論文93-30B-7-3

자.모 해석적 모델에 의한 高精度 한글 인식 알고리즘에 관한 연구

- 패턴정합법에 기초한 후보문자 선정 및 구조해석적인 방법에 의한 유사문자 판별 -

(A Study on the Highly Accurate Korean Character Recognition Algorithm by Analyzing Vowel and Consonant Models)

- Selecting of candidates using pattern matching method and discriminating similar characters by structural analysis -

姜 仙 美*, 金 俸 奭**, 金 悳 鎭***

(Sun Mee Kang, Bong Seok Kim and Duck Jin Kim)

要 約

본 논문은 패턴정합방식을 기초로 한 한글문자 인식에서, 유사문자에 대한 정확한 인식을 위한 새로운 방법을 제안하였다. 제안된 방법은 기존의 한글문자 인식 방법들의 장점을 살린 것으로서, 안정된 후보문자의 확보와 유사문자들의 정확한 판별을 가능케 한다. 알고리즘의 성능 평가를 위해 15 종류의 명조체 레이저 프린터의 인쇄물을 사용하여 실험한 결과 97%의 인식율을 나타내었다.

Abstract

In this paper, a new method is proposed to recognize a character from its similar characters, which are selected by pattern matching method in Korean character recognition. This new method, which couples the merits of already suggested methods, can choose the character to be in the candidate set and discriminate it from the others correctly. To evaluate performance of this algorithm, we used 15 kinds of different laser printer fonts and obtained about 97% of recognition rate.

Ⅰ. 서론

사무 자동화의 급속한 발달과 더불어 문서의 자동 입력장치의 필요성은 이미 알려진바 있다. 영문 알파벳이나 숫자, 부호 등을 대상으로 하는 인식은 이미 인쇄체를 넘어서 필기체 문서에도 가능하다. 한자를 많이 사용하는 일본어에 대해서는 성능 이 아주 우수한 시제품[1,2]을 선보이기도 했다. 한글의 경우는 한자와는 달리 비교적 적은 획으로 이루워져 획 간 공간의 변화가 다양하여 여러 종류의 폰트 개발이 가능하다. 상용화된 명조체만 보더라도 신명조, 세명조, 태명조, 견출명조, 신문명조 등 다양한 모양과 크기의 폰트가 있으며 그 외에도 그래픽 에서 제공되는 명조체 범주에 속하는 폰트들도 사용되고 있다. 또한 한글은 유사한 문자들 간의 차이가 매우 미세하므로, 정확하고 안정된 인식율의 보장에 있어서는 아직도

*正會員,高麗大學校 情報通信 技術共同研究所
**準會員, ***正會員, 高麗大學校 電子工學科
(Dept. of Elec. Eng., Korea Univ)
(※본 논문은 삼성전자의 산학협동연구 과제 지원으로연구되었음.)
接受日字: 1992年 8月 10日

많은 문제점들이 남아있다. 이에 대하여 지금까지 발표된 여러 알고리즘[3,4,5,6]의 결과를 분석하여, 그 장점을 살려서 입력된 문자영상에 대하여 안정된 높은 인식율을 보장할 수 있는 한글 문자인식 알고리즘을 개발하고저 한다.

본 논문에서는 패턴정합방식을 기초로 하여 한글 문자인식을 행할 경우에 발생하는 문제점인 유사문자에 대한 정확한 인식을 위한 새로운 방법을 제안하였다. 기존의 한글 문자인식 방법에는 주로 한글의 초성, 중성, 종성의 조합적인 특성을 이용한 구조해석법적인 연구가 되어왔으나 레이저 프린터의 출력물 등에서 자주 발생되는 자소간의 붙음 현상이 두드러져서 인식율의 저하를 가져왔다. 이에 반하여 패턴정합법은 한글 인식에 있어서의 유사문자 판별이 매우 어렵다는 결과를 가져왔다. 그러므로 본 연구에서는 패턴정합법에 의한 안정된 후보문자들의 확보와 후보문자간의 유사성을 고려한 구조해석법적인 기법을 도입함으로써 정확한 인식율을 보장할 수 있는 방법을 제안하였다.

문자 획의 방향및 위치정보로 구성된 특징벡터를 이용하여 유사도 계산에 의한 후보문자 10위 까지를 고려한다면 99.9%이상의 인식율도 기대할 수 있다. 분류된 후보문자들을 분석하면 주로 작은 획의 차이가 있음을 알 수 있다. 이러한 차이를 감지할 수 있는 상세분류 과정을 삽입함으로써 최종 인식율을 향상 시킬 수 있다.

Ⅱ. 제안된 인식 알고리즘

본 연구에서 사용된 문자 패턴의 특징추출에 대하여는 참고문헌 [7] 에 제안된 방법을 기본으로 하여 구성되었으며, 얻어진 특징벡터의 거리계산에 의한 분류 실험결과에서 나타난 후보문자들에 대한 구조해석적인 상세분류 과정을 삽입시켰다. 제안된 인식 알고리즘은 입력된 문서에서 개별문자를 추출 한후, 크기에 대한 정규화 작업(32×32 화소)을 수행한다. 정규화된 문자에 대하여 문자 윤곽선소의 방향 및 위치정보를 이용한 특징벡터를 구성한다. 표준특징벡터와의 거리계산에 의한 후보문자를 선정한 후에 유사문자 사전과 자.모 해석적 모델을 이용한 상세분류 과정을 수행한다.

본 연구에서는 문자 윤곽선소의 방향정보를 특징소로 사용하였다. 한 화소로 연결된 획의 경우와 문자 윤곽선부의 잡음을 고려하여 특징추출 템플리트를 제안하였다. 각 화소는 이웃 화소와의 관계에서 제안된 템플리트에 의해 4 방향정보를 얻게되며, 이에 각 화소의 위치정보를 포함하여 특징벡터를 구성한다. 특징벡터의 구성은 9진 트리의 계층적 구조로 이루어졌으며, 9진 트리의 각 노드는 9개 부노드의 각 방향별 특징소의 합으로 이루어졌다. 한글의 특성인 모음 및 자음의 미세한 변화에 의한 유사문자의 高精度 인식을 위하여 미리 작성된 유사문자 사전을 이용하여, 유사한 후보문자들에 투영 도법과 런랭스(run length)[4,8] 기법등을 적용하여 최종 인식문자를 결정하였다.

실험에 사용된 한글 영상 데이타는 본 연구실에서 만든 한글 데이타 베이스(DB 1-15)로 각 DB는 KS 완성형 2350자의 문자영상을 포함하고 있으며, 문자영상은 15 종류의 명조체 레이저 프린터의 인쇄물에 해당된다. 15개의 DB 중에 DB 1-10은 표준 특징벡터 사전을 만드는데 사용되었으며, DB 11-15는 시험용 문자 세트로 이용하였다.

1. 특징 추출부

본 연구에서 사용한 특징벡터의 구성에 있어서 특징소 추출 방법과 특징벡터 구성 방법에 관한 대략적인 방법은 참고문헌 [7] 에 자세히 소개된바 있으며, 수정된 사항에 대한 소개로 간략화 하였다. 본 연구에서는 일반적으로 세선화 템플리트가 방향성에 대응되는 점에 착안하여, 템플리트에 의하여 추출되어 버려지는 화소가 가지고 있는 정보에 중점을 두어 문자인식을 위한 특징소로 이용하였다. 추출된 특징소는 그 존재 위치에 따라서 9개의 중첩된 소영역으로 분할하여 각기 서로 다른 위치별 가중치를 갖고 36차원 혹은 324차원의 특징벡터로 구성되어 인식실험에 사용된다. 특징벡터의 구성은 인식 과정에서 필요한 精度에 따라서 사용될 수 있도록 계층구조로 이루어져 있으며, 각 영역별 가중치를 살펴보면 영역의 가운데 부분이 강조되어 문자의 변위에 강하도록 하였다.

1) 제안된 특징소 추출 방법

특징추출을 위해 문자 윤곽선부에 잡음이 있는 경우를 고려하여 그림 1과 같은 12개의 특징소 추출 템플리트들을 사용하였다. 1 로 표시된 부분은 8개 주변 화소와의 관계에서, 조건에 맞는 경우 해당되는 방향성분을 갖고 추출된다. 방향성분은 4종류로 수평, 수직, 사선과 역사선 방향을 가리킨다. 그림 1의 (a)-(h)는 문자의 윤곽선부의 방향 성분을 추출할 때 사용되며, 특히 수평과 수직성분을 추출하기 위해 사용되는 템플리트들은 윤곽선 부분의 잡음에 의한 영향을 최소화하기 위해 윤곽선과 인접한 내부의 안정된 방향성분을 이용하였다. 또한 (i)-(l)은 한 화소로 연결된 획의 방향성분을 추출할 때 사용된다.

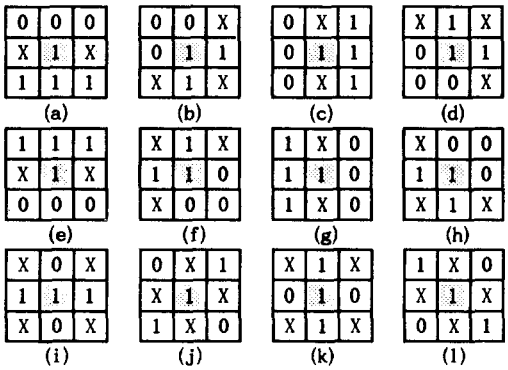
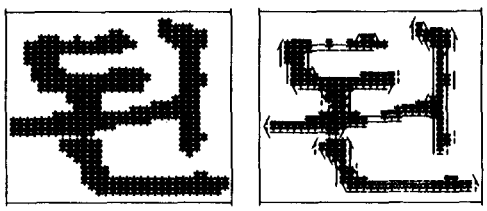


그림 1. 제안된 특징소 추출 템플리트 (X: don't care)

Fig. 1. Proposed templates for extracting feature primitives.

그림 2는 제안된 특징소 추출 템플리트에 의하여 벗겨진 특징소의 예로서 '|','/','-','\' 등은 각각 4가지 방향성분을 나타내고 있다.



(a) (b)

그림 2. 제안된 템플리트에 의해 추출된 특징소
(a) 정규화된 입력 영상
(b) 특징소의 추출

Fig. 2. Feature primitives extracted by proposed templates.
(a) Normalized input image,
(b) Extracted feature primitives.

2) 특징벡터 구성

본 논문에서 제안하고 있는 9진트리를 이용한 계층적 특징벡터 구성방법은 정규화된 입력문자(32×32)에 대하여 가로 및 세로 방향으로 1/4 길이(32/4=8)만큼 중첩하여 16×16 크기의 9개의 소영역을 만든다. 각 소영역을 계속하여 분할이 가능할 때까지 같은 방법으로 분할하면, 최종적으로 2×2 크기의 소영역을 얻게된다. 분할된 영상에서 얻어진 특징소들은 각 소영역의 크기별로 서로 다른 차원의 특징벡터를 얻을 수 있는데, 본 연구에서는 정규화된 문자 크기의 1/4에 해당되는 8×8의 영역에서 얻어지는 특징벡터 324차원(4방향×9^2의 소영역)를 이용하였다.

2. 특징벡터의 분류실험

구성된 한글 문자세트 중 표준문자 10개 세트(DB 1-10)에서 추출한 특징벡터의 평균값과 시험문자세트(DB 11-15)의 각 문자에서 추출된 특징벡터와의 거리를 비교하는 분류실험을 통하여, 본 연구에서 제안하고 있는 특징소 및 특징벡터의 구성체계가 유효함을 검증하였다. 분류실험에 이용된 거리 계산방법으로는 특징벡터를 N차원의 벡터로 간주하고 표준문자 데이타와 시험문자 데이타간의 각차원에서의 특징벡터 거리차의 제곱에 비례하는 Euclidean거리 계산법에 대해 적용해 보았다. 분류실험 결과는 그림 3과 같으며, 후보문자의 순위(Order)에 따른 누적분류율을 나타내었다.

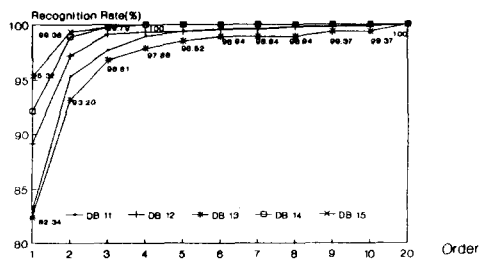


그림 3. 분류 실험 결과 (324차원)

Fig. 3. Results of classification experiments.

그림 3의 분류실험 결과를 살펴보면 10위 이내로 분류될 확률은 99.9%를 나타낸다. 그러므로, 본 연구에서 제안한 특징추출 방법에 의한 후보문자 확보는 안정되게 이루어졌음을 알 수 있다. 추출된 특징벡터를 적용하여 그 중에 입력문자와 가장 유사한 문자들을 인식 후보문자로 상세분류 과정으로 넘겨준다.

후보문자 3위까지를 고려한 누적분류율은 98.93%를, 후보문자 5위까지를 고려한 경우 99.80%이상의 분류율을 나타내었다. 즉, 대분류 과정을 통하여 선정된 후보문자 3개만을 고려할 경우에도, 98.93%의 인식율을 얻을 수 있다. 후보 순위에 따른 대분류 후의 결과를 표 1에 나타내었다. 선정된 후보문자들을

살펴보면 매우 유사성이 많은 것을 알 수 있으며, 이러한 유사한 문자들에 대한 분명한 판별을 할 수 있는 상세분류 과정이 요구된다. 후보문자의 선정에 있어서는 분류실험 결과와 표 1에 나타난 후보문자들의 유사성을 토대로 안정되게 이루어졌음을 확인할 수 있다.

표　1. 대분류 후 얻어진 후보 문자의 예
Table 1. Example of candidate characters after rough classification.

KS코드	입력문자	순위별 후보 문자				
b3af	날	날	낼	낳	달	낟
b5c1	들	들	둘	틀	돋	놀
b9ca	믐	믐	뭄	뭅	늠	놈
bbc0	뻐	며	뻐	뻐	띠	떠
bccf	센	션	쎈	선	신	센
c4e0	쿳	꿋	풋	풋	굿	쿳

3. 유사문자 사전을 이용한 상세분류

본 절에는 대분류 상에 나타난 후보문자들 중에서 인식문자를 정확하게 판별하기 위한 상세분류 과정에 관하여 기술하였다. 표 1에서 알 수 있듯이 대분류 상에 나타난 후보문자들은 문자세트에 따라서 다소 차이는 있으나, 대개 유사문자 간의 오류가 적잖게 일어나고 있다. 2위 혹은 3위에서 인식 되어지는 문자들을 분석해보면 일반적으로 한 획 혹은 짧은 두획 정도의 차이를 나타내고 있다. 간혹 잡음이 심한 경우나 전처리상의 처리과정에서 두꺼운 획들이 서로 붙어서 구분이 어렵게 되는 경우, 즉 'ㅃ' 혹은 'ㄸ'과 'ㅍ' 등은 전혀 다른 형태의 특징벡터로 추출되는 예도 있다. 또한 '웃'과 '읏', '웅'과 '응', '움'과 '쓤', '쎈'과 '쎈'의 경우에는 프린터 출력의 상태에 따라서 문맥을 읽어 보지 않고는 구별이 안되는 경우가 종종 발생한다. 그러나 대부분의 경우에 있어서, 한글의 특성인 모음의 변화, 즉 'ㅐ'와 'ㅒ' 혹은 'ㅏ'와 'ㅑ' 등과 자음에 있어서의 작은 변화, 즉 'ㅈ'과 'ㅊ' 혹은 'ㅂ'과 'ㅁ' 등에서 오 인식이 자주 일어난다. 그러므로 이러한 오 인식이 자주 일어나는 부분만을 미리 구별하여 필요한 부분에 대한 구조해석법을 적용한다면, 패턴정합방식에 있어서의 유사문자간의 오 인식을 해결해 낼 수 있게 된다.

이와 같이 구조정보를 이용한 모델을 선정하여 적용해 본 결과, 유사도 계산 만으로는 구별이 어려운 유사문자들을 분명하게 구별할 수 있었다. 제안된 상세분류 과정은 그림 4와 같다.

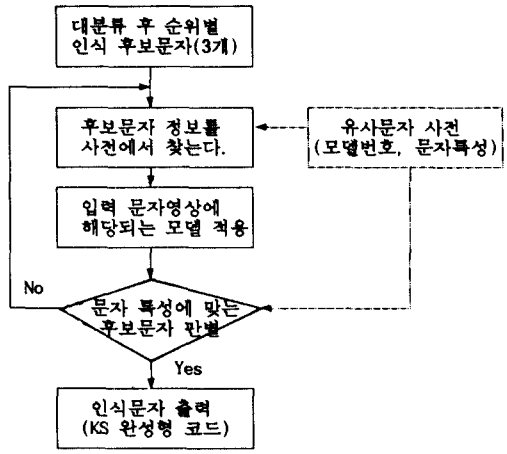


그림 4. 상세분류 과정
Fig. 4. The procedure of discrimination.

1) 유사문자 사전 작성

대분류 상에서 오 인식 가능한 유사문자들의 정보를 얻기 위하여, 먼저 분류 실험 결과를 이용하여 각 문자에 대한 유사한 문자들을 모아서 유사문자 사전의 후보문자들로 기록한다.[9] 후보문자들의 구조적 분석을 통하여 적용할 모델을 선정하고, 후보문자들에 대하여 모델 적용시 얻을 수 있는 여러가지 특징들을 수식화하여 유사문자 사전에 각 문자의 특징으로 기록한다. 이렇게 학습 문자세트를 이용하여 미리 사전을 작성한 후에 상세분류 실험의 결과를 이용하여 사전의 적응력을 향상시킨다.

이상과 같은 방법으로 작성된 유사문자 사전에는 각 문자마다 해당되는 여러 종류의 모델이 필요하다. 사전의 유효성을 검증하기 위해서 시험 DB를 이용하여 후보문자 3위까지에 해당되는 모델을 적용할 경우, 대분류 상에 나타난 오인식된 문자들의 인식이 가능함을 알 수 있었다. 유사문자 사전에 나타난 후보 문자들을 중심으로 만들어진 구조해석법적인 모델들은 일반 텍스트 실험에서 자주 발생하는 오인식 문자들을 중점적으로 다루었다. 이러한 과정은 유사문자가 많은 한글 문자 인식에서 高精度 인식을 위해서 필수적이다. 사전의 적응성을 개선하기 위해서는 많은 대상의 한글 문자세트에 대한 학습 효과가 필요하므로, 실험용으로 쓰이는 문자세트에 대해서도 사전의 부분적 수정 작업이 가능하도록 하였다.

2) 구조 해석을 위한 모델 설정

일반적인 한글인식에서 사용되는 구조해석법은 형

식분류를 기초로 해서 자소 해석을 하는 방식을 취하고 있다. 자소 해석에 있어서도 여러가지 방법들이 이용되나, 본 논문에서 제안한 방법은 투영도법과 획의 길이정보(run-length)를 사용한 비교적 간단한 방법으로 패턴정합법으로는 구별이 어려운 유사문자의 구별이 가능하도록 하였다.

유사문자 사전에 나타난 후보문자들은 특정한 부분에 대한 투영도법을 적용해 볼 때, 매우 간단하게 구별이 될 수 있음을 알 수 있다(받침의 유무에 따른 모음의 변화: ㅡ, ㅗ, ㅛ, ㅜ, ㅠ, ㅣ, ㅏ, ㅑ, ㅓ, ㅕ, ㅐ, ㅒ등). 또한 방향별로 처음 나타나는 획과 다음에 나타나는 획과의 상관 관계를 조사해 봄으로서 구별이 가능한 경우도 있다(자음의 변화; ㅈ, ㅊ와 ㄱ, ㅋ등). 이 외에도 비슷한 방법의 모델들이 있으며, 이는 같은 자소라도 구성에 따라서 그 모양이 변화되는 한글의 특성을 반영하여 각각의 경우에 따라서 약간의 변형을 가한 것이다. 현재까지 사용되는 모델은 15개이며 표 2에 사용된 구조 해석적 모델에 대하여 간단한 예와 더불어 소개되었다. 모델 적용 후에도 구별이 결정적이지 못한 경우는 대분류 상에 인식 후보문자 1위에 나타난 문자를 인식문자로 결정하였다. 실험용으로 사용된 일반 문서 1만자에 대하여 대분류 결과에 나타난 오 인식 문자들을 조사하여 해당되는 모델과 각 모델별 오 인식 빈돗수를 분석한 결과를 살펴보면 받침이 있는 긴 수평 모음의 경우는 전체 오 인식된 문자들의 35% 정도 차지하며, 받침이 없는 긴 수직 모음의 경우도 15% 에 달하고 있다.

표 2. 구조해석적 모델들의 예

Table 2. Example of structure analysis models.

모델	판별 자.모음	판별 방법	예	오인식 빈돗수
1	받침있는 긴 수평 모음: ㅡㅗㅛ ㅜㅠ	수평모음 중간 부분에서의 아래, 위 방향 흑화소 연속수	을, 울	35%
3	받침없는 긴 수직 모음: ㅏ ㅑ	수직모음 오른쪽에서의 획의 갯수	라, 랴	15%
5	받침없는 긴 수직 모음: ㅣ, ㅓ, ㅕ	수직모음 왼쪽에서의 획의 갯수	너, 녀	9%
9	초성: ㅈ, ㄱ 과 ㅊ, ㅋ	초성 상단점 아래에서 오른쪽 방향 흑화소 연속수가 큰 획의 유무	저, 처 강, 캉	5%

그 외에도 'ㅊ'과 'ㅈ'을 구별짓는 모델의 경우도 높은 빈돗수를 나타내고 있다. 이는 대개 자주 사용되는 모델들의 적용이 가능한 문자들의 유형은 패턴정합법에서는 구별이 매우 어려우며, 또한 일반 문서에서 자주 사용되는 문자들이기 때문이다. 실험에서 모델 종류별로 나타난 오 인식 빈돗수를 표 2에 나타내었다. 제시된 4개의 모델이 비교적 자주 발생하는 오 인식의 예이며 전체의 64%에 달함을 알 수 있다.

3) 상세분류 결과 분석

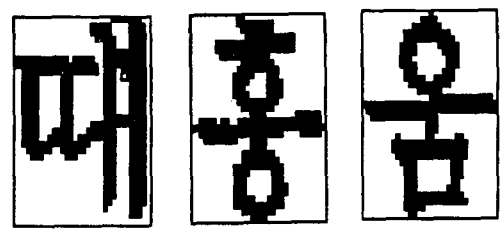
시험문자 세트로 대분류에서 후보 3위이내로 분류된 문자들에 유사문자 사전을 적용해 본 결과는 표 3에서 알 수 있듯이 문자세트에 따라 다소 차이는 있으나 95%정도의 인식율을 보이고 있다.

표 3. 상세분류 전.후의 인식 실험결과 비교 (인식대상문자: 2350자)

Table 3. Comparison of recognition rate before and after discrimination.

시험 문자 세트	상세분류후 인식률 (%)	상세분류전의 분류율 1위	2위	3위
DB 11	93.2	82.8	95.4	97.7
DB 12	95.2	89.2	97.3	99.2
DB 13	92.9	82.4	93.2	97.0
DB 14	96.1	92.2	98.9	99.8
DB 15	95.7	95.3	99.4	99.8

대분류에서 좀 더 많은 후보문자들에 대한 유사문자 사전을 적용한다면 더 높은 인식율도 기대할 수 있다. 그러나 여러 모델들을 적용할 경우 더 많은 처리시간이 요구되며, 잘못된 모델의 적용도 가능하므로 적절한 수의 후보문자를 고려해야한다.



(a) 때 (b) 홍 (c) 움

그림 5. 상세분류후의 오 인식 예
(a) 지면 상의 잡음 (b) 자음의 변형
(c) 자모가 붙은 경우

Fig. 5. Examples of misrecognition after discrimination.
(a)Noise in paper, (b)Modification in consonant, (c)Attached strokes.

상세분류에서 오 인식된 문자들을 살펴보면, 문맥에 대한 이해 없이 입력된 문자영상에서 추출된 특징만으로는 인식의 한계가 있음을 알 수 있다. 현재 오 인식된 후보들 중에는 문자체의 변화에 기인한 것도 있으나, 이에 대한 수정 작업이 가능하다 해도 그림 5에 나타난것과 같은 오 인식은 문맥을 어느정도 파악 할 수 있는 후처리 과정이 필요함을 시사한다. 후처리를 위해서는 한글 단어사전 및 문법과 조사의 사용법 등의 다양한 연구가 선행되어져야 하며, 이러한 작업이 완료되기까지는 국문학과의 지원이 필요하다고 본다. 연구된 자료에 대하여도 다양한 문법구조에 알맞는 구문 구조해석법적인 연구가 진행되어져야 한다. 많은 양의 정보를 포함하고 있는 한글사전을 탐색하여 판별하는 과정의 삽입은 막대한 처리시간을 요하게 된다. 본 연구에서는 후처리 과정은 생략하였으나, 계속적으로 연구되어져야 할 것으로 사료된다.

Ⅲ. 인식 실험 및 검토

1. 실험용 한글 데이타 베이스 구성

한글 인식 알고리즘을 개발하고 연구 평가하기 위해서는 한글 영상 데이타 베이스(DB)가 필요하다. 그러나 아직 국내에는 표준화된 한글 문자영상 DB가 정립되지 않았다. 본 연구실에서는 인식 실험용 문자세트로 한글 명조체 KS완성형 2350자에 대하여 시판되고 있는 5종류의 레이저 프린터를 이용하여 각 3종류씩 크기(40×40화소 - 56×56화소)별로 출력시킨 15개의 문자세트를 구축하였다. 삼성(명조), 삼보(명조)와 H.P(¡ 글) 레이저 프린터에서는 Bitmap 형식의 프린트 인쇄물이며, 맥킨토시(문조, 중명조)에서는 PostScript형식을 사용하였다. DB 1-10 까지의 10개를 표준 특징벡터를 구성하는데 이용하였고, 나머지 DB 11-15의 5개는 시험 문자세트로 이용하였다. 한글 데이타 베이스 구축에 관한 자세한 내용은 참고문헌 [10] 에 기술되어있다.

2. 실험용 문자세트에 의한 실험

구성된 한글 문자세트 중 표준데이타 세트에서 추출한 특징벡터와 시험 문자세트의 각 문자에서 추출된 특징벡터와의 인식실험을 통하여, 본 연구에서 제안하고 있는 인식 알고리즘의 성능을 평가해보았다. 인식대상 문자는 한글 KS-완성형 2350자와, 빈도수 순으로 우선 순위가 높은 522자와 1500자에 대하여도 각각 실험해 보았다.

그림 6의 인식대상 문자수에 따른 실험결과를 살펴보면 빈돗수가 낮은 문자들을 제외한 1500자나 522자를 대상으로 실험한 결과는 97% 이상의 인식결과를 얻을 수 있다. 사용 빈돗수는 낮으나 글꼴이 비교적 복잡한 의성어, 의태어 또는 자모의 유사문자가 많은 외래어 표기 등에서 오인식이 자주 일어남을 알 수 있었다.

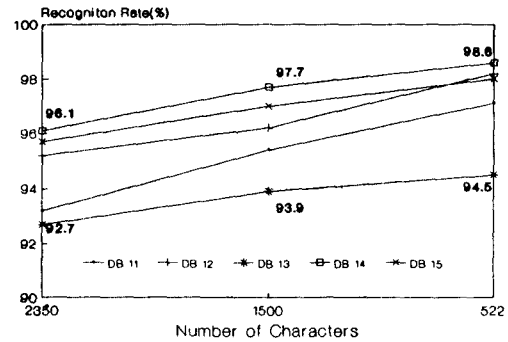


그림 6. 인식 대상문자 수에 따른 인식실험 결과 비교

Fig. 6. Recognition result according to number of korean character.

Ⅳ 결 론

본 연구에서 제안한 문자 인식알고리듬은 패턴 정합방식에 기초하여 유사문자 간의 구조적 분석을 병행한 방식으로, 한글 문자인식에 있어서 高精度 인식이 가능한 새로운 방법이다.

정규화된 입력 문자영상에 잡음에 의한 문자 윤곽선부에서 추출되는 방향 정보의 산란을 막을 수 있는 특징추출 템플리트와 추출된 특징소의 위치에 따라서 문자변위를 고려한 중첩된 소영역의 분할에 의한 특징벡터를 구성하였다. 이와 같은 특징벡터 구성의 유효성을 검증하기 위하여 간단한 거리계산에 의한 분류실험 결과 324차원을 사용했을때 후보문자 10위 이내에 99%이상의 분류율을 나타내었다. 분류실험 결과는 문자의 특징추출이 정확하게 이루어졌음을 입증한다. 대분류에서 분류되는 문자들에서 나타나는 유사문자간의 오 인식을 정확하게 판별하기 위하여, 한글 자모의 구조적 특징을 사용하였다. 통계적 방법에서는 구별해내기 어려운 문자들을 모아서 유사문자 사전을 작성하고, 사전에 나타나는 유사문자들의 특성을 분석하는 15개의 모델을 만들어서 실험해 본 결과, 전체적으로 대분류 후 결과에 비하여 5%정도 인식율의 향상을 가져왔다. 시험문자세트에 대한 인식

실험에서 KS 완성형 2350자에 대하여 95%이상의 분류율을 보이고 있으며, 사용 빈도수가 높은 1500자에 대한 실험결과는 97%정도이다. 일반 문서 1만자에 대하여 실험해 본 결과(인식대상문자 900자) 94.7%의 인식률을 나타냈다. 이러한 결과는 패턴 정합 방식의 특징인 학습된 문자체에 대하여는 적응력이 매우 높은 반면에, 학습되지 않은 문자체에 대하여는 다소 떨어지는 현상을 나타낸다. 그러므로 여러 문자체에 대한 학습이 필수적이며, 이를 위해 여러 종류의 문체에 대한 계속적인 재 학습을 위한 기능도 필요하다.

參 考 文 獻

[1] 酒巻 久 et. al., "認識速度 70字/秒の日本語 OCR,專用 LSIのDSPのペイプライン處理 高速化", キヤノンソフトウエア戰略本部, NIKKEI ELECTRONICS, pp.195-201, 1990. 7.

[2] 弘具, 大町眞一郎, 木村 正行, 勝山 裕, "高速高精度知的 認識システム SEIUN", 電子情報通信學會論文誌 Vol.J76-D- II No.3 pp. 474-484 1993. 3.

[3] 이 주근, "한글文字의 認識에 관한 硏究(Ⅳ)", 전자공학회지, 제 9권, 제 4호, pp.25-32, 1972. 9.

[4] 도 정인 외, "인쇄체 한글 문자의 인식을 위한 자소분리에 관한 연구", 한국정보과학회 가을 학술발표 논문집, Vol.17, No.2, pp.175-178, 1990.

[5] 김 진형 외, "문서 인식 및 처리기의 개발에 관한 연구", 연구보고서, 한국 과학 기술원, 1989. 4.

[6] 이승호 외, "한글 문서 인식 시스템 SIL-NOON", 한글 및 한국어 정보처리 학술발표 논문지, pp.132-136, 1989.

[7] 강선미 외, "윤곽선소의 방향정보를 이용한 특징소 추출 및 특징량 구성에 관한 연구", 전자공학회 추계종합학술대회, 14권, 2호, pp. 213-215, 1991. 11.

[8] Michio Umeda, "Recognition of Multi-font Printed Chinese Characters", CH1801-0/82/0000/0793 $00.75 1982 IEEE, pp.793-796, 1982.

[9] 孫 寧, 田原 秀, 阿曹 弘具, 木村 正行, "方向線素特徵量を用いた高精度文字認識", 電子情報通信學會論文誌 Vol.J74-D-II No.3 pp. 330-339,1991.3.

[10] 김 덕진 외, "병렬처리 기술을 이용한 인쇄문자 인식기의 구현에 관한연구", 연구보고서, 고려대학교 정보통신기술 공동연구소, 1992. 6.

著 者 紹 介

姜 仙 美(正會員) 第29卷 B編 第11號 參照
현재 고려대학교부설 정보.통신기술 공동연구소 연구조교수

金 惠 鎭(正會員) 第29卷 A編 第8號 參照
현재 고려대학교 전자공학과 교수

金 俸 奭(準會員)
1967年 10月 17日生. 1990年 8月 고려대학교 전자공학과 졸업. 1992年 8月 고려대학교 대학원 전자공학과 졸업. 주관심 분야는 영상 처리 및 패턴 인식 등임. 현재 삼성전자 근무.