論 文

인쇄체 한자에서 Radical의 구조적 정보를 이용한 형식분류 및 부분패턴 추출에 관한 연구

正會員 金 政 漢* 正會員 趙 鎔 周* 正會員 南宮 在 贊*

A Study on Type Classification and Subpattern Extraction Using Structural Information of Radical in Printed Hanja

Jung Han KIM,* Yong Joo CHO,* Jae Chan NAMKUNG* *Regular Members*

要 約 본 논문에서는 한자인식을 위한 전단계로서 인쇄체 한자를 대상으로 한자의 특성과 구조적 정보를 이용한 새로운 분류 알고리즘을 제안하였다. 한자는 자종이 많고 구조가 복잡하여 인식은 물론이고 분류하는 데에도 많은 어려움이 따른다. 이러한 문제점을 해결하기 위해, 본 연구에서는 한자패턴을 형식분류한 후 분류된 패턴에서 공통으로 존재하는 부분패턴을 추출하는 실험을 행하였다. 먼저 입력한 문자 패턴에 대해 전처리를 행한 후, 방향 세그먼트를 추출하여 4방향화면상에서 레이블링을 하고, 문자패턴의 부분패턴 존재 영역에 기초한 구조적 정보를 이용하여 12가지 형식으로 분류한 다음 그 부분패턴을 추출하였다.

중·고교 교육용한자 1800자에 대해서 실험을 행한 결과 93.07%의 형식분류율을 얻었고, KS C5601 표준 삼보 LBP 한자 폰트 4,888자에 대해서는 90.12%의 형식분류율을 얻었으며, 분류된 데이타로 부터 부분패턴을 추출하여 인식에의 적용가능성을 보였다.

ABSTRACT This paper proposes a new classification algorithm using characteristic and structural information of printed Hanja as preliminary stages of Hanja-character recognition.

Hanja is difficult for not only recogniton but classification as many character and complicated structure. In this paper, to solve this problem, extracted common subpattern in classified pattern after processing type classification for Hanja pattern.

First, we extracted subpattern, after we process preprocessing about input of character pattern, extracting directional segment, labeling on 4-directional pattern and 12 type classified using structural information based on the subpattern existing region of character pattern.

Through the experiment, this study obtained that classified rate of Hanja is 93.07% on 1800 character of educational Hanja and 90.12% on 4888 character of KS C5601 standard TRIGEM LBP Hanja font and saw that as extracting subpattern at classified data was this paper possibly applied to the recognition.

I. 서 론

최근 들어서 사회가 급격히 변화함에 따라 정보처리 시스템의 개발에 많은 관심이 고조되고 있다. 그중에서도 패턴인식 분야의 한 부류인 문자인식 분야의 연구는, 컴퓨터에 정보를 입력할 때, 타이프라이터 (typewriter)나 워드프로세서 (word processor)의 사용만으로는 처리속도에 한계성이 있기 때문에 그 필요성이 증대되어 왔다.

우리나라 문자인 한글에 관한 연구[1][2][3]는 현재 국내에서 활발히 진행중에 있으나, 한자문화권인 우리나라에서의 한자 인식에 관한 연구는 초보적

* 光云大學校 電子計算機工學科
Dept. of Computer Engineering Kwang Woon University
論文番號 : 91-21(接受1990. 11. 23)

인 단계에 머물러 있는 실정이다. 그러나 문서인식의 하나라고 할 수 있는 신문 문서인식의 측면에서 볼 때, 대부분의 국내 신문이 한글 위주로 이루어져 있긴 하지만, 문자의 의미표현을 위한 특성상, 한자의 존재는 결코 간과할 수는 없다. 또한 현재 한참 연구중에 있는 D.T.P.(desk top publishing)와 같은 연구에서 무시할 수 없는 것이 한자인식에 관한 연구라고 할 수 있다.

일본에서는 한자가 가나문자와 함께 사용어로 쓰이고 있기 때문에 이 분야에 대한 많은 연구가 있었는데, 특히 문자의 영역을 수치로 표현하기 위하여 여러가지 벡터(vector)의 형태로 표현하여 계산한 뒤, 거리의 계산을 하여 인식을 하는 방법인 전정통적인 방법[14]이 많이 연구되어 약 95% 정도의 인식률을 얻고 있다. 이러한 인식방법은 강력하기는 하나 너무나 계산 중심적인 경향을 갖는다는 단점이 있다. 그래서 현재에는 이에 대한 해결 방법으로 한자에 대한 특별한 지식을 training system이라고 하는 KBS(knowledge based system)에 집어넣어 효율을 높이고 속도를 개선시키는 새로운 방법이 도입되고 있다[6]. 한자는 글자의 수가 2만자가 넘는 방대한 문자로 구성되어 있는데, 인식을 하기 위해서는 먼저 한자의 구조적인 형태를 살핀 다음, 그 형태가 지니는 특징들을 살펴야 할 것이다. 우리는 일상생활, 예를 들면 한자사전을 참조하여 문자를 찾을 때에 부수색인에 의해 문자를 구하곤 한다. 한자는 부수(radical)을 포함하는 것이 대부분이므로 이러한 특성에 착안하여 분류를 한 다음 인식을 행함이 유효하다.

한자인식에는 크게 나누어 인쇄체 한자인식[10][12]과 필기체 한자인식[5][6]이 있는데, 인식시의 문제점은 자종이 광대하다는 양적인 문제와 대상자형의 구조가 복잡하고 유사문자가 많다고 하는 질적인 문제점을 들 수가 있다. 이러한 문제점을 해결하고, 인식부로의 대응을 피하기 위한 전단계로 한자패턴에서 서로 공통으로 가지고 있는 부분패턴(부수, radical)을 이용하는 방법이 있다. 기존에 발표된 대분류방법은 문자 패턴의 배경부의 특징 및 주변분포에 기초[7][12][13]하거나 세그먼트의 구조적 매칭에 의한 구조해석적인 방법[8][9]에 기초하여 분류를 한다.

본 논문에서는 이러한 한자 구성요소의 구조적 특성과 주변분포에 따른 문자 히스토그램의 특징성[12]을 함께 도입하고, 한자의 특성을 고려하여 부분패턴을 추출하였다. 먼저 한자패턴은 방향을 갖는 선소(방향 세그먼트 : direction segment)의 집합으로 구성되어 있으므로 각 방향별 세그먼트를 구하고[10], 각 문자에 해당하는 세그먼트의 영역을 조사하여[11], 한자패턴을 12종류의 form으로 형식분류한다. 또한 분류율을 높이기 위해 각 문자 세그먼트의 위치관계나 길이, 방향 등의 지식을 이용하여 분류된 형식으로부터 부분패턴을 추출하였다. 이러한 방법으로부터 문자패턴상에서 추출된 부분패턴의 존재 영역과 그 이외의 영역과의 지식 베이스적인 측면에서의 단계적 인상이 향후 과제라고 할 수 있다.

Ⅱ. 한자 패턴의 구조분석 및 인식방법

대상 문자인 한자패턴을 인식하기 위한 인식시스템을 설계하기 위해서는 한자구조와 구성에 관한 특징이 선행되어야만 한다. 따라서 본 장에서는 한자의 구조 및 특성에 대해 기술하였으며, 본 논문에 관계된 인식방법에 대해서 살펴본다.

1. 한자의 구조 및 특성

한자의 구조를 살펴보면, 吉와 같은 상하구조(North South structure)化와 같은 좌우구조(East West structure), 困과 같은 내외구조(compound structure) 등으로 구성되어 있는 경우가 아주 많다. 즉, 한자는 상형문자에서 시작되었다고 하지만 문명의 발달과 함께 문자에 의해 표현되는 내용이 복잡해지면서 문자의 형도 이미 사용된 문자의 개개의 형이 조합되기도 하고 원래 하나의 형이던 것이 나누이지는 등 복잡한 형으로 발전하였다. 이렇듯 한자는 도형적 계층성을 갖는다는 것을 알 수가 있다. 한자

는 자종이 광대하고 구조가 복잡하며, 유사문자가 많이 존재하기 때문에 인식은 물론이고 분류하는데에도 많은 어려움이 따른다. 본 논문에서는 이러한 문제점을 극복하기 위해서, 한자의 구성 및 구조상의 특징들을 살펴본 결과 한자패턴이 부분패턴(부수, radical)을 서로 공통으로 가진 점에 착안하여, 이들이 문자상에서 위치하는 영역을 중심으로 분류를 하는 방법을 사용하였다.

본 절에서는 한글이 그 구조적 특징에 의해 크게 6가지 형식으로 구분하여 인식부에 대응을 시켰을 때 그 유효성을 증명하기가 더 쉬웠다는 데 착안하여, 한자의 도형적 계층성에 따라 그림 1과 같이 12가지 형식으로 분할을 시도했다. 분할방법은 동 그림에서 처럼 도형적으로 2분할하며 공통 부분패턴을 소유하는 한자의 집합을 동일 그룹으로 분류한다. 단, 『利』, 『推』등과 같이 2분할된 부분의 어느 것도 공통 부분패턴인 경우에는 동일 그룹내의 한자의 갯수가 많은 쪽으로 분류한다.

본 논문에서는 대상 한자 패턴이 인쇄체라는 점에 착안하여 대상문자의 구조적 위치를 실험에 의해 정량적으로 분석하여 각각의 대상문자의 형식을 정하고, 다시 여기에서 지식의 개념을

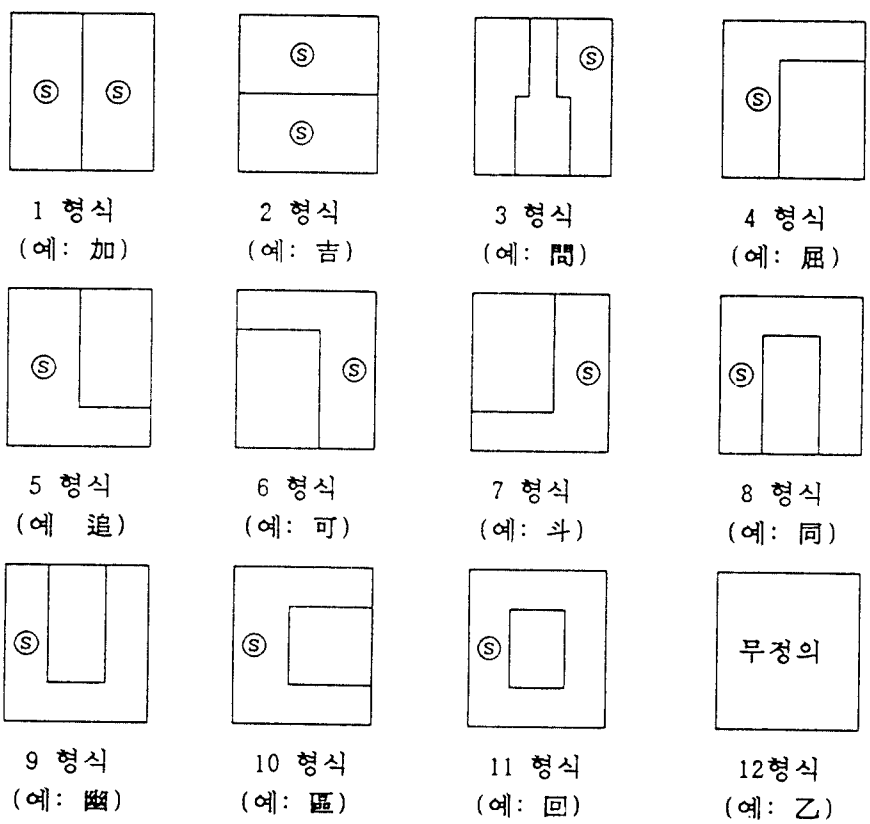


그림 1. 한자의 12가지 형식
Fig. 1. 12 type of Hanja

도입하여 분류된 형식으로 부터 부분패턴을 추출하는 식의 top-down적인 방법을 사용하여 한자 구조의 복잡성을 해결하였다.

2. 한자의 인식방법

본 논문에서는 문자의 특징을 이용하는 방법중의 하나인 주변분포를 이용하는 방법과 세그먼트 매칭(스트로크 매칭)방법을 혼용하였다. 주변분포는 문자를 X축 및 Y축등의 직선상에 투영한 것으로 인쇄체 한자인식등에 유효한 특징량이다[15]. 세그먼트 매칭방법은 인식시 비교적 안정된 세그먼트의 특징정보를 추출하여 방향별, 길이별로 사전패턴과 매칭하여 인식하는 방법이다[16]. 주변분포에 의한 방법은 양자화가 잘못될 경우에는 엉뚱한 분포특징을 가지게 되는 단점이 있으며 구조적 매칭 방법은 한자 패턴의 국소적인 유사성에 기초하고 있는 바, 한자패턴이 가지는 여러가지 정보를 활용치 못하는 측면이 많다. 그림 2는 각 기준축에서의 주변분포를 나타낸 것이다.

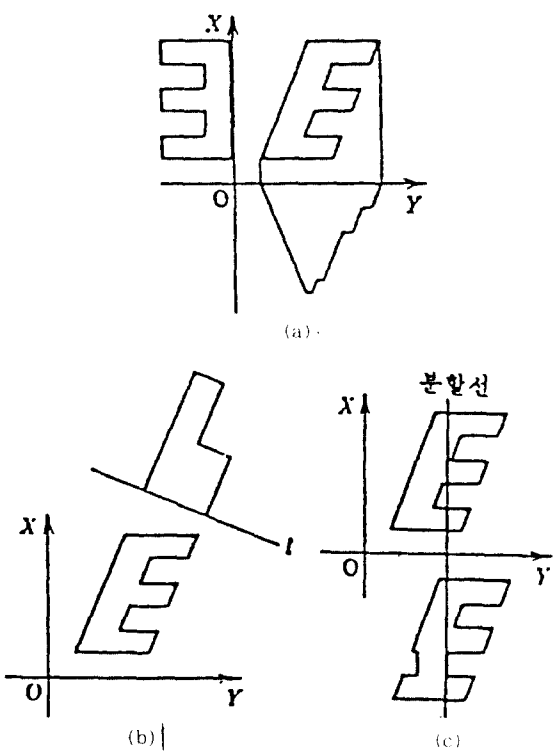


그림 2. 주변분포
Fig. 2. Peripheral Distribution

Ⅲ. 전처리와 세그먼트추출 및 레이블링

1. 처리와 흐름도

그림 3은 본 연구의 처리도를 나타낸 것이다.

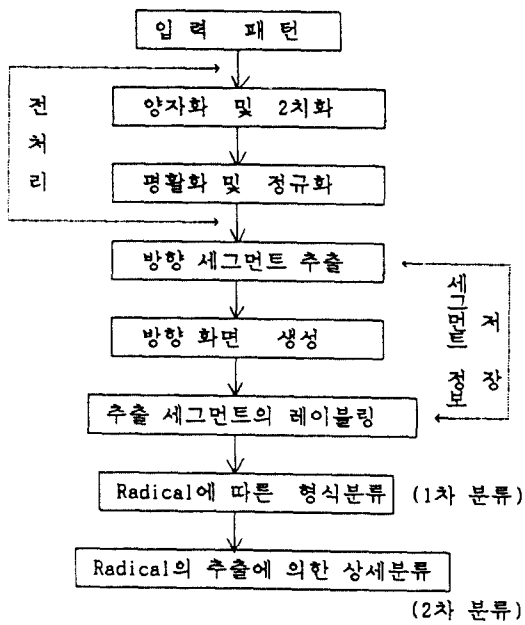


그림 3. 처리의 흐름도
Fig. 3. Flowchart of process

2. 영상의 전처리

영상 입력장치를 통해 입력되는 문자 패턴 정보는 흑화소는 1로, 백화소는 0으로 양자화되어 잡음과 함께 입력된다. 본 논문에서는 다음과 같은 방법으로 전처리하였다.

2.1. 평활화(잡음제거)

입력된 문서는 2차원 디지탈화에 따른 문서상의 잡음과 하드웨어상의 잡음을 제거하기 위해 평활화(Smoothing) 처리를 하여 매끄러운 패턴으로 만든다. 3×3 마스크(mask)를 사용하여

고립점을 제거하였으며 1×3 마스크를 사용하여 미세한 잡음을 제거하였다. 그림 4에 평활화 처리의 예를 나타내었다. (a)는 3×3 마스크와 1×3 마스크를 나타낸 것이고 (b)의 4×2 마스크와 3×2 마스크를 사용한 filling을 통해 X축을 평활화 했으며 (a), (b)경우 둘다 다시 마스크를 90° 회전시켜서 Y축도 평활화 하였다[15].

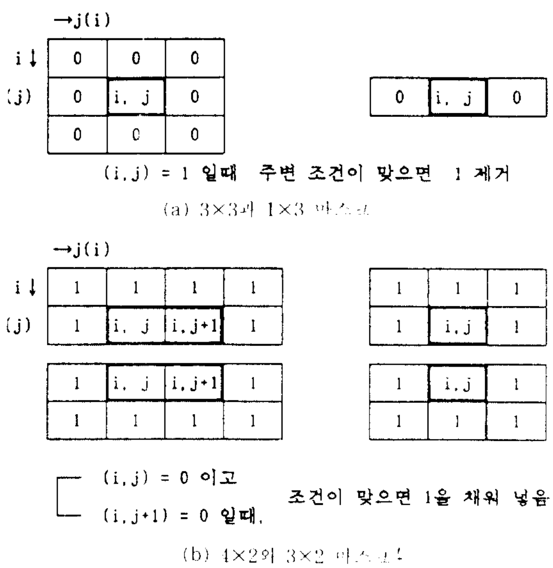


그림 4. 평활화 처리를 위한 마스크
Fig. 4. A mask for smoothing processing.

2.2. 정규화(Normalization)

문자의 정규화는 위치, 크기, 경사 및 문자 선폭등을 정형화하는 처리로서 인쇄체 매칭법에 있어서 특히 중요하다. 본 논문에서는 인쇄체 문자를 대상으로 하여 단지 위치의 정규화만을 행하였다. 그림 5(a)와 같이 문자들의 중심에 위치하지 않는 문자의 중심위치를 (b)와 같이 정규화 한다[15].

중심이 G인 문자 영역을 둘러싼 70×70의 장방형에서 문자영역의 가로길이를 x1, 세로길이를 y1이라 하면, 구하고자 하는 장방형의 중심 C(x, y)는 식 (1)에 의해 구할 수가 있다.

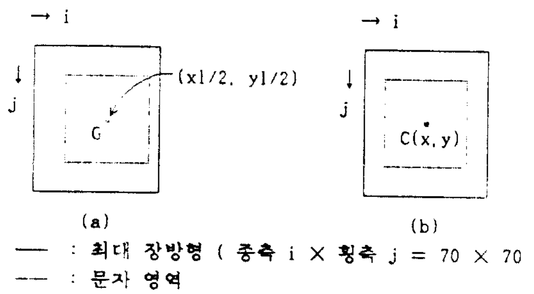


그림 5. 문자 위치의 정규화
Fig. 5. Normalization of character position

$$\begin{bmatrix} x=\{(70-x1)/2\}+(x1/2) \\ y=\{(70-y1)/2\}+(y1/2) \end{bmatrix} \quad (1)$$

3. 문자의 방향 세그먼트 추출과 레이블링

3.1. 코드화(방향 세그먼트 추출)

각 스트로크마다 방향성을 부여하기 위하여 문자패턴을 코드화한다. 본 논문에서는 정방격자상에서 간격이 √2 이내이고 등간격에 나란한 방향으로 $\theta_k=0°, 45°, 90°, 135°$ 즉, 수평, 대각, 수직, 역대각의 4개 방향만을 고려하였다[10]. 센서가 연결성을 잃어버릴 때 기준점(x_0, y_0)로 부터 특정각도 θ_k방향의 기하학적 거리 $L^{+}(x_0, y_0)$, θ_k 방향의 거리 $d_k(x_0, y_0)$는 식(2), (3)에 의해 얻어진다.

$$L^{+}(x_0, y_0)=\sqrt{(x-x_0)^2+(y-y_0)^2} \quad (2)$$

$$d_k(x_0, y_0)=L^{+}(x_0, y_0)+L^{-}(x_0, y_0) \quad (3)$$

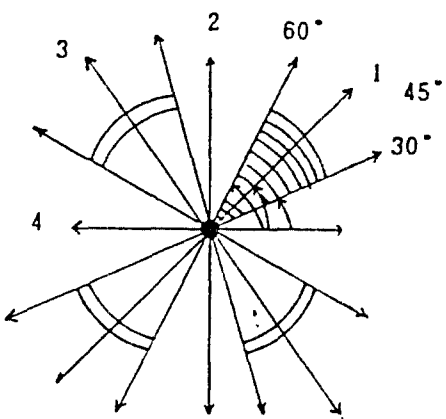
방향코드는 각 흑점에 대한 θ_k 방향의 기하학적 거리 $d_k(x, y)$에 의거하여, 흑점 $f_B(x, y)$의 값은 식(4)에 의해 구한다.

$$f_B(x, y)=k\,(k=1, \cdots, k) \quad (4)$$

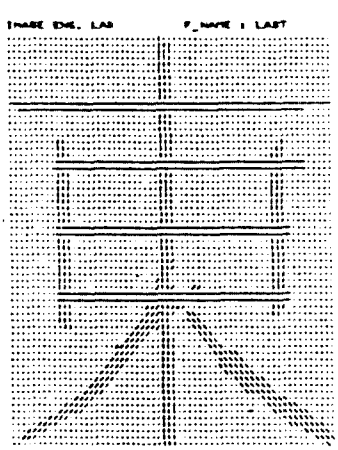
단, $d_k(x, y)=\max\{d_m(x, y)\}$ $(1\leq m\leq K)$

사선성분의 세그먼트는 실제로 각도가 너무

다양하여 정확히 45° 가 아닌 경우로 손쉽게 추출하기가 쉽지 않다. 본 논문에서는 사선 성분의 세그먼트의 각도에 적당한 임계치(Threshold)를 주었는데, 즉 역대각선 세그먼트의 각도를 θ_1°라 할 때 $\pm 30° \leqq \theta_1° \leqq \pm 60°$ 이면 $\theta_1° = \pm 45°$로 인정하여 그 방향코드를 k=1로 코드화 하였다. 그리고 대각선 세그먼트의 각도 θ_3°는 $\pm 120° \leqq \theta_3° \leqq \pm 150°$를 $\pm 135°$로 인정하여 그 방향코드를 k=3으로 코드화 하였다. 그림 6(a)는 각 방향세그먼트의 각도, 방향코드, (b)는 방향코드화 패턴을 보여준다.



(a)



(b)

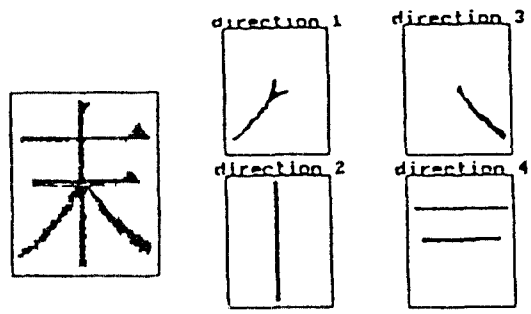
그림 6. 방향 코딩의 예
Fig. 6. Example of direction coding

3.2. 방향화면의 생성

식(4)에 의해 흑점의 값은 1, …K의 K값으로 변환된다. 이 값은 각 흑점에 대한 연결 흑영역 내부에서 가장 긴 거리의 방향을 나타낸다. 다음에 방향코드에 따라 흑영역을 분할하여, 식(5)에 의해 K개의 2치 패턴을 생성한다.

$$F_k = \left[(x, y, z) \middle| \begin{matrix} f_B(x, y) = k \text{일 때 } z = 1 \\ f_B(x, y) \neq k \text{일 때 } z = 0 \end{matrix} \right] \quad (5)$$

F_k를 제 k방향 화면이라 부르고, θ_k방향에서 거리가 가장 긴 점집합이다. 그림 7에 그림 6의 방향 코드화 패턴으로 부터 생성된 방향화면의 실례를 보여준다.

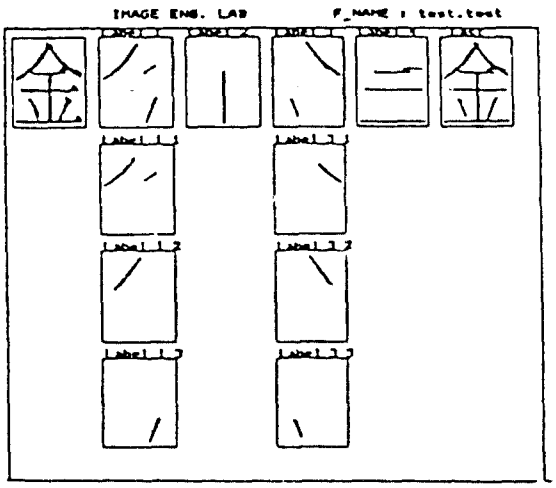


(a) 입력문자 (b) 4개의 방향화면

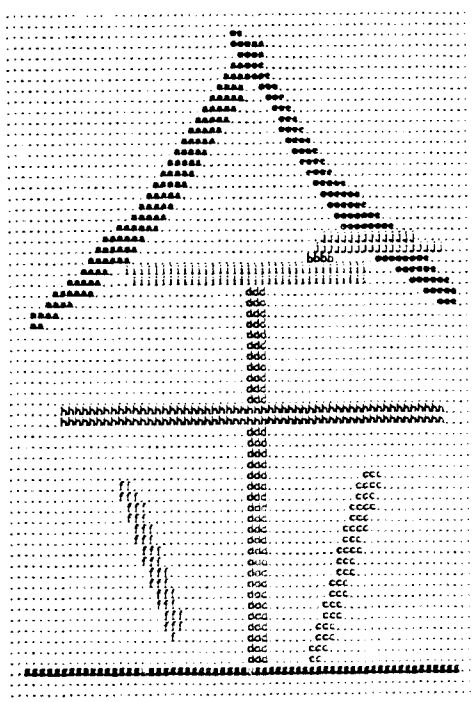
그림 7. 방향화면의 작성
Fig. 7. Direction pattern generation

3.3. 추출된 방향세그먼트의 레이블링 (Labeling)

4개의 방향 화면으로 부터 추출된 세그먼트에 대해서 레이블 값 m(m=1,…, M : M은 총 세그먼트수)을 주어 그 순서를 부여하는 레이블링을 행한다. 레이블링은 k=1인 방향화면 부터 방향코드가 작은 순으로 행하는데, 동일방향의 세그먼트에 대해서는 축방향으로 주사를 하면서 레이블링을 시도한다. 즉 역대각(↙), 수직(↕), 대각(↘) 방향의 세그먼트는 화면의 좌상에서 부터 우하로 주사를 하면서 세그먼트 출현순으로



(a) 그래픽 화면상의 레이블링



(b) 텍스트 화면상의 레이블링

그림 8. 문자 세그먼트의 레이블링
Fig. 8. Labeling of character segment

레이블링을 하고, 수평(↔)방향의 세그먼트를 좌하로부터 우상으로 주사를 하면서 레이블링을 행한다. 그리고 향후 인식부에 이용하기 위해서 레이블링된 정보, 즉 세그먼트의 양 끝점좌표, 길이, 방향성분등을 저장시키 놓는다. 그림 8의 (a)는 그래픽 화면상에서의 한 문자에 대한 레이블링 과정을 나타낸 것이고, (b)는 레이블링된 결과를 텍스트 화면상에 나타낸 것이다[10].

3.4. 문자의 블럭화

각각의 문자들이 모두 하나의 정방형으로 구성 가능하므로 문자를 한 문자씩 처리하여 그 위치 등을 파악하기 위해서 블럭화를 하게 되는데, 각 문자열 마다 시작점과 끝점을 찾아서 문자의 블럭을 구성한 결과를 그림 9에 보였다.

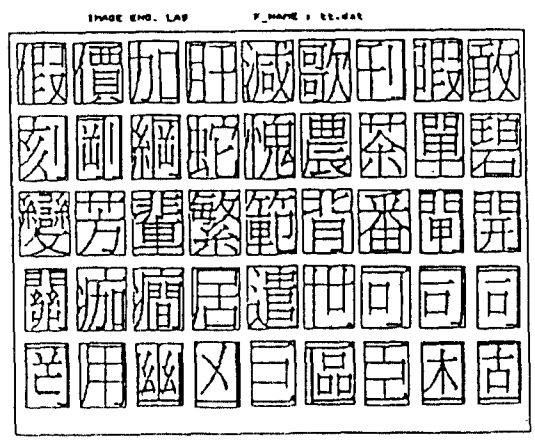


그림 9. 한자 데이타의 블럭화
Fig. 9. Block of Hanja data

Ⅳ. 한자의 형식분류 및 Radical의 추출

1. 한자의 형식분류

복잡한 구조를 가지는 한자를 인식하기 위해서 몇 번의 인식의 전단계인 대분류가 필요하다. 그래서 본 논문에서는 문자가 가지고 있는 고유한 구조를 분석하여 12가지(11가지 분할 가능한 형식과 1가지의 분할 불가 형식)으로 형식 분류를 하였다. 12가지로 분류를 행한 이유는 그 이상으로 형식 분류를 할 경우 후보 카테고리의

갯수는 줄어들지만 형식이 겹치는 등 분류가 잘 되지 않았으며, 12가지 미만으로 할 경우에는 분류는 잘 되나 후보 카테고리수가 너무 많아지는 문제가 발생하였기 때문이다.

일본의 경우는 한자 패턴을 9가지 class로 나누어 각 class 마다 부분 패턴이 존재할 것으로 생각되는 범위를 임의로 미리 정하여 그 영역내에서 특징을 찾아낸 반면, 본 논문에서는 인쇄체라는 특성을 살려 한자 패턴에서 부분패턴 존재가능 영역을 구조적 측면과 국소적 유사성의 측면에서 정량적으로 분석한 후 분류를 행하였다. 한자의 구조를 살펴보면 분할불가한 단순구조와 분할가능한 구조(좌우구조, 상하구조, 내외구조, 받침구조등)로 생각할 수 있지만 이외에도 복잡한 구조를 이루는 문자도 있기 때문에 문자마다 변과 부수들이 위치하는 장소에 따라 분류함이 유효하다. 변(偏), 방(旁), 관(冠), 각(脚)에 의해 분류하는 1, 2 형식은 한자의 형식분류시 가장 핵심이 되는 형식으로 교육용 한자 1800자를 대상으로 볼 때 1형식이 930여자, 2형식이 460여자로 거의 70%내지 80%를 차지한다.

1.1. 형식분류의 절차

구조적인 특성을 고려, 특징을 조사하여 먼저 3~11 형식을 분류한 후 여기서 분류가 되지 않는 문자에 대해서 1 또는 2형식으로의 분류를 행하였다. 그림 10은 그 처리도를 나타낸 것이다.

1.2. 특징조사 및 세그먼트 영역 추적

형식분류 알고리즘을 설명하기 위해서는, 특징영역에 특정 세그먼트가 존재하는지의 여부를 판단하기 위한 영역설정(그림 11)과 특징추출을 위한 조사(그림 12, 13) 과정이 필요하다.

(1) 영역설정

정방향의 문자 영역내에 그림 11과 같은 영역을 설정한다.

(2) 특징조사

형식분류를 위한 특징추출의 처리도와 그 예를 그림 12와 그림 13에 보였다. 그림 13에서 빗금

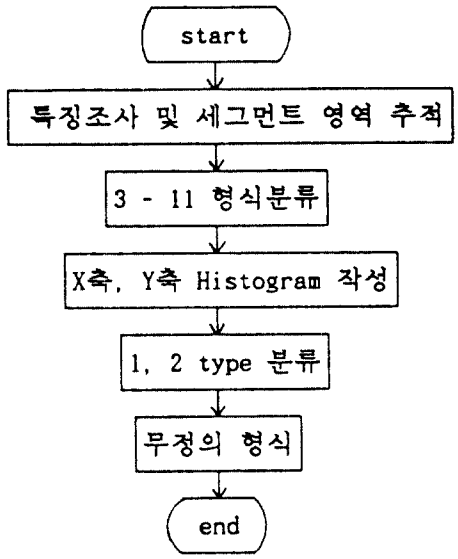
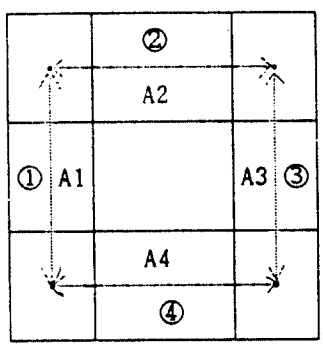
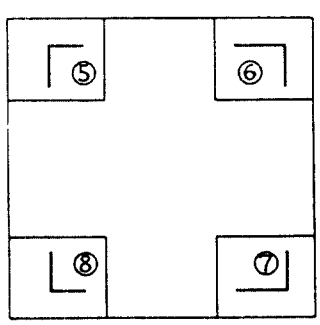


그림 10. 형식분류 처리 흐름도
Fig. 10. Flowchart of Type Classification



(a) 외곽선 조사를 위한 영역



(b) 내인점 조사를 위한 영역

그림 11. 형식분류를 위한 조사영역
Fig. 11. Test area for type classification

친 부분은 특징 영역을 나타낸다.

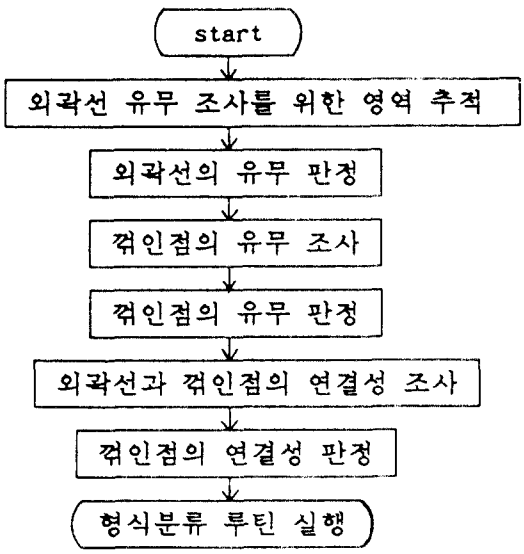
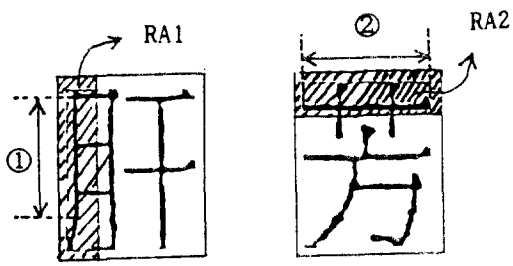
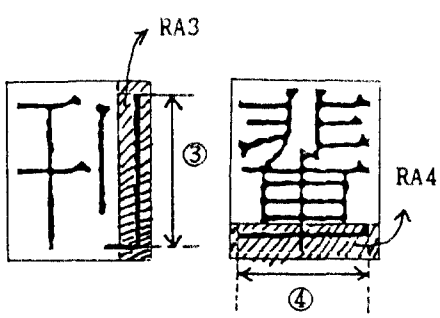


그림 12. 특징추출 처리도
Fig. 12. Flowchart of feature extraction

(가) 외곽선 유무조사(Flag A setting 유무)

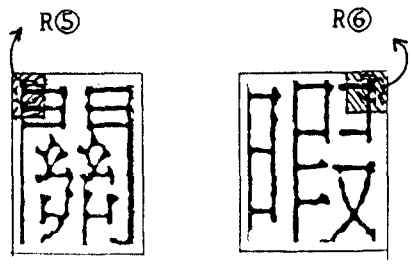


(a) a1 세팅 (b) a2 세팅

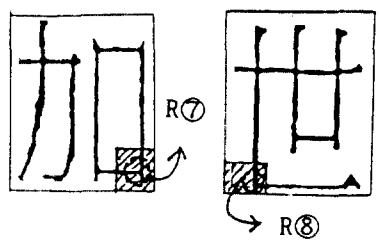


(c) a3 세팅 (d) a4 세팅

(나) 꺾인점 유무조사(Flag B setting 유무)

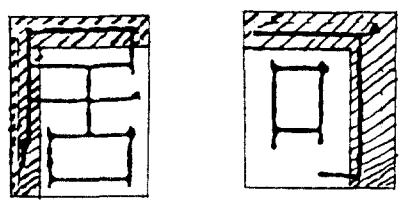


(e) b1 세팅 (f) b2 세팅

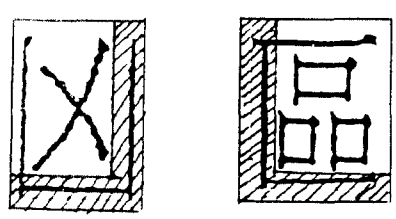


(g) b3 세팅 (h) b4 세팅

(다) 외곽선과 꺾인점의 연결화소 조사 (Flag C setting 유무)



(i) c1 세팅 (j) c2 세팅



(k) c3 세팅 (l) c4 세팅

그림 13. 특징 추출의 예
Fig. 13. Example of feature extraction

```
( 사용변수 정의 )
CS : Character segment(문자 세그먼트)
RA1 : Region of A1 (A1의 영역): X축면적의 20%이하
RA2 : Region of A2 (A2의 영역): Y축면적의 28%이하
RA3 : Region of A3 (A3의 영역): X축면적의 80%에서 100%사이
RA4 : Region of A4 (A4의 영역): Y축면적의 90%에서 100%사이
XL : X Length (정방형 문자의 X축길이)
YL : Y Length (정방형 문자의 Y축길이)
①'s length(①의 길이) : YL의 57%이상이 되는 길이
②'s length(②의 길이) : XL의 75%이상이 되는 길이
③'s length(③의 길이) : YL의 75%이상이 되는 길이
④'s length(④의 길이) : XL의 57%이상이 되는 길이
R⑤ : Region of ⑤ : RA1 과 RA2의 교집합 영역
R⑥ : Region of ⑥ : RA2 와 RA3의 교집합 영역
R⑦ : Region of ⑦ : RA3 과 RA4의 교집합 영역
R⑧ : Region of ⑧ : RA1 과 RA4의 교집합 영역
```

1.3. 형식분류 알고리즘

(1)3-11 형식 분류 알고리즘(특징이 만족되면 flag 세팅(=1))

i) ((Flag c1=c2=c3=c4=1))이면
→ 11 type
((Flag c1=1) && (Flag c4=1))이면
→ 10 type
((Flag c3=1) && (Flag c4=1))이면
→ 9 type
((Flag c1=1) && (Flag c2=1))이면
→ 8 type
((Flag c3=1) && (Flag c1=c2=c4≠1))이면
→ 7 type
((Flag c2=1) && (Flag c1=c3=c4≠1))이면
→ 6 type
((Flag c1=1) && (Flag c2=c3=c4≠1))이면
→ 4 type

ii) a1이 리세트, a4가 세트되면서 (0, y1/2)에서 (x1/5, y1)인 영역에 1 2, 3, 4 방향 세그먼트가 1개씩 존재하며 일정 위치정보를 가지면 5 type

iii) a1, a3와 b1, b2가 세트되면 y축 히스토그램을 탐색하여 Y길이의 50% 이내에 Y축 히스토그램의 크기가 X길이의 80%인 기둥이 3개 이상 존재하면 3 type
(예)困 : c의 모든 flag가 세트 11 type
癇 : flag c1만 세트, 나머지는 reset 4 type

iv) Sub type 분류 알고리즘으로 간다.

(2) X축, Y축 히스토그램 (Histogram) 작성

그림 14의 농도 히스토그램은 X축 및 Y축을 기준으로 하여 문자영역을 사영하였을 때, 문자상의 점들의 분포를 나타낸 것이다. 그림에서 우측 히스토그램은 실제 Y축 투영의 히스토그램을 +90° 회전시킨 것이다. 본 방법에서는 이들 점들의 분포를 이용하여 1, 2의 형식을 분류하였다.

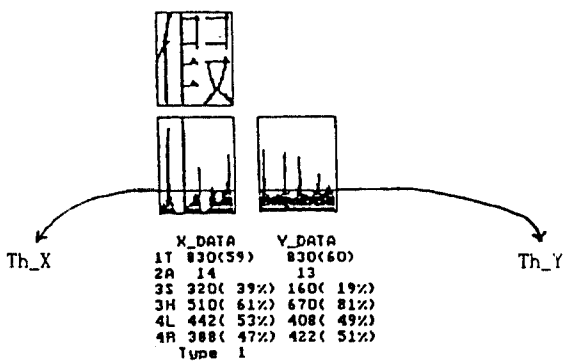


그림 14. Sub type 분류 예
Fig. 14. Example of Subtype classification

(3) 1, 2 type 분류 알고리즘

i) 그림 15에서 처럼 신으로 표시된 threshold (Th_X, Th_Y)를 경계로 하여 PX_S, PY_S와 PX_H, PY_H를 구한다.

ii) ((PX_H)<PY_H) && (PX_H>Th_3)) 이면 type=1

iii) ((PX_H)<PY_H) && (PY_H>Th_3)) 이면 type=2

iv) step 2, 3이 아니면 12 type으로 판정하되, PX_S가 PY_S 보다 크면 1 type, 그 반대이면 2 type으로 판정.

v) step ii, iii, iv에 다 해당되지 않으면 12 type(부정의 type)으로 판정.

사용 변수 정의
Th_X : x축 Histogram상의 모든 점면적을 xlength로 나눈 값 Th_Y : y축 Histogram상의 모든 점면적을 ylength로 나눈 값 (그림 14에 그 예를 나타냄) PX_S : x축 Histogram상의 상부 점면적의 비율 PY_S : y축 Histogram상의 상부 점면적의 비율 PX_H : x축 Histogram상의 하부 점면적의 비율 PY_H : y축 Histogram상의 하부 점면적의 비율 Th_3 : 문자 전체 점면적의 42%의 값 (PX_H와 PY_H가 거의 같을 경우에 사용)

이상의 알고리즘에 의해서 형식분류된 데이타를 그림 15에 보였다.

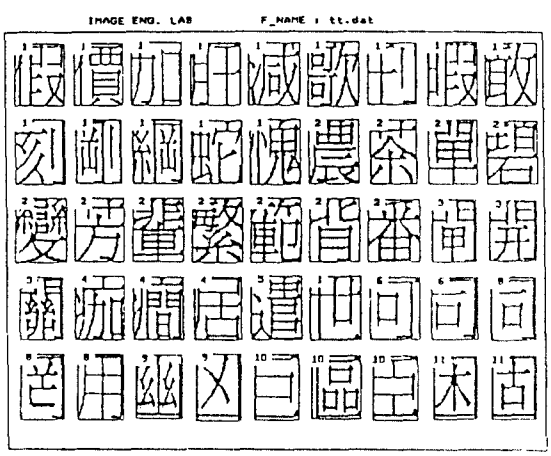


그림 15. 형식분류된 데이타
Fig. 15. Type classified data

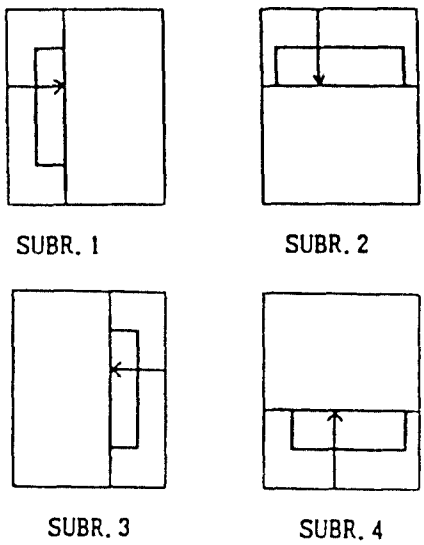


그림 16. Radical 구성 세그먼트 추출 영역
Fig. 16. Extraction area of Radical configuration segment

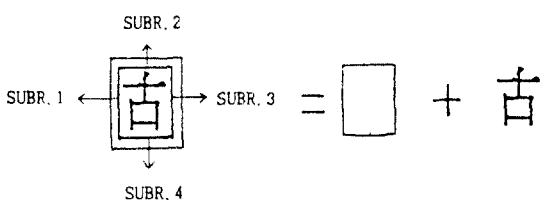


그림 17. 문자 패턴 분리
Fig. 17. Devision of character pattern

2. Radical(부분패턴)의 추출

형식 분류된 각 type에서 부분패턴(이하 SP)을 추출하기 위해서는 그림 16과 같은 영역할당이 필요하다. Radical 추출을 위해 다음 알고리즘을 사용한다.

2.1. 추출 알고리즘 1(3-11 type 추출) (그림 17 참조)

i) type 분류된 문자의 각 세그먼트의 위치, 방향정보를 미리 기억해 둔다(레이블링시 저장).

ii) 각 type의 Radical 추출영역을 설정.

iii) 영역 설정 알고리즘을 수행.

iv) 각 SUBROUTINE으로 부터 구한 추출영역에서 레이블링시에 기억해둔 정보 세그먼트와 동일한 정보가 있으면 그 세그먼트를 추출.

v) SP 추출 알고리즘 2를 수행한다.

2.2. 영역 설정 알고리즘

i) 그림 16처럼 수직 또는 수평 화면에서 정량적으로 임의의 한점을 고정시키고, 그 점부터 값을 증가시키면서 스캔한다.

ii) 스캔하는 중에 3×1 또는 1×3의 마스크의

조건에 맞으면 그때의 마스크 중심점까지를 추출영역으로 설정한다.

「5 type의 경우는, A2영역이 더 넓어지는 경우("尺" 같은)를 고려하여 추출영역을 설정한다.」

iii) SP 추출 알고리즘 1의 iv)로 간다.

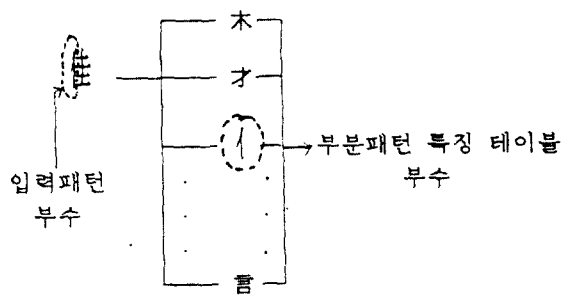
2.3. SP 추출 알고리즘 2(1, 2 type 추출)

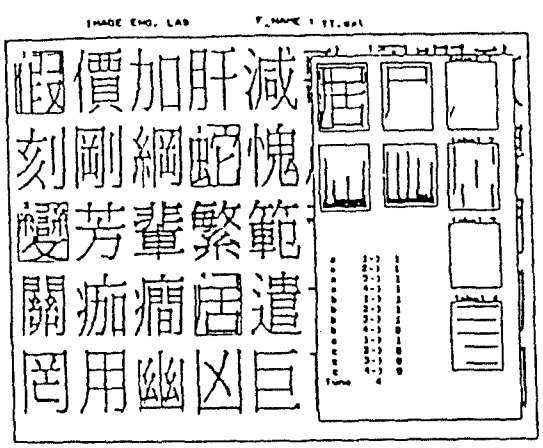
i) 1 type 부수들의 표준 특징 테이블을 만든다. 그 특징으로는 세그먼트의 갯수, 방향정보, 길이정보, 중심좌표가 있다.

ii) 2 type 부수들에 대해서 step i)반복.

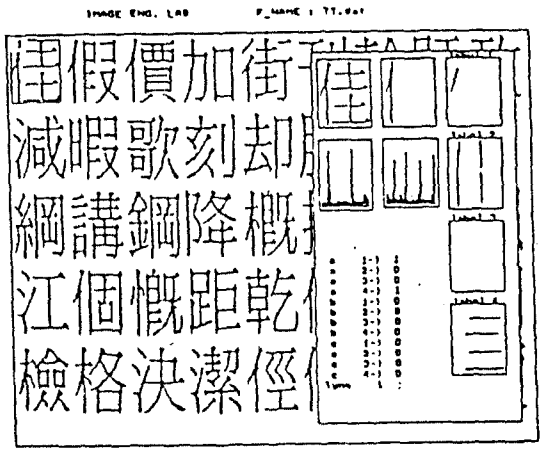
iii) 분류된 type이 1 type이면 step i)에서 만든 특징 테이블과 입력 패턴의 세그먼트 특징들을 비교하여 가장 근접 세그먼트를 추출.

(Ex) 1 형식 "隹"의 경우





(a) 1형식 (隹)의 부분패턴 추출 예



(b) 1형식 (隹)의 부분패턴 추출 예

그림 18. 부분패턴의 추출 예
Fig. 18. Example of subpattern extraction

표 1. 부분패턴 특징 테이블 작성 예
Table 1. Example of subpattern feature table

형식	부수	방향	세그먼트수	중심좌표	길이
좌우 구조	亻	1	1	(6.20)	24
		2	1	(9.32)	59
	口	1	1	(8.37)	16
		2	1	(13.32)	65
		3	1	(12.14)	25
	木	1	1	(5.32)	26
		2	1	(10.32)	65
		3	1	(13.27)	20
		4	1	(11.14)	21
	制	2	2	(56.31)	42
				(42.28)	62
상하 구조	花	2	2	(22.12)	11
				(22.36)	11
		1	1	(58.29)	27
	教	1	2	(35.20)	30
				(44.44)	40
		3	1	(49.47)	35
		4	1	(49.15)	27

iv) 2 type에 대해서 iii)을 반복.

표1은 1, 2 type 부수들의 특징 테이블 작성

예를 나타낸 것이고, 그림 18에는 그래픽 화면상에서의 부분패턴 추출예를 나타내었다.

V. 실험 및 고찰

1. 실험 시스템

본 실험에 사용된 대상 문자 패턴은 KS C 5601 표준 심보 LBP 한자 4,888자와 중, 고등학교 교육용 한자 1,800자이며, SQ(System Quality) 사의 IS 300 image scanner로 부터 240 dpi 해상도의 문자 데이타를 받아 인터페이스를 통해 NEC 9801 computer에 입력한다.

입력된 문자는 RS 232 C 인터페이스를 통해 IBM PC / 386 machine에 640×400의 데이타 크기로 전송하여 처리하였다. 처리에 사용된 언어는 C Language를 사용하였으며, 그림 19에 본 실험 시스템의 구성도를 보였다.

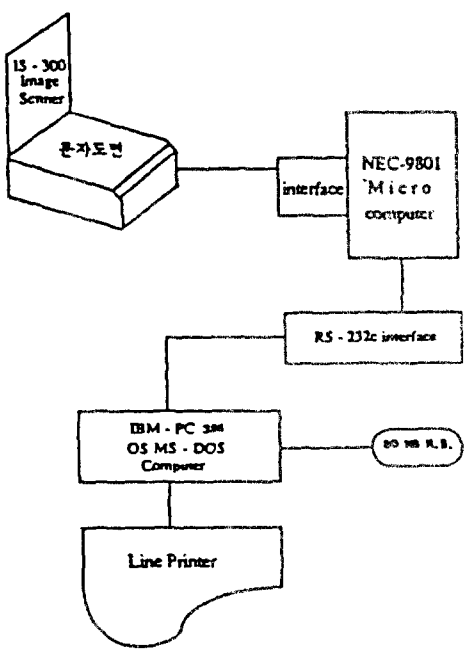


그림 19. 실험 시스템의 구성도
Fig. 19. Block diagram of Experiment system

2. 실험 및 결과 고찰

본 실험에서는 3-11 형식분류시 2가지 방법을 사용하여 비교실험을 하였는데, method 1(md 1)은 4장에서 언급한 외곽선 조사 영역을 그대로 사용한 방법이고, method-2(md-2)는 RA1, RA 2, RA3, RA4를 각각 5%씩 줄인 영역을 사용한 방법이다. 실험을 통해, 표2에서 처럼 md 1의 유효성을 입증할 수가 있었다. 표3에는 대상한자 중 분할가능한 문자와 불가능한 문자(무정의 형식)의 갯수를 표시하였다.

KSC 5601 표준한자 4,888자와 교육용 한자 1,800자를 대상으로 12가지 형식으로 분류한 형식분류율과 분류된 데이타로부터 추출한 부분패턴 추출율은 표4와 같다.

1,2 형식 분류실험시, 문자 밀도가 높은 경우(문자가 복잡한 경우) 오분류가 많이 발생했고, 나머지 형식의 분류에서는 주 분류원인 직선성분이 곡선부와 접하는 등 굴곡이 심한 경우 정의한 Flag를 세팅시키지 못하여 오분류가 발생하였다. 이와 같이 데이타 질이 나쁜 경우에는 분류가 잘 되지 않았는데 이를 보완하기 위해서는 문자 세그먼트가 가지는 폭을 비교하여 적당한 값을 넘는 경우, 강제적으로 세그먼트를 분리하는 방법등이 고려되어야 할 것이다. 또한 1, 2형식의 분류시의 오류를 줄이기 위한 해결방법으로 가능한한 1, 2 형식의 복합 형태등의 새로운 형식의 추가도 고려해 볼만하다.

부분패턴 추출시 오추출이 발생하기 쉬운 문자중은『驗』처럼 짧은 세그먼 문자는 부수가 단순한 구조를 가진 것, 즉『付』,『廣』,『土』,『抽』,『川』,『元』과 같 세그먼트로 구성되어 있는『口』등의 부수로 되어 있는 문자들이다.

12형식(무정의 형식)의 경우는 특정부분패턴이 없으므로 바로 인식부로 넘겨서 특징 테이블과 비교하여 인식하면 된다. 그리고 나머지 형식분류시의 에러는 수직성분과 사선성분의 연결점을 찾아서 연결되는 두 성분을 직선화시키는 방법을 적용시키면 좀 더 유효한 추출이 이루어지리라 생각된다. 표 5에는 오분류된 데이타와 오추출된 데이타의 예를 나타내었다.

표 2. method 1과 method 2의 비교
Table 2. Comparison of method 1 and method 2

대상 / 방법	표준한자 4888자(%)		교육용 한자 1800자(%)	
형식	md 1	md 2	md 1	md 2
3	97.2	85.3	98.6	86.2
4	92.4	84.2	97.5	85.5
5	91.1	82.5	93.5	83.1
6	82.3	79.8	84.2	80.6
7				
8	86.4	80.4	87.6	81.2
9	89.8	82.5	90.3	84.1
10	98.2	83.2	100	88.7
11	90.1	85.5	90.2	85.5

*7형식은 자형이 너무 작아 제외

표 3. 대상 한자의 분할 가능성
Table 3. Partition capability of the object Hanja

대 상 / 분 할	표준 한자 (4888자)	교육용 한자 (1800자)
가 능	4523자	1747자
불 가	365자	53자

표 5. 오분류된 데이타와 오추출된 데이타의 예
Table 5. Example of misclassified data and misextracted data.

형 식	오분류된 데이타	오추출된 데이타
1	幹(2), 個(7), 技(4), 奴(12)	拾, 減, 網,…
2	甲(4), 多(12), 罰(6)	姦, 看,…
3	鬪(1)	開
4	斤(1), 斥(1)	屋, 庸
5	勉(1), 魅(1)	邊, 邈
6	刀(1), 力(12), 耳(4)	句, 匀
7	대상 데이타가 아주 적음	…
8	丹(1), 鳳(4), 商(2)	再
9	甘(1), 曲(11), 面(5)	西
10	없 음	…
11	助(1), 血(8)	血

Ⅵ. 결 론

본 논문에서는 KS C5601 표준한자와 중, 고등학교 교육용 한자를 대상으로 하여 문자의 형식분류와 분류되어진 문자로 부터 부분패턴을 추출

표 4. 각 형식의 분류율 및 부분패턴 추출율
Table 4. Cassification rate of type and Subpattern extraction.

대 상	KS C5601 표준 한자 4888자		교육용 한자 1800자	
형 식	형식 분류율	부분패턴 추출율	형식 분류율	부분패턴 추출율
1	85.2	84.3	90.5	87.5
2	88.4	86.5	97.8	88.5
3	97.2	95.2	98.6	98.2
4	92.4	91.2	97.5	92.4
5	91.1	90.6	93.5	88.7
6	82.3	81.5	84.2	80.4
7				
8	86.4	82.5	87.6	85.2
9	98.8	84.4	90.3	87.4
10	98.2	97.5	100	99.2
11	90.1	85.4	90.2	87.6

※12형식은 무성의 형식이므로 제외 (단위 : 백분율)

하는 연구를 하였다. 한자패턴의 구조적인 특성을 고려하여 모두 12개의 형식으로 분류하여 기존의 방법과는 전혀 다른 새로운 방법을 제안하였다. 형식분류시 이들 형식이 가지는 각 부분의 구조적인 정보를 이용하였고, 형식분류된 데이타에 대해서 인식의 중간단계인 부분패턴을 추출하였다.

KS C5601 표준한자 4,888자와 교육 한자 1,800자를 대상으로 분류를 행한 결과 각각 90.12%, 93.07%의 형식분류율과 분류된 형식으로부터 각각 87.91%, 89.5%의 추출율을 얻었다. 본 논문은 한자 인식을 위한 전단계로서 인식부로의 단계적 매칭을 위한 적용 가능성을 발견하였다.

향후과제로, 부분패턴 이외의 부분과의 단계적 매칭을 위해 입력문자 스트로크의 획순에 의한 스트로크 추출 및 분석을 고려한다든지, 문자 세그먼트를 상호관계등의 연구를 지식 베이스적인 시스템 환경에서 수행할 수 있는 보다 세심한 연구가 검토되어야 할 것이다.

參 考 文 獻

1. J.K. Lee, "Korea 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 부분분리와 인식에 관한 연구", 영남대학교 박사학위 논문, 1982.
3. 이주근, 남궁재찬, 김영건, "한글 Pattern에서 Subpattern 분리와 인식에 관한 연구", 대한전자공학회지, Vol. 18, No. 3, 1983.
4. Guo Chunbiao, Xuan Guorong, "Automatic recognition of printed Chinese characters by four corner codes", Xian Jiaotong University, China, 1986.
5. Zhang Xinzhong, Xia Ying, "The automatic recognition of handprinted Chinese characters A method of extracting an ordered sequence of strokes", Qinghua University, Peking, China, 1983.
6. C.H. Leung, Y.S. Cheung, Y.L. Wong, "A Knowledge Based stroke Matching Method for Chinese Character Recognition", IEEE Trans. On System, Man, and Cyber. Vol. SMC 17, No. 6, Nov./Dec. 1987.
7. Naoki Tanaka, Yoshinori Sakurai, Hiromi Aota, Hidchiko Sanada, Yoshikazu Texuka, "A Basic Study of Handprinted Kanji Characters Recognition Method Based on Identification of Sub-patterns", Osaka University, 信學技報, PRL 83-81.
8. Noboru Babaguchi, Tsunehiro Aibara, Hidehiko Sanada, Yoshikazu Tezuka, "Identification and extraction of radicals from hand-printed KANJI characters by segment correspondence method", Ehime Univ., Osaka Univ., 信學技報, PRL 83-59, 1983.
9. Yoshiyuki Yamashita, Koichi Higuchi, Youichi Yaqmada, Yunosuke HAGA, "Classification of handprinted Kanji characters by the structured segment matching method", PRL, July, 1983.
10. Noboru Babaguchi, Yoshihiro Kitamura, Mitsuru Shiono, Hidehiko Hidehiko Sanada, Yoshikazu Texuka, "A Method of Direction Segments Extraction from Character Pattern without Thinning Process", Osaka University, 信學論, Vol. 65-D No. 7, 1982.
11. Noboro Babaguchi, Tsunehiro Ajbara, Hidehiko Sanata, Yoshikazu Tezuka, "On Identification and Extraction of Radicals from Handprinted KANJI Characters by Structural Segment Matching", Osaka University, Japan., I.E.C.E., Vol. 68, 1985.
12. Yasuaki Nakano, Kazuo Nakata, "Recognition of Chinese Characters using Peripheral Distributions and their Amplitude Spectra", 日本電子通信學會論文誌, Vol. 56-D, No. 3, 1973.
13. Norihiro Hagita, Isao Masuda, "Classification of Handprinted Chinese Characters by Global and Local Stroke Density", 信學技報, PRL 80-23, 1980.
14. Toshiaki Ejima, Yosuke Nakamura, Masayuki Kimura, "The Characteristic Feature Based on Four Types of Structural Information and Their Effectiveness for Character Recognition", 日本電子通信學會論文誌, Vol. J 68-D No. 4, 1985.
15. 남궁재찬, "畵像工學의 基礎", 청전 인쇄사, 1989.

金 政 漢(Jeong Han KIM) 正會員
1965년 2월 1일생
1989년 2월 : 광운대학교 전자계산기공학과 졸업(공학사)
1991년 2월 : 광운대학교 대학원 전자기공학과 졸업(공학석사)
1991년 ~ 현재 : (주)한국산업기술본부 연구3실 근무중
관심분야 : 패턴인식

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

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 Tohoku대학 객원교수
1979년 ~ 현재 : 광운대학교 전자계산기공학과 교수
관심분야 : 패턴인식, 컴퓨터비전, 인공지능