6-유형 별로 적응적 계층 구조를 갖는 인쇄 한글 인식

Printed Hangul Recognition with Adaptive Hierarchical Strutures Depending on 6-Types

함대성, 이득용, 최경웅, 오일석
전북대학교 전자정보공학부 컴퓨터공학

Dae-Sung Ham(ham890@chonbuk.ac.kr), Duk-Ryong Lee(dr_lee@chonbuk.ac.kr),
Kyung-Ung Choi(pinksmup@ourtech.co.kr), Il-Seok Oh(isoh@chonbuk.ac.kr)

요약

한글 인식은 부류 수가 많다는 특성을 가지며 이 특성으로 인해 6-유형으로 사전 분류하는 것이 일반적이다. 사전 분류 후 각 유형들은 초성, 중성, 종성으로 분리하여 인식할 수 있다. 초성, 중성, 종성 각각은 부류의 수는 적지만 'ㅐ','ㅒ', 과 같이 서로간의 유사도가 높아 오 인식 되는 경우가 종종 발생한다. 따라서 본 논문에서는 6-유형 각각에 대해 다단계 트리 구조를 가진 계층적 인식 방법을 제안 하였다. 또한 초성, 중성, 종성의 서로 간의 간섭을 줄이기 위해, 초성과 종성의 인식 결과를 중성 분류기의 특징으로 사용하였다. PHD08 데이터베이스의 테스트 집합에 대해 98.96%의 정확률을 보였다.

■ 중심어 : | 한글 인식, | 신경망 | 특징 추출 | 문서 영상 | 웹 영상 |

Abstract

Due to a large number of classes in Hangul character recognition, it is usual to use the six-type preclassification stage. After the preclassification, the first consonent, vowel, and last consonent can be classified separately. Though each of three components has a few of classes, classification errors occurs often due to shape similarity such as 'ㅐ' and 'ㅒ'. So this paper proposes a hierarchical recognition method which adopts multi-stage tree structures for each of 6-types. In addition, to reduce the interference among three components, the method uses the recognition results of first consonents and vowel as features of vowel classifier. The recogntion accuracy for the test set of PHD08 database was 98.96%.

■ keyword : | Hangul Recognition | Neural Network | Feature Extraction | Duocument Images | Web Images |

I. 서 론

예전 문서를 스캔한 형식 문서 영상이나 웹에 게시되어 있는 웹 문서는 중요한 콘텐츠이다[1]. 이들에 포함되어 있는 문자는 이미지 형태로서 코드화되어 있지 않기 때문에 콘텐츠를 색인하는 등의 작업이 불가능하다. 이들 콘텐츠가 이러한 상태로 남아 있는 한 생명력을 가진 콘텐츠라 볼 수가 없고 활용도는 매우 낮을 수밖에 없다. 따라서 이들 영상에서 문자를 추출하고 인식하는 작업은 매우 중요하다 할 수 있다. 이들은 양이 매

* 본 연구는 교육과학기술부와 한국산업기술진흥원의 지역혁신인력양성사업으로 수행된 연구결과임

접수번호 : #091008-010
접수일자 : 2009년 10월 08일
심사완료일 : 2010년 01월 07일
교신저자 : 오일석, e-mail : isoh@chonbuk.ac.kr

우 방대하고 보다 복잡한 형태가 많이 발생하고 있다.

하지만 기존의 한글 인식은 대부분 형식 문서를 대상으로 하였다. 이들 응용에서는 일정한 형식을 가진 영상에서 단락, 줄, 단어, 글자 분할을 수행한 후 인식을 하는 방식을 사용한다. 하지만 현재의 응용은 형식화된 문서의 인식에 국한되지 않는다. 간판인식, 웹 영상의 문자인식, 우편물의 주소인식, 데이터 마이닝, 그리고 전자 도서관 등으로 응용분야가 확대되었다. 이러한 응용에서의 문자 인식은 훨씬 어렵다. 따라서 인쇄 한글 인식은 여전히 중요하고 미 해결인 문제로 보아야 한다.

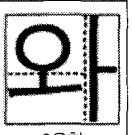
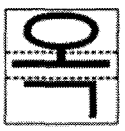
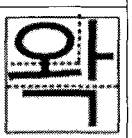
1. 관련 연구

한글 인식이 어려운 이유 중 하나는 부류의 수가 매우 많다는데 기인한다. 한글 문자의 조합에 관한 연구[2]에 따르면, 자음과 모음의 조합에 의해 생성 될 수 있는 문자는 14,364자에 달한다. 이것은 매우 방대한 숫자로 한글 인식의 어려움을 단편적으로 보여 준다. 하지만 디지털로 표현 가능한 한글의 수는 한국 산업 표준[3] KS X 1001에 따르면 2350글자로 국한된다. 즉 문자 인식 프로그램이 인식해야하는 문자의 수가 2350자로 제한된 것이다. 하지만 이것 또한 적은 수가 아니며, 하나의 인식기로 인식하는 것은 매우 어려운 문제이다.

이러한 문제점을 극복하기 위해 많은 연구들이 이루어 졌다. 이들 연구들의 공통점은 한글이 자음과 모음의 조합으로 만들어지며, 조합 규칙을 가지고 있다는데 기인한다. 한글은 초성 19개, 중성 21개, 종성 28개의 조합으로 구성되며, 각 문자들은 초성+중성 또는 초성+중성+종성이 결합되어 만들어진다. 이러한 특징은 표 1과같이 한글을 6가지 유형으로 구분할 수 있도록 해준다. 각 유형들은 중성의 위치와 종성의 유무에 따라 구분되어진다. 이러한 유형 분류는 각 유형을 독립적으로 인식할 수 있도록 해주며, 이것은 한 번에 인식해야 하는 대상의 수를 줄일 수 있도록 해준다. 이러한 접근 방법은 매우 타당하며, 많은 연구들이 이를 뒷받침 하고 있다.

6 유형으로 분류하는 방법은 크게 규칙 기반 분류 방법과 인식기를 이용한 방법으로 나눌 수 있다. 규칙 기반 분류 방법으로는 허프 변환을 이용한 방법[4], 모음의 구조적 형태에 기반한 트리 분류 방법[5-7], 비선형 분할 경로를 이용한 방법[8], 신경망을 이용한 방법[9-12] 등이 있다. 허프 변환을 이용한 방법과 트리 분류 방법, 비선형 분할 경로를 이용한 방법들은 공통적으로 모음의 위치와 형태를 이용해 유형을 분류한다. 이때 자음과 모음의 간섭이 심할 경우 모음의 위치를 찾지 못해 오분류 하는 경우가 발생한다. 또한 다양한 폰트에는 적용 할 수 없다는 단점을 가지고 있다. 이에 반해 신경망을 이용한 방법은 문자 전체를 입력으로 사용하여 학습 시키거나[9], 각 유형별로 특정 위치에서 추출된 특징을 입력으로 사용하여 유형을 분류한다[10-12]. 신경망 방법은 비교적 안정적인 성능을 보인다.

표 1. 한글의 6 유형

	종모음	횡모음	종모음 횡모음
받침 없음	 1 유형	 2유형	 3유형
받침 있음	 4유형	 5유형	 6유형

(a) 자소의 획 분할 (b) 자소 영역

그림 1. 인쇄 한글 인식 방법

6유형으로 분류 한 후 각 유형별로 인식하는 과정은 자소 분할 여부에 따라 두 가지로 나눌 수 있다. [그림 1](a)와 같이 자음과 모음을 완전히 분리하여 자소를 독립된 형태로 인식하는 것과, [그림 1](b)와 같이 자소의 영역의 위치를 지정하고 인식하는 것이다.

자소를 독립적으로 인식하는 방법은 자소 분리 후 인식을 위해 초성 19개, 중성 21개, 종성 28개를 위한 분류

기만 설계하면 된다. 이 방법은 자소 분리 성능이 인식 성능에 매우 큰 영향을 미친다. 자소를 분리하기 위해 모음의 구조적 형태를 이용한 방법[5][6], 신경망을 이용한 방법[9], 허프 변환[4]을 이용한 방법 등이 연구 되었으나, 이것은 매우 어려운 문제이며 아직도 많은 연구가 필요한 실정이다.

두 번째 방법은 이러한 자소 분리 문제를 우회하는 방법으로 [그림 1](b)와 같이 각 자소가 나타날 수 있는 영역을 지정한 후 인식하는 방법이 있다. 이 방법은 글자 전체를 신경망의 입력으로 사용하는 방법[5][7]과 초성, 중성, 종성으로 분리하여 독립적으로 인식하는 방법[9] 으로 나눌 수 있다. 글자 전체를 입력으로 사용하는 방법은 6유형 분류 후에도 많은 부류를 가지는 유형이 존재하기 때문에 단일 인식기로 인식하기는 여전히 어렵다. [그림 2]와 같이 4유형과 5유형의 경우에는 1069개와 585개의 부류를 가진다. 초성 중성 종성으로 분리하여 인식하는 방법은 자소를 독립적으로 인식하는 방법의 단점인 자소 분리 문제를 우회하며, 부류의 수 또한 적어 획 분할 방법과 자소 영역 분할 방법의 장점을 모두 가진다. 하지만 중성 경우 21부류로 인식 대상의 개수는 많지 않으나, 각 모음들끼리 유사도가 매우 높아 높은 성능을 기대하기 어렵다. 이것은 종성도 다르지 않다.

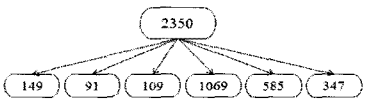


그림 2. 6-유형 분류에 의한 문제 축소

2. 제안하는 방법

본 논문에서는 6유형으로 분류한 후 자소 영역의 위치를 지정하고 초성, 중성, 종성을 독립적으로 인식하는 방법을 사용 한다. 이때 중성과 종성이 부류 내 유사도가 높아 인식기의 성능이 저하되는 것을 방지하기 위해, 중성과 종성을 세부 유형으로 분리하여 인식하는 아이디어를 제시한다. 예를 들어, 모음의 경우 세부 유형 분류는 모음의 긴 획의 모양과 삐침의 유무에 따라 분류된다. 즉 "ㅣ", "ㅓ", "ㅏ"는 삐침의 방향에 따라 세 부류로 구분 된다. 이와 같이 계층적으로 한글 인식기를 만들면, 하나의 인식기가 인식해야 하는 대상의 수가 매우 줄어들어 효과적인 인식을 할 수 있다. 여러 유형 별 인식기 중 4유형의 인식 트리는 다음과 같이 동작한다. 이것은 한 예로서 다른 유형의 다른 자소는 그에 적합한 계층 분류 구조를 가진다.

1) 6 유형 분류기가 4 유형으로 분류하면 (6-부류 분류)
2) {ㅏ, ㅑ, ㅣ-형}인지 분류하고 (3-부류 분류), ㅏ또는 ㅑ로 분류되면 거기서 멈추지만 ㅣ-형으로 분류되면 3)으로 간다.
3) {{ㅓ,ㅕ,ㅣ},{ㅐ,ㅒ,ㅔ,ㅖ}}의 두 부류 중 하나로 분류한다.
4) {ㅓ,ㅕ,ㅣ}이면 3-부류 분류를 하고 {ㅐ,ㅒ,ㅔ,ㅖ}이면 4-부류 분류를 한다.

위와 같은 계층 구조를 가진다 하더라도, 자소 영역의 위치를 정하고 인식기의 입력으로 사용하기 때문에 자소간의 간섭을 피할 수 없다. 본 논문에서는 자소간의 간섭을 최소화하기 위해, 초성과 종성의 인식결과를 모음 인식기의 입력으로 사용하는 아이디어를 제시한다. 즉 모음 인식 시 초성과 종성의 영향을 최소화한 것이다.

본 논문의 2장에서는 제안하는 계층적 한글 인식기의 구조를 설명하고, 3장에서는 인식기의 성능을 제시한다. 또한 4장에서는 웹 영상에서 추출한 문자 영상에 대한 성능을 제시한다.

II. 계층적 한글 인식기

[그림 2]는 제안한 계층적 인식기를 보여 준다. 입력된 샘플을 우선 6-유형 분류기로 6 유형 중의 하나로 분류한다. 6 유형 각각이 유형별 분류기를 갖는데, 이 분류기는 해당 유형의 특성에 따라 설계되어 있다. 예를 들어, [그림 3]이 보여주는 바와 같이 유형 4는 초성 분류기, 중성 분류기, 그리고 종성 분류기로 구성된다.

중성 분류기의 경우 먼저 'ㅏ', 'ㅑ', 그리고 나머지의 세 부류로 분류하며, 나머지는 {ㅓ,ㅕ,ㅣ}와 {ㅐ,ㅒ,ㅔ,ㅖ}의 두 부류로 분류한다. 이후 {ㅓ,ㅕ,ㅣ}는 세 부류로, {ㅐ, ㅒ,ㅔ,ㅖ}는 네 부류로 분류한다.

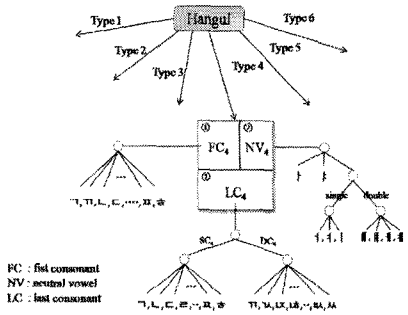


그림 3. 계층적 한글 인식기의 구조

1. 특징추출과 분류기

문자 데이터에서의 특징 추출은 방향 거리 분포(Directional Distance Distribution) 특징[13]을 사용하였다. 이는 흰 화소와 검은 화소 모두 거리 계산에 참여하는 방법이다. 또한 맵 타일링을 적용하여 타일된 맵에서 패턴의 거리 값이 다른 부분에 영향을 받게 한다. 유형별 분류기의 초성, 중성, 종성 인식기는 자소 영역에서 추출된 DDD특징을 입력으로 사용한다.

계층적 인식기의 모든 분류기는 다층 신경망(Multi Layer Percpetron)을 사용한다. 6-유형 분류에 1개, 초성에 6개, 중성에 20개, 종성에 9개를 사용하여 모두 36개의 MLP를 사용하였다. 훈련은 오류 역전파 학습 알고리즘을 사용하였으며, 양극 시그모이드 활성 함수를 적용하였다. 신경망의 파라미터인 학습률은 0.05로 일정하게 하였으며 관성항은 사용하지 않았다. MLP의 학습은 검증 집합에 대한 성능의 개선이 없을 때까지 하였다.

2. 6-유형 분류

6-유형 분류는 인쇄 한글 인식에서 사전 분류로서 일반적으로 사용되는 방법이다. 자소의 구조적 특성으로 나누어진 6 유형에 따라 한글을 분류하며 효과적임이 입증되어 있다. 이 사전 분류는 [그림 2]에서 보는 바와 같이 2,350자의 인식 문제를 각 유형별 글자 수에 해당하는 만큼의 작은 문제로 줄일 수 있다. 유형 분류의 절차는 아래 [그림 4]와 같다.

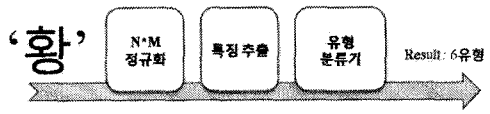


그림 4. 6-유형 분류 절차

문자 영상이 입력되면 N*M의 크기로 정규화를 한다. 그리고 정규화 영상에서 DDD 특징 추출 방법으로 특징을 추출한다. 추출된 특징을 분류기에 입력하여 출력에 따라 문자 영상의 유형을 결정하는 절차이다. 본 논문에서는 N과 M을 동일하게 64로 하였다.

3. 유형별 자소 특징

초성, 중성, 그리고 종성의 특징추출 영역은 문자 영상에서 각 자음과 모음이 표기되는 위치를 충분히 포함할 수 있도록 한다[11]. 각 유형별로 지정된 자소 인식 영역에서 DDD 특징을 추출한다. 예를 들어 1 유형 중성의 특징은 좌측 반절로 정했으며, 따라서 64*32의 영역을 점유한다. DDD 추출 알고리즘은 8*8 크기의 블록으로 나누고 각 블록에서 8방향과 흑백 각각에 대해 특징 값을 추출하므로 (64/8)*(32/8)*8*2=512개의 특징이 추출된다.

하지만 자음과 모음은 미리 지정된 영역을 대상으로 특징을 추출하므로 다른 자소의 획에 의한 간섭을 피할 수 없다. 이것은 초성, 중성, 그리고 종성을 인식하는데 상호간에 획 간섭이 있음을 의미한다. 다음의 [그림 5]는 이러한 간섭의 예를 보인다.

그림 5. 획 간섭의 예

첫 번째 글자인 '텐'은 중성이 'ㅔ'이지만 초성 자음의 영향으로 'ㅖ'의 오인식 가능성이 존재한다. 네 번째 글자인 '티'는 모음이 'ㅣ'임에도 불구하고 'ㅓ'로 인식될

수 있으며, 마지막 글자인 '쭝'은 'ㅡ'모음이지만 초성자음의 영향으로 'ㅗ'로 오인식될 가능성이 존재한다.

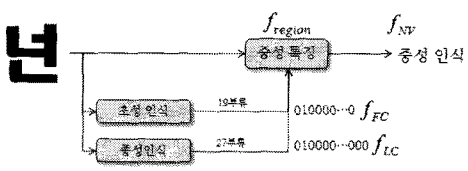


그림 6. 중성을 위한 특징

이러한 간섭은 자음보다 모음의 인식에 더욱 치명적이다. 한글은 구조적으로 모음은 초성 또는 종성의 획 간섭을 더 받으며 오인식 가능성이 더 크다. 이러한 오인식을 줄이기 위해 [그림 6]과 같이 중성 인식에 초성과 종성의 인식 결과를 이용한다.

[그림 6]에서 fregion은 중성 영역에서 추출한 특징 벡터를 의미하고, fFC와 fLC는 초성과 종성의 인식 결과를 이진 코딩한 것이다. 예를 들어 [그림 6]의 초성은 'ㄴ'으로 인식되었으며 따라서 두 번째 이진 특징 값은 1이고 나머지 18개 특징은 0 값을 가진다. 이 중성 인식기는 fregion은에 46 개의 특징을 추가한 특징 벡터를 사용한다.

4. 유형별 인식

유형별 인식에서는 초성, 중성, 그리고 종성을 따로 인식한다. 각각의 인식은 최소 1에서 최대 3단계의 과정을 거치는 계층적인 구조를 가지고 있다. 비록 인쇄체이지만 다양한 글꼴을 가정하면 자음과 모음의 형태와 위치가 다양하다. 그러므로 유형별로 그에 적합한 분류기를 설계하는 것이 효과적이다. 유형별 인식의 절차는 [그림 7]과 같다.

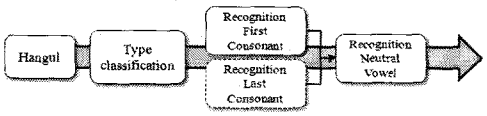


그림 7. 유형별 인식 절차

유형 분류가 되면 초성 또는 선택적으로 종성을 인식한다. 그리고 인식된 결과를 [그림 6]과 같이 중성의 입력에 포함시켜 중성을 인식한다. 각 자소에서의 인식 결과를 조합하여 출력을 결정한다.

4.1 초성 분류기

초성 분류기는 각 유형별 초성을 분류한다. 초성 분류기는 유형별로 독립적으로 설계되었다. [그림 8]은 초성 분류기의 구조를 나타낸다.

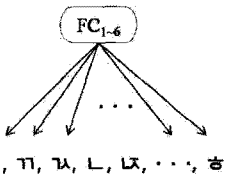
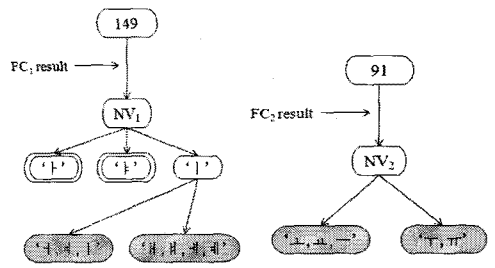


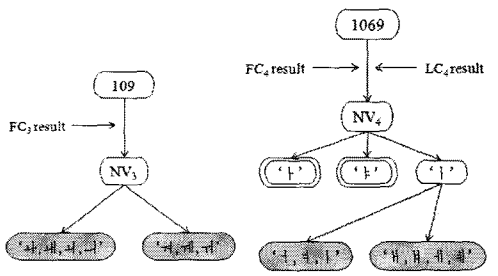
그림 8. 초성 분류기의 구조

4.2 중성 분류기

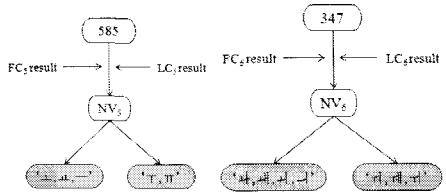
중성 분류기는 최대 3단계의 인식과정을 거친다. [그림 9]는 6-유형 각각에 대해 중성 분류기의 구조를 나타낸다.



(a) 1유형 중성 (b)2유형 중성



(c) 3유형 중성 (d) 4유형 중성



(e) 5유형 중성 (f) 6유형 중성

그림 9. 중성 분류기의 구조

4, 5, 6유형은 [그림 6]이 설명한 바와 같이 초성과 종성 인식 결과를 특징 벡터에 추가하여 사용하였다. 구조적으로 같기 때문에 1과 4유형, 2와 5유형, 그리고 3과 6유형은 동일한 구조를 적용하였다.

1과 4유형의 인식 과정은 다음과 같다.

1) {ㅏ, ㅑ, ㅣ-형}인지 분류하고 (3-부류 분류), ㅏ또는 ㅑ로 분류되면 거기서 멈추지만 ㅣ-형으로 분류되면 2)로 간다.
2) {{ㅓ,ㅕ,ㅣ},{ㅐ,ㅒ,ㅔ,ㅖ}}의 두 부류 중 하나로 분류한다.
3) {ㅓ,ㅕ,ㅣ}이면 3-부류 분류를 하고 {ㅐ,ㅒ,ㅔ,ㅖ}이면 4-부류 분류를 한다.

2와 5유형은 두 단계를 거친다. 모음의 횡 획을 기준으로 종 획의 방향을 구분하는 첫 단계를 거쳐 각각의 모음을 구분한다.

3과 6유형의 첫 번째 단계는 2와 5유형에서 사용한 횡 획을 기준으로 종 획의 유무를 판단하고, 두 번째 단계에서는 1과 4유형의 분류기준을 적용하여 'ㅘ,ㅙ,ㅚ, ㅝ'와 'ㅓ,ㅔ,ㅟ'를 구분한다.

4.3 종성 분류기

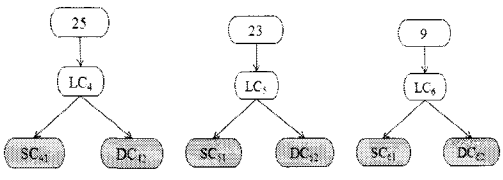


그림 10. 종성 분류기의 구조

종성은 단일 자음인 'ㄱ', 'ㄹ'등과 복합 자음 'ㄲ', 'ㅄ' 등으로 구성되어 있다. 이러한 사실에 따라 첫 번째 단계에서는 단자음 (SC)과 복자음 (DC)으로 2 부류 분류한다. 두 번째 단계에서는 단자음과 복자음 각각을 분류한다.

III. 실험 및 결과

1. 실험 환경

MLP의 학습은 PHD08 한글 데이터베이스[14]를 이용하여 수행하였다. PHD08은 9종류의 글꼴을 포함하고 있으며, 부류별 샘플의 수는 2,187개이다. 글꼴 별 10개의 샘플을 훈련 집합으로 하였고 그 외의 5개의 샘플을 검증 집합으로 정하였다. 나머지 2,052개 샘플은 시험 집합으로 정하였다.

기존의 한글 데이터베이스인 CBNU00[15]와 ETRI[16]에 대해 성능을 측정하였다. CBNU00 데이터베이스는 3가지 글꼴로 구성되어 있으며 글자당 1,200개의 샘플을 보유한다. ETRI 한글은 10가지 글꼴을 대하여 생성되었으며 글자당 193개가 테스트에 사용되었다.

2. 6-유형 분류와 유형별 분류 성능

6-유형 분류와 유형별 분류 성능은 PHD08의 시험 집합, CBNU00의 글자별 1,000개 샘플, ETRI데이터베이스의 글자별 193개의 샘플을 대상으로 측정하였다. [표 2]와 [표 3]은 6-유형과 유형별 정확률을 보여준다.

표 2. 유형 분류의 성능: 단위(%)

종류	1유형	2유형	3유형	4유형	5유형	6유형	인식률
PHD08	99.86	99.94	99.80	99.94	99.97	99.75	99.91
CBNU	99.25	99.28	96.87	99.65	99.91	97.05	99.23
ETRI	99.74	99.57	96.71	98.48	99.98	98.65	98.92

세 가지 데이터베이스에 대한 성능은 99%정도로 6-유형 분류의 성능이 비교적 높음을 알 수 있다. 3유형과 6유형이 다른 유형에 비해 성능이 낮다.

표 3. 유형별 인식 성능: 단위(%)

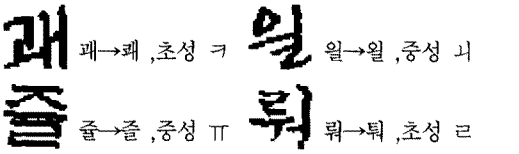
종류	1유형	2유형	3유형	4유형	5유형	6유형	인식률
PHD08	99.16	99.70	98.90	99.31	98.47	98.50	98.96
CBNU	93.19	95.54	91.06	88.57	87.03	87.00	88.63
ETRI	96.76	96.53	90.80	89.84	90.92	86.26	90.31

유형별 인식의 정확률은 PHD08은 98.96%이다. 하지만 훈련에 전혀 참여하지 않은 또한 다른 환경에서 수집한 CBNU00과 ETRI 데이터베이스에 대해서는 각각 88.63%와 90.31%이다. 종성을 가진 4, 5, 그리고 6유형이 인식 성능이 낮았다.

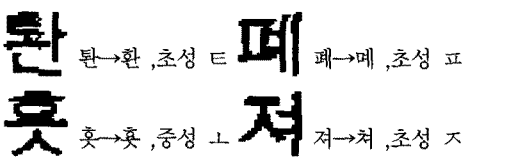
3. 결과 분석

[그림 11]은 각 데이터베이스에 대해 틀리게 분류된 샘플들을 보여 준다.

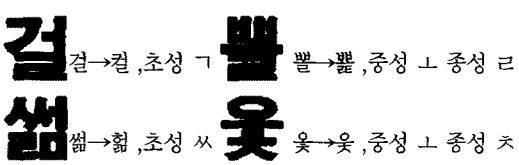
오분류를 분석해보면 초성, 중성, 종성 중의 하나에서 틀린 경우가 대부분이다. 예를 들어 [그림 11](a)의 첫 번째 샘플은 'ㄱ'을 'ㅋ'으로 오분류 하였다.



(a) PHD08



(b) CBNU00



(c) ETRI한글

그림 11. 오분류된 샘플

제안한 시스템은 분할 정복 (divide and conquer)과 같은 구조를 가진다. 따라서 특정 부분에서 심하게 오분류가 발생한다면 그곳만 대상으로 특징 또는 분류 알고리즘을 개선할 수 있다. 이것이 제안한 방법의 장점 중의 하나로 볼 수 있다.

V. 웹 영상의 문자 데이터에 대한 적용

이 절에서는 제안한 인식 시스템을 웹 영상의 문자 인식에 적용한 결과를 제시한다.

1. 실험 환경

웹 영상에서 텍스트를 추출하는 모듈은 [17]을 사용하였다. 이 모듈은 2-단계 컬러 분산 맵을 바탕으로 단어를 하나 또는 몇 개를 포함하는 이진 영상을 출력한다.

2. 실험 데이터

문자 샘플은 [그림 12]와 같은 다양한 광고 영상 등에서 추출하였다. [그림 13]은 이들 영상에서 추출한 텍스트 영역을 보여 준다.

그림 12. 광고 영상의 예

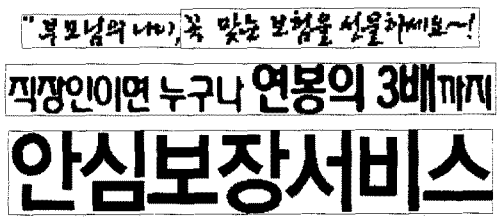
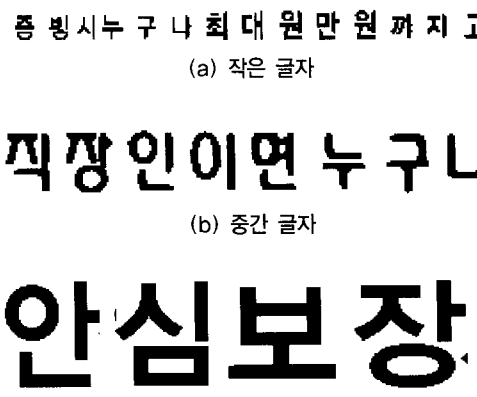


그림 13. 텍스트 추출 모듈의 출력



(a) 작은 글자

(b) 중간 글자

(c) 큰 글자

그림 14. 분할된 글자 샘플

추출 모듈이 출력한 이진 영상을 문자 단위로 분할하는 것은 수작업으로 수행하였다. 샘플의 크기는 최소 12*8, 최대 68*64이었다. [그림 14]는 이렇게 얻은 샘플을 보여 준다. 성능 측정의 편의를 위해 작은 문자, 중간 문자, 큰 문자의 3가지로 구분하였다. 작은 문자는 크기가 15*13 이하이며, 중간 문자는 30*30 미만, 큰 문자는 30*30 이상의 크기를 가진다. 샘플은 총 173개이며, 작은 글자는 95개, 중간 글자 39개, 그리고 큰 글자 39개이다. 작은 글자들은 글자 추출 과정에서 획이 병합되고 일부 인식에 불리한 노이즈가 삽입된 것을 관찰할 수 있다. 중간 글자의 경우에는 다른 크기의 글자들보다 다양한 글꼴이 포함되어 있음을 알 수 있다.

3. 인식 성능

웹 영상의 문자 데이터에 대한 성능 측정은 173개 샘플에 대하여 측정되었다. [표 4]는 글꼴 크기에 따른 인식 성능을 나타낸다. 작은 글자는 95개중 48개를 옳게 인식하여 50.5%의 정확률을 보이며, 작은 글자와 큰 글자는 각각 27개와 26개를 옳게 인식하여 69.2%와 66.7%의 정확률을 보였다. 전체 정확률은 58.3%이다.

표 4. 웹 영상 내 글자 인식 성능: 단위 (%)

	작은 글자	중간 글자	큰 글자
성능	50.5	69.2	66.7

웹 영상은 낮은 성능을 보였다. 웹 영상은 미적인 요소를 중요한 요소로 보고 디자인하므로 다양한 모양의 폰트가 나타난다. 또한 압축 영상으로 저장되기 때문에 획이 왜곡되는 현상이 나타난다. 제안한 인식 알고리즘은 이러한 현상을 충분히 극복하지 못하였다.

IV. 결 론

본 논문은 계층적 구조의 인쇄 한글 인식기를 제안하였다. 인식기는 6-유형 분류기와 유형별 분류기로 구성되어 있다. 유형별 분류기는 초성, 중성, 그리고 종성에 대해 그들에 적합한 방식으로 분류하였다. 또한 획의 간섭을 고려하기 위해 초성과 종성을 먼저 인식하고 그 결과를 중성 인식에 반영하였다.

향후 연구로는 자소의 인식 영역을 동적으로 조절하는 방법의 개발하고 분류 단계별로 최적의 특징 추출 방법을 적용하는 연구가 필요하다. 특히 웹 영상의 특성을 반영하여 웹 영역에서의 인식률을 높이는 연구가 뒤따라야 한다.

참 고 문 헌

[1] 정인숙, 오일석, "분산맵을 이용한 웹 이미지 텍스트 영역 추출," 한국콘텐츠학회논문집, 제9권, 제9호, pp.68-79, 2009.

[2] 이주근, "한글 문자의 인식에 관한 연구," 전자공학회지, 제9권, 제4호, pp.35-32, 1972.

[3] http://www.kssn.net/

[4] 최환수, 정동철, 공성필, "잡영과 왜곡이 심한 한글 문자의 자소분리 및 인식에 관한 연구", 한국통신학회논문집, 제22권, 제6호, pp.1160-1169, 1997.

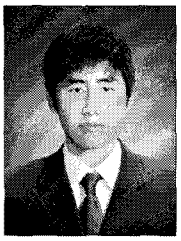
[5] 김민기, 권오성, 권영빈, "모음의 구조적 형태와 조합 규칙에 충실한 한글 문자의 유형 분류", 정보과학회논문지, 제25권, 제4호, pp.685-695, 1998.

[6] 이관호, 장희동, 남궁재찬, "동적자소분할과 신경망을 이용한 인쇄체 한글 문자인식에 관한 연구",

한국통신학회논문지, 제19권, 제11호, pp.2133-2146, 1994.
[7] 김병기, "유형의 상대적 크기를 고려한 한글문자의 유형 분류", 컴퓨터정보학회논문지, 제11권, 제6호, pp.99-106, 2006.
[8] 이성훈, 조규태, 김진식, 김진형, 정철곤, 김상균, 문영수, 김지연, "저해상도 인쇄체 한글 영상 인식을 위한 자소 분할 방법", 한국컴퓨터종합학술대회 논문집, 제33권, 1호, pp.382-384, 2006.
[9] 홍순재, 김백섭, "신경망을 이용한 인쇄체 한글의 자소 분리 및 인식", 한국정보과학회논문집, 제23권, 제1호, pp.285-288, 1996.
[10] 조성배, 김진형, "인쇄체 한글문자의 인식을 위한 계층적 신경망", 한국정보과학회 논문지, 제17권, 제3호, pp.306-316, 1992.
[11] 이진수, 권오준, 방승양, "개선된 자소 인식 방법을 통한 고인식률 인쇄체 한글 인식", 정보과학회논문지, 제23권, 제8호, 1996.
[12] 임길택, 김호연, "문자형식 분류 기반의 인쇄체 문자인식에 관한 연구", 전자공학회논문집, 제40권, 제5호, pp.266-279, 2003.
[13] 오일석, Ching Y.Suen, "광학 문자 인식을 위한 거리 특징", 정보과학회논문지, 제25권, 제7호, pp.1028-1043, 1998.
[14] 함대성, 이득용, 정인숙, 오일석, "한글 문자 데이터베이스 PHD08 구축", 한국컨텐츠학회논문지, 제11권, 제8호, 2008.
[15] http://cv.chonbuk.ac.kr/
[16] http://ai.kaist.ac.kr/
[17] I. S. Jung, D. S. Ham, D. R. Lee, and I. S. Oh, "Two-Level Text Segmentation from Web Document Images based on Variance Analysis," Korea and Japan Joint Workshop on Pattern Recognition, pp.107-108, 2008.

저 자 소 개

함 대 성(Dae-Sung Ham) 정회원



- 2007년 2월 : 전북대학교 컴퓨터공학(공학사)
- 2009년 3월 : 전북대학교 컴퓨터공학 석사
- 2009년 3월 ~ 현재 : 피엔에스테크놀러지(주) 주임연구원

<관심분야> : 컴퓨터비젼, 패턴인식, 한글인식

이 득 용(Duk-Ryong Lee) 정회원

- 2004년 8월 : 전북대학교 컴퓨터공학(공학사)
- 2006년 8월 : 전북대학교 컴퓨터공학 석사
- 2006년 9월 ~ 현재 : 전북대학교 컴퓨터공학 박사 수료

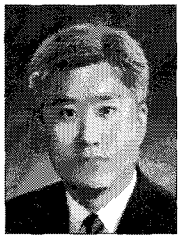
<관심분야> : 컴퓨터비젼, 패턴인식

최 경 웅(Kyung-Ung Choi) 정회원

- 1996년 2월 : 호원대학교 정보통신공학과 학사
- 1998년 2월 : 전북대학교 영상정보공학과 석사
- 2002년 2월 : 전북대학교 정보통식공학과 박사 수료
- 2004년 6월 ~ 현재 : (주)아워텍 대표이사

<관심분야> : 텍스트 마이닝, 영상검색

오 일 석(Il-Seok Oh) 정회원



- 1984년 : 서울대학교 컴퓨터공학과(공학사)
- 1992년 : KAIST 전산학과 박사
- 1992년 9월 ~ 현재 : 전북대학교 컴퓨터공학 석사과정

<관심분야> : 컴퓨터비젼, 패턴인식