[5-6]

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

°김정한, 조용주, 이성범, 남궁재찬
광운대학교 전산기공학과 화상공학연구실

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

°J.H.Kim, Y.J.Cho, S.B.Lee, J.C.Namkung
Dept. of Computer Eng., Image Engineering Lab.
University of Kwang Woon

ABSTRACT

This paper is proposed on new classification algorithm using characteristic and structural information of printed hanja as preliminary stages for 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 structual 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 saw that as extracting subpattern at classified data was this paper possibly applied to the recognition.

1. 서 론

최근에 들어서 사회가 급속히 변화함에 따라 정보처리 시스템의 개발에 많은 관심이 고조되고 있다. 컴퓨터와 사람과의 맨 머신 인터페이스(man machine interface)를 통한 정보통신을 위해서 무엇보다도 중요한 일은 사람들이 사용하기 용이한 통신 수단으로 하는 것이다. 이렇듯 최근의 O.A.(office automation)와 정보처리 환경에서도 많은 발전을 가져오고 있으며 활발한 연구가 진행중에 있다. 그중에서도 패턴인식 분야의 한 부류인 문자인식 분야의 연구는, 컴퓨터상에 대량의 문서를 고속으로 받아들이는 경우, 타이프라이터(typewriter)나 워드 프로세서(word processor)의 사용만으로는 처리속도에 한계가 있기 때문에, 그 필요성이 증대되어 왔다.

중국, 일본, 동남아의 여러나라와 같이 한자문화권인 우리나라에서의 한자 인식에 관한 연구는 연구의 진행속도가 초보적인 단계에 머물러 있는 실정이다. 그러나 문서인식중의 하나라고 생각할 수 있는 신문 문서인식을 위한 측면에서나, 또는 현재 한참 연구중에 있는 전자출판 시스템(Desk Top Publishing)과 같은 문서 인식 연구에서도 결코 배제할 수 없는 것이 한자인식에 관한 연구라고 할 수 있다. 한자는 글자의 수가 2만자가 넘는 방대한 문자로 구성되어 있다. 한자인식을 하기위해서는 먼저 한자의 구조적인 형태를 살핀다음, 그 형태가 지니는 특징들을 살펴야 할 것이다. 우리는 일상생활, 예를 들면 한자사전을 참조하여 문자를 찾을때에 부수색인에 의해 문자를 구하곤 한다. 한자는 부수(Radical)을 포함하는 것이 대부분이므로 이러한 특성에 착안하여 분류를 한다음 인식을 행함이 유효하다. 한자인식시의 문제점은 자종이 광대하다는 양적인문제와 대상자형의 구조가 복잡하며 유사문자가 많다고하는 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 부분패턴 분류방법에서는 국소적인 유사성에 기초하거나 또는 세그먼트의 구조적 매칭에 의한 구조해석적인 방법에 기초하여 행하기 때문에 특정 부분패턴이외는 분류가 잘 되지 않는 경우가 많았다.

본 논문에서는 이러한 한자 구성요소의 구조적 특성과 주변분포에 따른 문자 히스토그램의 국소적인 유사성을 함께 도입하여 부분패턴을 추출하였다. 먼저 한자패턴은 방향을 갖는 선소(방향 세그먼트:direction segment)의 집합으로 구성되어 있으므로 각 방향별 세그먼트를 구하고, 각 문자에 해당하는 세그먼트의 영역을 조사하여, 한자패턴을 12종류의 form으로 형식분류 한다. 또한 분류율을 높이기 위해 각 문자 세그먼트의 위치관계나 길이, 방향등의 지식을 이용하여 분류된 형식으로부터 부분패턴을 추출하였다. 이러한 방법으로부터 문자패턴상에서 추출된 부분패턴의 존재 영역과 그 이외의 영역과의 지식 베이스적인 측면에서의 단계적 인식이 향후 과제라고 할 수 있다.

2. 한자 패턴의 구조분석

대상 문자인 한자패턴을 인식하기 위한 인식 시스템을 설계하기 위해서는 한자구조와 구성에 관한 특징이 선행되어야만 한다. 따라서 본 장에서는 한자의 역사적인 생성과정을 고찰하고 그 구조및 특성에 대해 기술하였다.

2.1 한자의 역사적인 생성과정

중국 문자의 구성은 그 생성과정에 따라 분류하면 크게 6가지(육서)로 나눈다. 표 1은 6서에 의한 한자의 자구성을 나타낸다. 그림 1은 명조체를 기준으로한 엘리먼트(element)를 나타낸 것으로 각각의 엘리먼트는 나름대로의 이름과 개성을 가지고 있다. 한자는 한글처럼 조합 문자인데 대개 하나이상의 세그먼트(segment) 로 구성된 것이 그 특징이다. 또한 한자는 한글처럼 정방형(사각형)의 형태로 이루어져 있음을 알 수 있다.

표 1. 한자의 자구성

⊙상형문자 : 물체의 형태를 모방하여 간략화한 문자로 그 자신이 각각 原義(원뜻)와 音을 지니고 있을 뿐만 아니라 다른 글자의 변, 발등에 쓰여 새로운 문자를 구성한다.

(예) 牌 → ⺧ , 竹 → 竹

⊙지사문자 : 象形을 기본으로하여 의미에 따라 자획을 증감시켜 만든 文字

(예) 朝 → 旦

⊙회의문자 : 既成의 문자를 결합하여 音에 관계없이 뜻만을 취하여 만든 文字

(예) 日 + 月 → 明 , 力 + 田 → 男

⊙형성문자 : 두개 이상의 既成文字를 합하여 만든 점은 회의문자와 동일하나 한은 의미를 나타내고, 반은 音을 표시한 문자

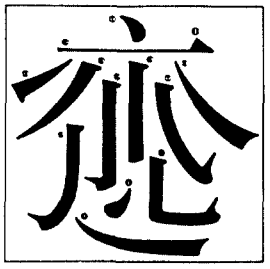
(예) 門 + 口 = 問 , 亡 + 女 = 妄

⊙전주문자 : 어떤 자의 본뜻을 확대시켜 새로운 의미내용을 도출하는 문자

(예) 道 → 道理, 道路 樂 → 音樂, 娛樂

⊙가차문자 : 原字의 音을 빌려서 그와 同音의 다른 의미를 가지는 문자

(예) 皮革 → 皮革



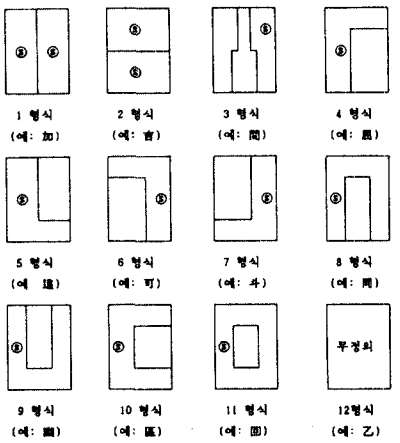
엘리먼트의 명칭
①점
②가로선
③왕침선
④좌 내림선
⑤세로 삐침
⑥세로선
⑦우 꺾어 삐침
⑧우 내림선
⑨좌 꺾어 삐침
⑩마감 삐침
⑪악점
⑫가로 내림선
⑬혹음

그림 1. 한자의 엘리먼트와 그 명칭

2.2 한자의 구조 및 특성

한자의 구조를 살펴보면, 台와 같은 상하구조(North-south structure), 化와 같은 좌우구조(East-west structure), 困과 같은 내외구조(Border-Interior), 그리고 이 세 구조의 복합구조등으로 구성되어 있는 경우가 아주 많다. 한자는 자종이 광대하고 구조가 복잡하며, 유사문자가 많이 존재하기 때문에 인식은 물론이고 분류하는데에도 많은 어려움이 따른다. 그래서 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.

한자의 도형적 계층성에 의거해 그림 2와 같이 12가지 형식으로 분할을 시도하였다.



⑧ : 부분패턴 존재 가능 영역

그림 2. 한자의 12가지 형식

3. 전처리와 세그먼트추출 및 레이블링

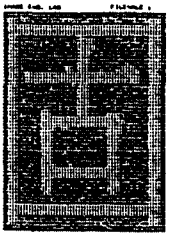
3.1 영상의 전처리

영상 입력장치를 통해 얻어진 문자 패턴 정보는 입력시의 용지의 오염 및 센서부등에 의한 잡음이 혼입되는데 이러한 잡음을 인식과정으로 들어가기 전에 제거하는 과정을 전처리라 부른다. 본 논문에서 사용한 인쇄체 문자의 전처리에는 다음의 방법이 시도된다.

3.1.1 양 자 화 및 2 치 화

영상 입력 장치인 image scanner를 통해 입력된 문자 패턴은 배경부를 0 , 문자부를 1로하여 2치화된다. 그림 3는 입력 패턴을 양자화한 예를 나타내었는데 (a)는 입력 패턴을 (b)는 양자화된 패턴을 나타낸 것이다.

(a)



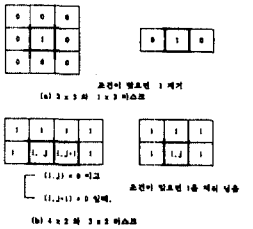
(b)

그림 3. 입력패턴 및 양자화된 패턴

3.1.2 평 활 화 (잡 음 제 거)

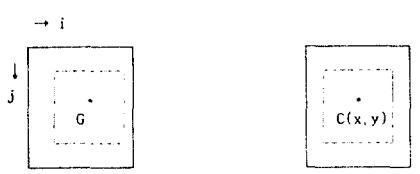
입력된 문서는 2차원 디지탈화에 따른 문서상의 잡음과 하드웨어상의 잡음을 제거하기 위해 평활화(Smoothing)처리를 하여 매끄러운 패턴으로 만든다. 3 x 3 마스크(mask)를 사용하여 고립점을 제거하였으며 1 x 3 마스크를 사용하여 세그먼트(segment)들의 불음을 최대한 방지하였다. 그림 4에 평활화 처리의 예를 나타내었다.

그림 4.
평활화 처리를 위한 마스크



3.1.3 정 규 화 (Normalization)

문자의 정규화는 인쇄체 매칭법에 있어서 특히 중요한데, 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.



(a) (b)

—— : 최대 장방형 (종축 i × 횡축 j = 70 × 70)
--- : 문자 영역

그림 5.. 문자 위치의 정규화

3.2 문자의 방향 세그먼트 추출과 레이블링

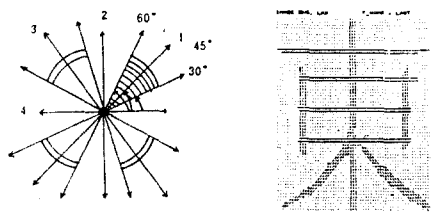
3.2.1 코드화

문자패턴에 대해 각 스트로크마다 방향성을 부여하기 위하여 문자패턴을 코드화한다. 본 논문에서는 정방격자 상에서 간격이 $\sqrt{2}$이내이고 등간격에 나란한 방향으로 θ_k = 0°, 45°, 90°, 135° 즉 수평, 대각, 수직, 역대각의 4개 방향만을 고려한다. 센서가 연결성을 잃어버릴때 기준점(x_0, y_0)로부터 특정각도 θ_k방향의 기하학적 거리$L^+(x_0, y_0)$는,식(1)(2)와 같이 된다.

$$L^+(x_0,y_0) = \sqrt{(x-x_0)^2+(y-y_0)^2} \quad \cdots\cdots\cdots\cdots (1)$$

$$d_k(x_0,y_0) = L^+(x_0,y_0) + L^-(x_0,y_0) \quad \cdots\cdots\cdots\cdots (2)$$

그림 6.(a)는 각 방향세그먼트의 각도,방향코드, 그림 7(b)는 방향 코드화 패턴을 보여준다.



(a) (b)

그림 6. 방향 코딩의 예

3.2.2 방향화면의 생성

그림 7.에 그림 6의 방향 코드화 패턴으로부터 생성된 방향화면의 실례를 보여준다.

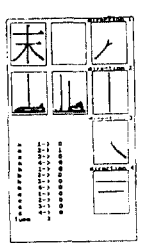
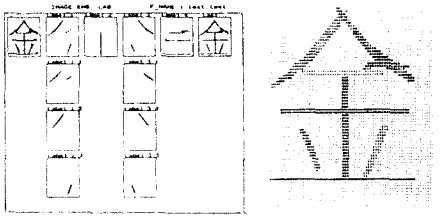


그림 7. 방향화면의 작성

3.3 추출된 방향세그먼트의 레이블링 (Labeling)

4 개의 방향 화면으로부터 추출된 세그먼트에 대해서 레이블 값 m(m =1, · · ·, M : M은 총 세그먼트수)을 주어 그 순서를 부여하는 레이블링을 행한다.

그림 9의 (a)는 그래픽 화면상에서의 한 문자에 대한 레이블링 과정을 나타낸 것이고, (b)는 레이블링된 결과를 텍스트 화면상에 나타낸 것이다.



(a) (b)

그림 8. 문자 세그먼트의 레이블링

제 4 장 한자의 형식분류 및 Radical의 추출

4.1 한자의 형식분류

복잡한 구조를 가지는 한자를 인식하기 위해서는 먼저 인식의 전단계인 대분류가 필요하다. 그래서 본 논문에서는 문자가 가지고 있는 고유한 구조를 분석하여 12가지(11가지의 분할 가능한 형식과 1가지의 분할 불가 형식)으로 형식 분류를 하였다.

4.1.1 문자의 블럭화

각각의 문자들이 모두 하나의 정방형으로 구성 가능하므로 문자등의 위치등을 파악하기 위해서 블럭화를 행하고, 문자의 블럭을 구성한 결과를 그림 9.에 보였다.

그림 9. 한자 데이타의 블럭화

4.1.2 형식분류 알고리즘

구조적인 특성을 고려하여 먼저 3 – 11 형식을 분류한 후 여기서 분류가 되지 않는 문자에 대해서 1 또는 2 형식으로의 분류를 행하였다. 형식분류 알고리즘을 설명하기 위해서는, 그림 10과 같이 특정영역에 세그먼트가 존재하는지의 여부를 판단하는 2개의 플래그 세팅 영역과 3개의 플래그(A, B, C)가 필요하다. (a)는 플래그 A를 세트하기 위한 영역이고, (b)는 플래그 B를 세트하기 위한 영역이다.

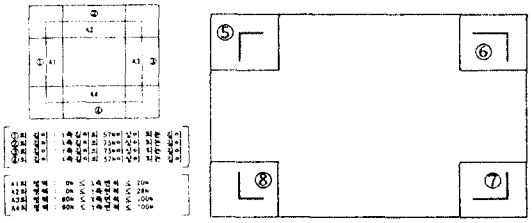
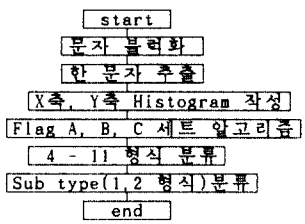


그림 10. 형식분류를 위한 조사영역

Flag A, B, C 세트 알고리즘

step 1. (a)그림에서 A1(A2, A3, A4)영역에 존재하면서 ①(②, ③, ④)의 길이를 만족하면 Flag A1(A2, A3, A4)을 세팅시킨다.
step 2. (b)그림에서 ⑤(⑥, ⑦, ⑧)영역내를 스캔하면서 꺾인점의 화소를 찾은 다음 그 화소를 중심으로 우(또는좌)로 10화소이상, 하(또는상)로 10화소이상 연결되는 흑화소가 존재하면 Flag B1(B2, B3, B4)을 세팅시킨다.
step 3. B1(B2, B3, B4)이 세트되어 있으면서 step 2와 같은 방법으로 조사하되 10화소가 아니라 그점으로부터 연결된 점열이 X길이의 75%(75%, 57%, 57%) Y길이의 57%(75%, 75%, 57%)이상이면 Flag C1(C2, C3, C4)이 세팅된다.

형식 분류 알고리즘



형식분류 알고리즘

step 1. (1)Flag C1, C2, C3, C4가 모두 세트되면 11 type이 된다.
(2)Flag C1, C4만 세트되면 10type이 된다.
(3)Flag C3, C4만 세트되면 9 type이 된다.
(4)Flag C3, C4만 세트되면 8 type이 된다.
(5) Flag C3만 세트되면 7 type이 된다.
(6) Flag C2만 세트되면 6 type이 된다.
(7) Flag C1만 세트되면 4 type이 된다.
step 2. Flag A1이 세트되지 않고 A4가 세트되면서 (0, ylength/2)에서 (xlength/5, ylength)인 영역에 각각 1, 2, 3, 4방향 세그먼트가 1개씩 존재하며 일정한 위치정보를 가지면 5 type이 된다.
step 3. Flag a1, a3와 b1, b2가 세트되면 y축 히스토그램을 탐색하여 Y길이의 50%이내에 Y축 히스토그램의 크기가 X길이의 80%이상인 기둥이 3개이상 존재하면 3 type이 된다.
step 4. Sub type분류 알고리즘으로 간다.

Sub type 분류 알고리즘

step 1. 그림 11에서 볼 수 있듯이 라인으로 표시된 threshold를 경계로하여 각 히스토그램상의 상부 및 하부 점면적의 비율 PX_H, PY_H와 PX_L, PY_L를 구한다.
step 2. PX_L과 PY_L이 거의 같을 경우를 위해 또 하나의 threshold를 구한다. (Th_3)
step 3. PX_L가 PY_L보다 작고 Th_3보다 크면 1 type으로 판정한다.
step 4. PX_L가 PY_L보다 크고 Th_3보다 크면 2 type으로 판정한다.
step 5. step 3, step 4이 아니면 12 type으로 판정하되, PX_S가 PY_S보다 크면 1 type으로 판정하고, 그 반대이면 2type으로 판정한다.
step 6. step 3, 4, 5에 다 해당되지 않으면 12 type(부정의 type)으로 판정한다.
threshold는 x, y축상의 모든 점면적을 각각 xlength, ylength로 나눈 값이다.
(Th_X, Th_Y)

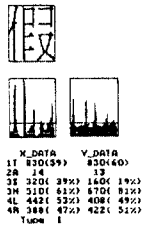


그림 11. Sub type 분류 예

이상의 알고리즘에 의해서 형식분류된 데이타를 그림 12에 보였다.

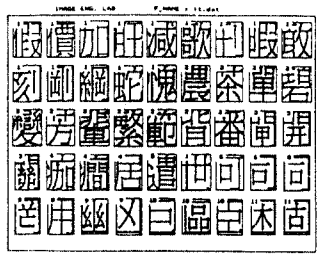


그림 12. 형식분류된 데이타.

4.2 Radical(부분패턴)의 추출

형식 분류된 각 type에서 부분패턴을 추출하기 위해서는 그림 13와 같은 영역할당이 필요하다. Radical추출은 2개의 서브 알고리즘을 사용한다.

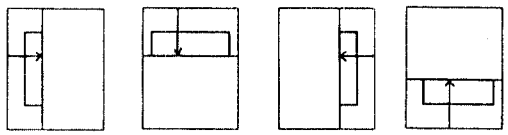


그림 13. Radical 구성 세그먼트 추출 영역

부분패턴 추출 알고리즘 1 (4 — 11 type 추출)

step 1. type분류된 문자의 각세그먼트의 위치, 방향정보를 미리기억해 둔다.(레이블링추출시)
step 2. 각 type의 Radical 추출영역을 임의로 찾아낸다.
step 3. 영역 설정 알고리즘을 수행한다.
step 4. 각 SUBROUTINE으로부터 구한 추출영역에서 레이블링시에 기억해둔 정보 세그먼트와 동일한 정보가 있으면 그 세그먼트를 추출한다.
step 5. 부분패턴 추출 알고리즘 2를 수행한다.

영역 설정 알고리즘

step 1. 그림 14처럼 수직 또는 수평 화면에서 정량적으로 임의의 한점을 고정시키고, 그점부터 값을 증가시키면서 스캔을 한다.
step 2. 스캔하는 중에 3×1 또는 1×3의 마스크의 조건에 맞으면 그 때의 마스크 중심점까지를 추출영역으로 설정한다.
(5 type의 경우는, A2영역이 더 넓어지는 경우를 고려하여 추출영역을 설정한다.)
step 3. 부분패턴 추출알고리즘 1의 step4로 간다.

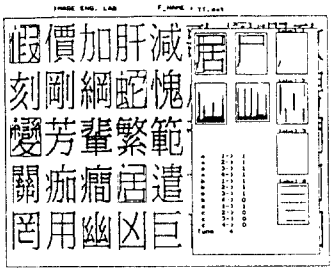
부분패턴 추출 알고리즘 2 (1,2 type 추출)

step 1. 1 type 부수들의 특징 테이블을 만든다. 그 특징으로는 세그먼트의 방향, 길이, 중심점이 있다.
step 2. 2 type 부수들에 대해서 step 1 반복
step 3. 형식 분류된 type이 1 type이면 step 1에서 만든 특징 테이블과 입력패턴의 세그먼트 특징들을 비교하여 가장 근접한 세그먼트를 추출한다.
step 4. type 2에 대해서 step 3를 반복한다.

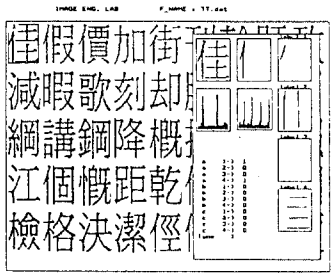
표 2은 1, 2 type 부수들의 특징 테이블의 작성예를 나타낸 것이고, 그림 15 에는 부분패턴 추출의 예를 나타내었다.

표 2. 부분패턴 특징 테이블 작성 예

형식	부수	입력	세그먼트수	중심좌표	길 이
좌우구조	亻	1	1	(8,20)	24
		2	1	(9,32)	58
	扌	1	1	(6,37)	18
		2	1	(13,32)	65
		3	1	(12,14)	25
	木	1	1	(5,32)	28
		2	1	(10,32)	65
		3	1	(13,27)	20
		4	1	(11,14)	21
	忄	2	2	(36,31)	43
				(42,28)	62
상하구조	艹	2	2	(22,12)	11
				(22,36)	11
		4	1	(50,28)	27
	灬	1	2	(35,20)	30
				(44,44)	40
		3	1	(40,47)	35
		4	1	(49,15)	27



(a) 4 형식(居)의 부분패턴 추출 예



(b) 1 형식(佳)의 부분패턴 추출 예

그림 14. 부분패턴의 추출예

5. 실험 및 고찰

5.1 실험

본 실험에 사용된 대상 문자 패턴은 KSC 5601 표준 삼보 LBP 한자 4888자와 중, 고등학교 교육용 한자 1800자이며, SQ(System Quality)사의 IS-300 image scanner로 부터 240 dpi 해상도의 문자 데이타를 받아 인터페이스를 통해 NEC-9801 Micro computer에 입력한다. 입력된 문자는 RS 232-C 인터페이스를 통해 IBM PC / 386 machine에 640 × 400의 데이타 크기로 전송하여 처리 하였다. 처리에 사용된 언어는 C-Language 를 사용 하였으며, 그림 15에 본 실험 시스템의 구성도를 보였다.

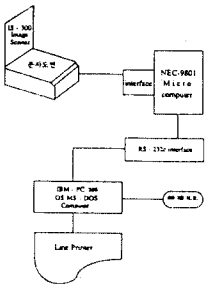


그림 15. 실험 시스템의 구성도

5.2 실 험 결 과

표 3에 나타낸 바와 같이 KSC 5601 표준 교육용 한자 1800자를 대상으로 12개 형식으로 분류한 결과 93.07%의 형식 분류율을 얻었으며, 분류된 형식으로부터 추출한 부분패턴 추출율은 89.5%를 얻었다. 1,2형식의 분류실험시 문자 밀도가 높은 경우(문자가 복잡한 경우) 오분류가 발생했고, 나머지 형식의 분류에서는 주 분류원인 직선성분이 곡선부와 겹치는 등 굴곡이 심한 경우에 정의한 Flag를 세팅시키지 못하여 오분류가 발생하였다.

표 3. 각 형식의 분류율 및 부분패턴 추출율

대상 / 형식	교육용 한자 1800자	
	형식 분류율	부분패턴 추출율
1	90.5	87.5
2	97.8	88.5
3	98.6	98.2
4	97.5	92.4
5	93.5	88.7
6	84.2	80.4
7	-	-
8	87.6	85.2
9	90.3	87.4
10	100	99.2
11	90.2	87.6

(단위 : 백분율)

5.3 고 찰

본 논문에서는 인쇄체 한자의 인식을 위한 전단계로서 한자의 직선성에 착안하여 문자내에 존재하는 세그먼트의 위치, 방향, 길이등의 정보를 분류의 주안점으로 사용하였다. 1, 2 형식의 오분류를 보완하기 위해서는 1형식과 2형식이 복합된 형태를 새로운 형식으로 추가하는 것이 필요하다. 『茄』,『茨』,『苑』의 경우는 1형식으로 오분류되었다. 이 경우에는 『艸』부수의 특성을 고려한 분류가 이루어져야 할 것이다.

6. 결 론

본 논문에서는 KSC 5601 표준한자와 중, 고등학교 교육용 한자를 대상으로 하여 문자의 형식분류와 분류되어진 문자로 부터 부분패턴을 추출하는 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 추출 하였다.

표준 교육 한자 1800자를 대상으로 분류를 행한 결과 93.07%의 형식분류율과 89.5%의 부분패턴 추출율을 얻었다.