한 결과 98.3%의 형식 분류율을 얻었으며, 형식이 분류된 것을 대상으로 95.2%의 인식율을 얻었다.

5.3 고찰

본 연구에서는 세선화를 하지 않고 세이프 그 자체를 가지고 처리를 함으로써, 세선화시에 필요한 시간을 절약하였으며, type 분류시에 오분류는 입력시에 왜곡된 문자경우, shape 패턴의 사용으로 잡음이 과도하게 들어 갔거나 모음 부분이 많이 손상된 경우에 오 분류가 일어났다. 또한 초성 자음이 불력의 x축 가장 왼쪽에 불으며, 그 자음 바로 옆에 중성이 붙은 경우 긴 보로 인식 하여 오분류가 발생하였으며(값, 값), 초성 "ㅇ"의 경우 기존의 방법들에서는 횡모음의 짧은 보로 추출되어 오분류가 발생하였으나 본 연구에서는 4.2 절에서 제안한 방법에 의해 좋은 효율을 얻을 수 있었다. 본 연구에서는 98.3%를 보였다. 인식 시에는 패턴의 일부가 잘라져 입력된 경우 또는 문자의 부분이 심하게 왜곡된 경우에 오인식되었으며 95.2%의 인식 결과가 나왔으나 패턴이 잘라져 입력된 데이터를 제외하면 약 4% 이상 향상할 수 있다.



오분류와 오인식의 실제 데이터

제 6 장 결 론

본 연구에서는 문서 인식과 한글 인식을 위하여 한글의 형식(type) 분류와 한글 인식에 관한 연구를 하였다.

구조적 정보를 이용한 형식 분류는 기존의 다른 방법들 보다 매우 우수하였으며 인식에 있어서도 매우 정확한 인식을 하였으나, 시간적 효율면에서 템플릿 매칭 방법의 한계성과 인식 대상의 자소가 많아 그리 빠르지 못하였다.

앞으로의 한글인식은 효율과 성능 면에서, 자소를 추출하고 추출된 자소를 인식하여, 인식의 대상을 줄임으로써 좀 더 빠르고 높은 인식율을 가져올 수 있다.

또한 실시간 인식(real time recognition)을 위하여서 분류된 자소에 대해 템플릿의 비교 시간이 빨라야겠으며, 좀 더 다양한 서체와(typeface) 멀티 폰트(multi-font)에서 이러한 연구가 더 확장되고 좀 더 정확한 형식 분류 알고리즘이 개선되어야 할 것이다.

참 고 문 헌

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 부분 분리와 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 석사 학위 논문, 1982
5. 조성배, 김진형, "한글 문자 인식을 위한 신경망 기법의 개선에 관한연구", 제 2회 영상 처리 및 이해에 관한 워크샵 발표 논문집, pp. 56-62, 1990
6. 장희돈, "필기체 한글 패턴의 실시간 인식에 관한 연구", 광운대학교 석사 학위 논문, 1986
7. 김진평, "한글의 글자 표현", 미진사, 1989
8. 오인권, "영문이 혼합된 한글 문서에서의 문자 및 특수 문자 추출에 관한 연구", 광운대학교 대학원 석사 학위 논문, 1988.
9. YAMASHITA, HIGUCHI 등, "Classification of handprinted Kanji characters by the structured segment matching method", Pattern Recognition Letters Vol 1, No. 5,6, 1983
10. H.Niemann, "Pattern Analysis", Springer Series in Information Sciences, Springer Series, pp. 98-100.
11. 長谷川純一, 興水大和, 中山晶, 横井茂樹 共著, "画像處理의 基本 技法", 加藤 文明社, 1986.
12. 권중익, "한글 shape 문자 패턴에서의 구조적 정보를 이용한 형식 분류와 인식에 관한 연구", 광운대학교 석사 학위 논문, 1991