韓國 컴퓨터情報學會 論文誌
第11卷 第6號, 2006. 12

2006-11-6-1-11

유형의 상대적 크기를 고려한 한글문자의 유형 분류

김 병 기*

Type Classification of Korean Characters Considering Relative Type Size

Pyeoung-Kee Kim*

요 약

한글과 같이 문자집합이 큰 조합 문자의 인식을 위해서는 문제공간을 줄여주는 유형분류가 큰 도움이 된다. 기존 연구들이 한글 구성원리에 치중하여 한글 유형을 정한 결과 복모음 문자에 대한 정확한 분류가 어려웠고 문자집합이 상대적으로 큰 종성 있는 문자들에 대한 세분류가 부족하여 문제공간의 분배에 어려움이 많았다. 본 논문에서는 이러한 문제들을 해결하고자 수평 투영 프로파일을 이용하여 안정적 추출이 가능한 횡모음을 우선 추출하고, 수평 투영 프로파일과 연결요소를 이용하여 종성 있는 문자들에 대하여 종성을 5가지 그룹 중 하나로 세분류 하는 유형분류 방법을 제안하였다. 기존의 유형분류 방법들이 유형간 크기 불균형을 갖는 6개 혹은 15개의 유형을 가진 반면에 제안한 방법은 균형 있고 안정적 분류가 가능한 19개의 유형을 갖는다. 한글 잦기순 1,000자에 대한 7개의 상용 글꼴자료를 사용하여 분류 시스템을 만들고 월간지에서 스캔(Scan)한 30,614자에 대한 유형 분류 실험을 통하여 제안한 방법이 다양한 글꼴과 큰 문자집합을 갖는 한글 문자의 유형분류에 효율적임을 확인하였다.

Abstract

Type classification is a very needed step in recognizing huge character set language such as korean characters. Since most previous researches are based on the composition rule of Korean characters, it has been difficult to correctly classify composite vowel characters and problem space was not divided equally for the lack of classification of last consonant which is relatively bigger than other graphemes. In this paper, I propose a new type classification method in which horizontal vowel is extracted before vertical vowel and last consonants are further classified into one of five small groups based on horizontal projection profile. The new method uses 19 character types which is more stable than previous 6 types or 15 types. Through experiments on 1,000 frequently used character sets and 30,614 characters scanned from several magazines, I showed that the proposed method is more useful classifying Korean characters of huge set.

▸ Keyword : 유형 크기(Type Size), 투영 프로파일(Projection Profile), 연결요소(Connected Component).

• 제1저자 : 김병기
• 접수일 : 2006.11.17, 심사일 : 2006.11.18, 심사완료일 : 2006. 12.25
* 신라대학교 컴퓨터정보공학부 교수

Ⅰ. 서론

한글 문서 자동인식에 대한 연구는 1960년대부터 꾸준히 진행되어 왔다. 문서 인식기 개발을 위해서는 문서의 기울기 보정, 문자 영역 추출, 문자 인식, 후처리 등의 다양한 처리 루틴의 개발이 필요하다. 이러한 처리 루틴 중에서 가장 인식률에 직접적인 영향을 주는 루틴은 문자 인식 루틴이다. 99% 이상의 인식률을 보이는 영문자 인식기들의 경우와는 달리 한글 문자 인식기는 여전히 인식률 개선이 필요한 실정이다. 2000년에 조사한 한글문서 상용 인식기 중 최고 평균 인식률은 92.96%였고, 상용인식기임에도 불구하고 어떤 인식기의 경우에는 평균 인식률이 83.27%에 불과한 것이 실정이다[1].

문자 인식에서 93%라는 인식률은 실용적으로 사용하기에 무리가 있는 수치이다. 더욱이, 잡영이나 인쇄상태가 흐린 경우와 같은 특징을 갖지 않는 깨끗한 문서에 대한 인식률조차 최고 인식률이 97% 이하라는 사실은 한글 문자 인식에 대한 연구가 여전히 필요하다는 점을 시사한다. 예를 들어, 2000자로 구성된 문서 한 쪽을 인식할 경우 60여자가 오인식 되거나 미인식 된다는 뜻이다. 어떤 문자에 대하여 오인식인지 미인식이 발생했는지 정확하게 알 수 없기 때문에 원칙적으로 전체 인식대상 문자를 모두 사람이 확인해야 하므로 이로 인한 수작업은 생각보다 많은 시간을 요하게 되어 문서 자동인식기의 효율을 크게 떨어뜨리게 된다.

이와 같이 영문자의 경우와 달리 한글문자의 인식이 어려운 것은 한글 문자가 조합 문자이기 때문에 인식대상 문자 집합의 수 자체가 많을 뿐만 아니라 글꼴과 한글 유형에 따라 문자를 구성하는 자소의 크기, 모양, 위치 등이 다르기 때문에 결과적으로 패턴 문제 공간이 매우 크기 때문이다. 한글 문자 인식이 갖는 이러한 문제를 해결하기 위하여 입력된 한글 문자에 대하여 인식단계 이전에 유형을 먼저 분류하려는 다양한 연구들이 있어 왔다[2-10]. 기존 연구들이 나름대로 의의가 있지만 크게 세 가지 면에서 안정적인 유형 분류가 어려운 것이 사실이다. 첫째, 기존 연구들은 다양한 글꼴에 대한 고려가 부족한 것이 사실이다. 이는 기존 연구에 사용된 실험 데이터가 명조, 신명조, 고딕과 같은 고전적인 몇 개의 글꼴로 이루어진 것이 주요한 원인이다. 따라서 동일한 문자라도 글꼴에 따라서 상당히 다른 모양을 갖기 때문에 특정 글꼴 문자는 잘 인식하지만 다른 글꼴 문자에 대하여는 형편없는 인식률을 보일 수 있다. 현재의 문서 편집기와 같은 소프트웨어들은 수십 개의 글꼴을 내장하고 있으므로 이는 개선되어야 할 부분이다. 둘째, 기존 연구들에서는 복모음 문자를 구분하기 위하여 횡모음 존재 여부를 확인하는 방법을 주로 사용하는데, 신명조체와 같이 모음 가로획에 대한 투영 프로파일이 뚜렷하지 않은 많은 글꼴의 경우에 이에 대한 안정적인 검사가 매우 어려운 것이 사실이다. 셋째, 문제 공간을 크게 만드는 요인 중의 하나인 종성 분류에 대한 연구가 거의 없었다. 모음과 종성유무에 따른 분류가 제대로 된다 하더라도 종성을 갖는 문자 집합의 크기가 다른 집합보다 여전히 커서 유형분류의 가치를 반감시키게 되는 것이다.

패턴 분류나 패턴 인식 시에 방법론에 앞서 중요한 것이 대상 패턴의 정보량(Entropy)이나 특징(Feature)이 무엇이며 어디에 존재하는가를 파악하는 것이다. 한글 문자는 구조상 다수의 모음과 자음에서 수평 투영 프로파일(Projection Profile)을 많이 갖고, 정보량은 글꼴에는 거의 무관한 대신 자소 구성의 구조적 유형에 따라 변하며, 문자들 간을 구별하는 대부분의 정보량이 문자의 가장자리 부분에 위치한다[11]. 따라서 문자 인식기 개발 시 글꼴에 의존적인 인식 보다는 문자의 구조적 특징을 이용하는 방법이 유리하고 문자 내부나 전체 영역을 동일한 비중으로 처리하기 보다는 가장자리의 특징을 이용하는 것이 유리하다.

Ⅱ. 한글문자의 유형 분류

2.1 패턴인식 대상으로서의 한글 문자

한글 문자는 초성 19개, 중성 21개, 종성 27개의 자소로 구성된 조합 문자이고, 모음의 종류와 종성 유무에 따라 11,172자의 조합 가능한 문자 집합을 갖는다. 예를 들어, 10개 글꼴에 대응 가능한 한글문자 인식기를 만든다면 총 11만 문자에 해당하는 문제 공간(Problem Space)을 갖는 패턴인식 문제를 해결하는 것이 된다. 따라서 문자 전체를 대상으로 하는 문자 인식기 보다는 입력 문자를 자소 조합을 기준으로 적절한 개수의 유형으로 분리하는 유형 분류를 하는 것이 정확하고 속도가 빠른 인식기를 개발하는데 도움이 클 것이다.

인식 대상으로서 한글 문자의 구조적 특징으로는 다음과 같은 것을 들 수 있다. 첫째, 한글 문자는 초성+중성, 또는 초성+중성+종성으로 구성되고 구성방법에 따라 전통적으로 6가지 내지 15가지 유형으로 분류된다[2]. 둘째, 한글

문자는 동일 글꼴 동일 크기의 문자 내에서도 종성 유무와 중성의 종류에 따라 초성과 종성의 모양, 크기, 위치가 달라진다. 셋째, 종모음을 갖지 않고 횡모음만 갖는 문자의 경우에는 글꼴에 거의 무관하게 횡모음은 가장 긴 가로 프로파일을 형성하여 다른 자소보다 상대적으로 그 존재유무를 확인하기 용이하다.

자음과 모음 구성에 따른 한글 유형들과 문자 수를 분류하면 표 1에 나타낸 것과 같다. 표 1을 보면, 자소 중에서 종성의 개수가 초성이나 모음과 같은 다른 자소 보다 상대적으로 많기 때문에, 유형 1, 2, 3에 종성이 추가된 유형 4-6의 문자 집합이 큰 것을 알 수 있다. 따라서 입력 문자에 대한 정확한 유형 및 자소 분리는 크고 복잡한 문제 공간을 보다 작고 간단한 소수의 문제 공간으로 변환하는 효과를 가지므로 인식률과 속도 개선에 매우 의미 있는 방법이다. 특히, 모음보다 자음의 개수가 많으므로 자음을 세분하면 문제 공간을 보다 적절한 크기로 줄일 수 있을 것이다.

표 1. 자소구성에 따른 한글 문자 유형별 문자수
Table 1. Character numbers in Korean characters based on grapheme composition

유형	유형 1	유형 2	유형 3	유형 4	유형 5	유형 6
	초성+종모음	초성+횡모음	초성+복모음	초성+종모음+종성	초성+횡모음+종성	초성+복모음+종성
조합가능 문자 수	171 초성 19 x 종모음9	95 초성 19 x 횡모음 5	133 초성 19 x 복모음 7	4617 유형1 171 x 종성 27	2565 유형2 95 x 종성 27	3591 유형3 133 x 종성 27
문자 예	가, 허, 기	고, 구, 그	과, 워, 의	각, 연, 일	공, 윤, 을	관, 궐, 월

2.2 제안한 유형 분류의 개관

패턴 인식을 할 경우에는 가장 중요한 요소 중의 하나가 특징(Feature)의 추출이다. 그런데 한글 문자의 경우에는 각 자소를 구분하는 특징이 문자 영역 중 가장자리에 산재하고, 문자 유형에 따라 동일 글꼴에서조차 자소의 모양, 위치, 크기가 달라지기 때문에 문자영역 전체에 대한 통계값의 사용이나 신경망과 같은 방법론적인 해결책의 사용보다는 자소 구성에 기반 하는 구조적 특징의 추출이 보다 중요하다.

한글 문자에서 모음은 비교적 긴 프로파일을 가지므로 기존 연구들에서 종모음을 기준으로 하는 유형 분류 시도가 많이 있었다. 그러나, 종모음 우선 추출의 경우에는 종성 'ㄱ', 'ㄹ', 'ㅁ', 'ㅂ' 과 이들의 조합으로 구성되는 복자음 종성의 일부분이 종모음으로 간주 되는 문제가 있다. 복모음과 종모음 문자를 구분하기 위하여 복모음에서 횡모음의 존재 유무를 확인하는 방법도 많이 사용되어 왔다. 복모음의 경우에는 글꼴과 문자 유형에 따라 자음이 복모음의 횡모음 요소로 오분류 되거나, 글꼴에 따라 복모음의 모음 부분 존재를 알지 못하는 경우가 많았다.

따라서 본 논문에서는 이러한 유형 분류에 관한 문제점을 해결하고자 다음과 같은 유형분류 방법을 사용한다. 첫째, 기존의 종모음 추출에 기반한 유형 분류 대신에 종모음이 없는 횡모음 문자인지를 먼저 결정하는 횡모음 우선인식 방법을 사용한다. 종모음은 종성과의 연결 관계나 기울기를 포함한 글꼴에 따라 그 존재여부를 정확히 추출하기가 어려운 경우가 많은 반면에, 횡모음의 경우에는 가장 긴 가로 프로파일을 갖는다는 특징을 갖기 때문에 글꼴에 관계없이 그 존재여부를 정확히 알 수 있다.

둘째, 복모음 문자에서 횡모음을 별도로 추출하지 않고 초성과 횡모음을 하나의 추출 단위로 간주하여 분류한다. 기존의 관련 연구에서는 입력문자가 복모음 문자인지 알기 위하여 횡모음 부분을 추출하는 방법을 사용하기 때문에, 글꼴에 따라 횡모음 부분의 안정적 추출이 불가능한 경우가 많고, 초성이나 종성의 일부가 횡모음 영역으로 잘못 간주되는 경우도 많다. 뿐만 아니라, 서로 다른 종모음과 복모음을 갖는 두 개의 문자가 매우 유사하여 이 둘 간의 분류가 어려운 경우도 많다.

셋째, 유형 분류 시에 모음의 종류와 종성의 유무에 따른 분류 외에 종성 분류를 추가적으로 사용하였다. 기존 한글 문자의 유형 분류 연구들에서는 모음의 종류와 종성의 유무를 근거로 6가지 혹은 15가지의 유형으로 분류하여 왔고, 자음의 세부 분류에 대한 연구가 거의 없었다. 그러나 실제적인 한글 유형 분류를 위하여 다음과 같은 이유에서 자음의 분류가 필요하다.

먼저, 종성의 숫자가 개별 모음군에서의 숫자 보다 많다. 횡모음, 종모음, 복모음이 각각 5개, 9개, 7개로 구성된 반면에 종성은 27개의 자음들로 이루어져 있다. 표 1에서 본 바와 같이 종성 때문에 유형 4-6의 문자집합이 커진 것이다. 따라서 적절한 종성군의 유형 분류는 큰 문제 공간을 적절한 크기로 분산된 다수의 작은 문제 공간들로 변환할 수 있는 것이다. 다음으로, 기존 연구들에서 종모음 문자와 복모음 문자를 서로 다른 유형으로 분리하여 왔으나, 신명조나 명조체에서와 같이 복모음에서의 횡모음 수평획이 오른쪽 윗 방향으로 휘어져 있는 등의 이유로 수평 프로파일이 제대로 형성되지 않는 경우가 많아서 횡모음 성분의 존재여부를 알기가 매우 어렵다. 마지막으로, 글꼴에 따라 초성, 복모음, 종성을 갖는 상대적으로 획이 많이 포함된 문자

들에서 초성이나 종성 자음의 일부가 횡모음으로 오인되는 경우도 많다. 결과적으로, 복모음 문자인지 종모음 문자인지를 알기 위하여 횡모음 존재 여부를 알고자 하는 문제가 자소나 문자 인식에 버금가는 난이도를 갖게 되어, 문자 인식기의 부담을 덜어 주려는 유형 분류 본래의 취지가 훼손될 수 있다. 따라서 본 논문에서는 복모음과 종모음을 하나의 모음 군으로 간주하여 이러한 종모음-복모음 오분류가 발생하지 않도록 한다.

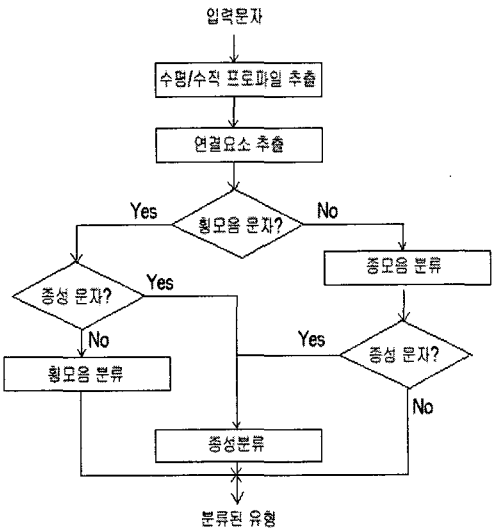

그림 1. 한글문자 유형분류 전체 흐름도
Fig 1. Flow of the proposed system

그림 1에 유형분류에 대한 전체적인 흐름을 보였다. '고', '으', '는', '을' 자와 같이 종모음이 없이 횡모음만 갖는 문자를 횡모음 문자라고 칭하자. 입력문자에 대하여, 횡모음이 있는지 확인하고 횡모음이 있는 경우에는, 표 2에 나타낸 바와 같이, 종성없는 'ㅗ,ㅛ,ㅡ' 횡모음 문자(유형 3), 종성 없는 'ㅜ,ㅠ' 횡모음 문자(유형 4), 종성있는 'ㅗ,ㅛ,ㅜ,ㅠ,ㅡ' 횡모음 문자의 세가지 중 하나로 분류한다. 이 중, 세 번째 경우인 종성 있는 횡모음 문자는 다시 각각 종성의 종류에 따라 유형 15에서 유형 19중 하나로 세분된다.

횡모음이 없는 문자는 종모음이나 복모음을 갖게 되고, 종모음 부분을 검사하여 'ㅏ,ㅑ,ㅘ' 종모음을 갖는지와 종성 유무를 검사하여 표 2에 나타낸 유형 1, 2, 5-14 중 하나로 분류한다. 종성이 있는 경우는 횡모음 문자에서와 같이 종성의 종류에 따라 유형 5-14 중 하나로 세분된다.

표 2. 모음과 종성에 따른 19가지 한글 유형
Table 2. 19 Korean character types based on vowel and last consonant

한글문자 유형	사용되는 모음	자소 구성
유형 1	ㅏ,ㅑ,ㅘ 종/복모음	초성+종/복모음
유형 2	ㅏ,ㅑ,ㅘ 모음을 제외한 모든 종/복모음	초성 + 종모음
유형 3	ㅗ,ㅛ,ㅡ 횡모음	초성 + 횡모음
유형 4	ㅜ,ㅠ 횡모음	초성 + 횡모음
유형 5-9	ㅏ,ㅑ,ㅘ 종/복모음	초성+종/복모음 + 종성
유형 10-14	ㅏ,ㅑ,ㅘ 종모음을 제외한 모든 종모음	초성+종/복모음 + 종성
유형 15-19	모든 횡모음	초성 + 횡모음 + 종성

2.3 횡모음 문자의 분류

횡모음 문자에 대하여 수평 프로파일을 구하면 횡모음의 수평획 부분에서 글자 폭과 거의 같은 크기의 프로파일을 갖는다. 횡모음 문자의 존재 여부와 위치는 다음과 같은 간단한 투영 프로파일 조사만으로도 알 수 있다.

(1) 횡모음 문자는 문자 높이의 1/3 지점 아래에서 문자폭의 90%가 넘는 "가장 긴" 수평 투영 프로파일(Profile_Max_H)을 갖는다.

(2) Profile_Max_H 좌우 끝 부분의 아래 위쪽으로는 각각 문자 상단과 하단에까지 이르는 긴 수직 프로파일이 연결되지 않는다.

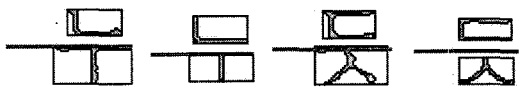

그림 3. 횡모음 문자에 대한 수평 프로파일의 예
Fig 3 Example profiles of horizontal-vowel characters

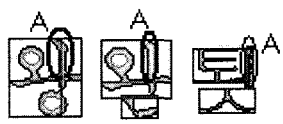

그림 4. 횡모음 분류의 예외조건 처리
Fig 4. Processing of exceptional conditions in horizontal-vowel classification

그림 3에 횡모음 문자에 대한 수평 프로파일의 예를 나타내었다. 그림에서 문자 왼쪽에 표시된 수평선이 횡모음 존재를 나타내는 수평 투영 프로파일이다. 횡모음 분류시에 복모음의 종모음 일부가 횡모음과 연결되어 수평 프로파일 값이 문자폭과 같이 되어 횡모음 문자로 오인될 수 있다. 위의 횡모음 검사 조건 두 번째는 이러한 경우를 알기 위한 것이다. 첫 번째 조건을 만족하여 입력문자가 횡모음 문자로 분류되면 횡모음과 연결되어 있으면서 아래쪽이나 위쪽 방향으로 긴 세로 프로파일이 있는지 검사하여 그러한 프로파일이 있으면 횡모음 문자가 아닌 것으로 예외 처리한다. 궁서체 '왕', '완' 글자와 고딕체나 굴림체에서 '툇'과 같은 문자들에서 이러한 경우를 볼 수 있는데, 그림 4에 예들을 나타내었다. 그림에서 타원으로 둘러싸고 'A'로 표시한 부분에 위쪽 방향으로 횡모음과 연결된 긴 프로파일이 있으므로 이 문자들은 첫 번째 조건을 만족하지만 횡모음 문자가 아닌 것으로 분류된다.

횡모음 문자는 'ㅗ', 'ㅛ', 'ㅜ', 'ㅠ', 'ㅡ' 의 5개 횡모음을 갖는데 작은 유형들로 세분하기 위하여 종성 유무를 확인한다. 종성이 없는 횡모음 문자에서 'ㅗ', 'ㅛ', 'ㅡ' 횡모음은 문자의 최하단에 위치하므로 수평 프로파일의 세로상의 위치만으로도 이들 모음의 존재 여부와 종성 유무를 알 수 있다. 종성이 없는 'ㅜ', 'ㅠ' 횡모음 문자는 횡모음 수평 프로파일의 가운데 아래쪽에 각각 1,2 개의 세로 프로파일의 존재 여부를 검사하여 확인할 수 있다. 따라서, 횡모음 문자는 (1)종성이 없는 'ㅗ', 'ㅛ', 'ㅡ' 횡모음 문자, (2) 종성이 없는 'ㅜ', 'ㅠ' 횡모음 문자, (3) 종성 있는 횡모음 문자로 나누고, 종성 있는 횡모음 문자는 다시 종성의 수평 프로파일 개수와 그 위치에 따라 5가지로 세분하여 횡모음 문자로 나눈다. 결국, 횡모음 문자는 총 7가지 횡모음 문자 유형(유형 3,4,15-19) 중 하나로 분류된다. 종성과 같이 초성을 세분하지 않는 이유는 종성의 경우와는 달리 초성은 그 하단 위치를 알기 어려워서 횡모음 일부가 초성 수평 프로파일에 포함될 수도 있기 때문이다.

2.4 종모음/복모음 문자의 분류

앞 단계에서 횡모음 문자로 분류되지 않은 모든 문자는 종모음 문자이거나 종모음과 횡모음을 동시에 갖는 복모음 문자이다. 본 논문에서는 종모음 문자와 복모음 문자를 따로 분류하지 않는다. 그 이유는, 전술한 바와 같이, 이 두 유형간 분류 작업의 복잡도가 유형 분류가 아닌 자소나 문자인식 수준의 정밀성이 요구되는 경우가 있기 때문이다.

예를 들어 그림 5의 고딕체 문자 '판'과 '관'을 보자. (a)의 '판' 자는 종모음 문자이고 (b)의 '관' 자는 복모음 문자이며, (c)는 이 두 문자를 겹쳐 놓은 것이다. (c)에서 보는 바와 같이, 두 문자는 자모 구성상 모음에 관한한 전혀 다른 문자이지만 패턴의 입장에서 보면 거의 동일한 문자이며 그 차이가 프로파일을 이용하여서는 거의 구분하기 어려운 수준임을 볼 수 있다. 이러한 예는 '귀'와 '커' 자를 비롯하여 다양한 글꼴의 여러 문자에서 나타나고, 이러한 문자들 간의 구분은 분류가 아닌 인식의 단계에서 처리해야할 일이다.

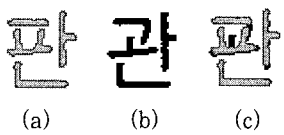

그림 5. '판' 과 '관' 문자의 유사성
Fig 5. Similarity between '판' and '관'

그림 6. 'ㅏ,ㅑ,ㅘ' 종/복모음 분류의 예
Fig 6. Example classification of 'ㅏ,ㅑ,ㅘ' vowel

종모음은 'ㅏ,ㅑ,ㅘ' 종/복모음과 그 밖의 종/복모음으로 나눈다. 이 세 가지 모음은 긴 수직 프로파일 우측에 작은 수평 프로파일의 존재 여부를 파악함으로써 간단히 모음 여부를 알 수 있다. 그림 6의 '가' 문자에서 종모음 종류를 알기 위한 검색 영역의 예를 나타내었다. 이 세가지를 제외한 그 외의 종/복모음은 종모음과 초성간의 경계를 정확하게 알기 힘들기 때문에 종/복모음을 오분류 할 수 있는 소지를 갖는다. 예를 들어 '뻬' 나 '뻔' 자와 같이 초성으로 'ㅃ'을 갖는 문자의 경우에는 종모음을 정확히 분류하는 것이 매우 어렵다.

2.5 종성의 확인과 분류

종/복모음 문자는 종성 부분의 수평 프로파일 정보와 연결 요소 정보를 이용하여 종성이 있는 문자와 종성이 없는 문자로 나눈다. 모든 자소가 연결된 경우를 제외하면 연결 요소 정보는 종성 유무 확인에 많은 정보를 제공한다. 종성의 존재 여부는 다음과 같은 방법을 사용하여 확인한다.

(1) 문자 영역 최 하단에서의 긴 수평 프로파일의 존재 유무 정보를 이용한다.

(2) 연결요소를 이용하여 종성을 갖는 유형인지를 검사하여 종성 존재 여부를 확인한다.

먼저, 'ㄴ', 'ㄷ', 'ㄹ', 'ㅁ' 과 같은 종성은 문자 최 하단에 일정 크기 이상의 수평 프로파일을 가지므로 이 수평 프로파일의 확인만으로도 종성 유무를 알 수 있다. 두 번째, 연결 요소정보를 이용하여 종성이 있는지 확인할 수 있다. 문자내 모든 자소가 연결된 경우를 제외한 경우에 종성을 갖는 문자들은 종성을 갖지 않는 문자들과 달리 고유한 연결요소 모양들을 가진다.

그림 7에 다양한 유형의 연결요소의 경우 중 일부를 나타내었다. (a)의 경우에는 초성, 종모음, 종성이 분명하게 구분된 경우이다. (b)는 종모음과 종성이 연결되어 하나의 연결요소로 추출된 경우로서, 우측의 문자높이 만큼의 높이를 갖는 연결요소와 그 왼쪽 상단 또는 상단 안쪽의 상대적으로 작은 연결요소의 존재로써 확인할 수 있다. (c)에는 초성 영역에 두 개의 연결요소를 갖는다는 점을 제외하곤 (b)의 경우와 같다. (d)는 초성과 종성이 큰 하나의 연결요소를 형성하고 그 우측에 작은 종모음 연결요소를 갖는 경우이다. (e)는 복자음 종성을 가짐으로써 자음 영역이 두 개의 연결요소를 갖는 경우이고, (f)는 초성과 종모음이 하나의 연결요소를 형성한 경우이다.

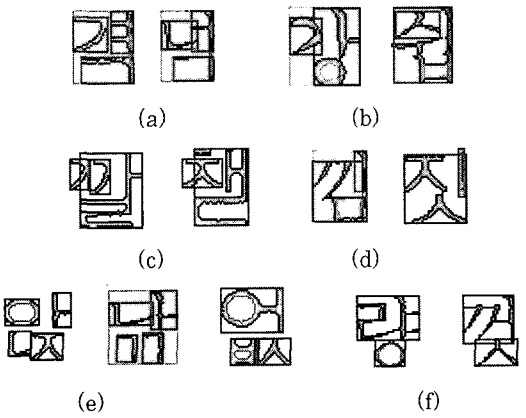

그림 7. 연결 요소를 이용한 종성 확인
Fig 7. Verification of last consonant using connected component

종성의 분류는 문자 폭의 1/3 크기 이상을 갖는 수평 투영 프로파일의 개수와 위치로 세분할 수 있다. 투영 프로파일은 유형분류나 문자간 분류에 많이 사용되는 정보이다[12]. 표 3에 가로 투영 프로파일(수평특징) 개수와 위치에 따른 종성 유형을 나타내었다. 결과적으로 종/복모음 문자는 종성이 없는 2개의 유형과 종성 있는 10개의 유형을 포함하여 총 12개의 유형중 하나로 분류 된다.

표 3. 종성 그룹 유형
Table 3. Type of last consonants

	수평특징 갯수	구성 자음(종성)
그룹1	상단수평특징 1	ㄱ,ㄲ,ㅈ,ㅉ,ㅋ,ㅊ,ㅎ
그룹2	하단수평특징 1	ㄴ
그룹3	수평특징 2	ㄷ,ㄸ,ㅁ,ㅂ,ㅍ,ㅃ,ㅆ,ㅅ,ㅎ
그룹4	수평특징 3	ㄹ,ㄺ,ㄻ,ㄼ,ㄽ,ㄾ,ㅌ
그룹5	수평특징 0	ㅅ,ㅆ,ㅇ

III. 실험 및 결과

본 논문에서 제안한 방법의 효용성을 알아보기 위하여 한글문자 분류 실험을 하였다. 유형 분류기를 구성하기 위하여 한글 잦기순 상위 1,000자를 대상으로 한글 워드프로세서의 7가지 글꼴을 사용하였다. 한글 잦기순 상위 1,000자에는 한글 문자 집합에서 사용되는 모든 자소가 포함되어 있으므로 유형분류기 구성에는 충분한 문자 집합이다. 10 포인트 크기의 총 7,000 글자를 레이저 프린터로 출력한 후, 400 DPI(Dot Per Inch) 해상도로 스캔(Scan)하여 획득한 문자 영상을 이용하였다.

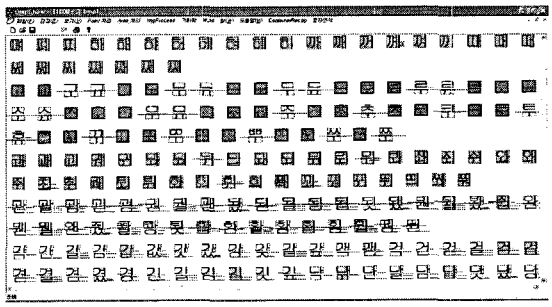

그림 8. 1,000자에 대한 유형분류 실행 예
Fig 8. Example classification of frequently used 1,000 characters

7개의 글꼴은 문서 작성시 일반적으로 많이 사용되는 고딕, 굴림, 궁서, 명조, 신명조, 바탕, 중고딕을 사용하였다. 그림 8에 한글 잦기순 상위 1,000자에 대한 유형분류 실행 예를 보였다. 유형 분류기 개발시 사용한 7개 글꼴 7,000자에 대한 분류 실험에서는 획이 가늘게 인쇄된 명조체 문

자를 제외하고 다른 6개 글꼴로 인쇄된 문자들에서는 모두 100%의 분류율을 보였다. 제안한 방법이 일반적인 인쇄문서에서도 좋은 성능을 발휘하는지 알아보기 위하여 "월간 마이크로소프트"와 같은 잡지 50장으로부터 구한 30,614자에 대한 유형분류 실험을 한 결과 30,597자에 대한 정분류를 보여 99.94%의 높은 유형 분류율을 보였다.

본 논문의 실험 결과를 기존 연구와 비교하면, 참고문헌 [13]의 논문에서 동일한 실험 데이터에 대한 유형분류 실험을 한 결과 99.90%의 유사한 인식률을 보인바 있다. 그러나 이 연구에서는 서로 다른 두 유형 간 인식 대상 문자의 수가 각각 880자와 13360자로 15.18배나 차이가 나는 경우도 있어서 유형 분류 다음 단계에서 문자 인시기에 큰 부담을 줄 수 있다는 문제가 있다. 또한 종성 없이 초성과 복모음으로 구성된 문자들을 종성 없는 종모음 문자로 오분류하는 경우가 많았다. 본 연구에서와 같은 데이터는 아니지만 참고문헌 [14]의 실험에서는 바탕체와 굴림체 글꼴에 대한 한글 잦기순 상위 1000자에 대한 신경망을 이용한 실험에서 98.15%의 분류율을 보인바 있다. 이는 신경망에 주어진 입력 특징이 문자 가장자리에 산재한 한글 유형간 특징을 충분히 나타내지 못했기 때문이라 사료된다. 따라서 본 논문에서 제안한 방법이 유형 분류율도 높을 뿐만 아니라 이전 방법들에서보다 유형 간 크기 불균형이 개선되어 문자 인식기의 부담을 들어준다는 점에서 의미가 크다.

본 논문 실험에서의 오분류 대부분은 신명조나 바탕체와 같은 글꼴에서 인쇄 상태가 불량하여 횡모음 프로파일 추출이 제대로 안되어 모음 추출에 실패한 경우였다. 그림 9에 오분류의 예를 보였다. (a)는 횡모음 영역에서 획이 제대로 인쇄되지 않아서 수평 프로파일 추출에 실패함으로써 모음 추출에 실패한 경우이다. 이러한 경우를 방지하기 위하여 흐리게 인쇄된 문자에는 횡모음 추출 전에 획을 보강하는 전처리가 필요할 것이다. (b)는 굵게 인쇄된 복자음 종성의 경우에 수평 프로파일이 과도하게 추출되어 종성 분류에 실패한 경우이다. 이렇게 두꺼운 수평 프로파일이 발생한 경우에는 수평특징 3개를 갖는 종성으로 분류하는 예외 조항을 추가하여 해결할 수 있다.

(a) (b)

그림 9. 유형 오분류의 예
Fig 9. Example of type misclassification

Ⅳ. 결론

한글이나 한자와 같은 문자집합이 큰 조합문자를 인식하기 위해서는 문자 집합 크기와 다양한 글꼴로 인한 인식기의 부담을 줄이기 위하여 문자 인식단계 이전에 문자의 유형분류를 하는 것이 유리하다. 기존 한글문자의 유형분류 연구들에서는 한글 구성원리에 따라 초성, 중성, 종성의 유무와 모음의 종류를 기준으로 한글 문자를 6가지 혹은 15가지로 분류하였다. 그러나 패턴인식기의 입장에서 볼 때 이러한 분류는 다분히 사람 중심의 분류방법으로서 같은 유형 내에서 자소의 모양, 크기, 위치가 달라 인식률을 저하시키는 원인이 될 수 있고, 종성 있는 문자가 포함된 유형에는 상대적으로 큰 문제 공간이 할당되어 유형 크기의 균형성이 문제가 된다. 본 논문에서는 수평 투영 프로파일을 이용한 횡모음 우선 추출과 종성의 세분류를 이용하여 같은 유형에 속하는 문자들에 대한 자소인식기의 부담을 들어주고 인식률을 높일 수 있는 새로운 한글 문자 유형 분류 방법을 제안하고 실험을 통하여 제안한 방법의 효율성을 보였다.

기존의 연구들이 종성의 세부 분류를 등한시 한 반면, 본 연구에서는 종성에 대한 일정크기 이상 수평 프로파일의 개수와 위치 정보, 그리고 연결요소 정보를 이용하여 종성을 세분함으로써 최종 유형의 개수를 19개로 하여 문제 공간을 보다 적절한 크기로 나눌 수 있었다. 한글 잦기순 1,000자에 대하여 7가지 폰트에 대한 개발과 월간지에서 스캔한 30,614 문자에 대한 실험에서 99.94%의 정분류율을 얻어서 제안한 방법이 글꼴에 무관한 한글문자의 유형분류에 효과적임을 확인하였다.

기존 관련 연구들이 한글 문자 생성 규칙 중심의 유형 분류를 하여 다양한 폰트에 대한 분류율이 낮고 복모음 문자 분류에 대한 어려움이 있는데 반하여 제안한 방법은 실제 패턴 중심의 유형 분류를 함으로써 분류율을 높일 수 있었다. 향후 제안한 유형 분류기를 이용하여 인식률이 높은 한글 문자 인식기를 개발할 예정이다.

참고문헌

[1] 김민수 외 4인, "인쇄체 문자 인식기의 성능 평가에 관한 연구," 한국정보처리학회논문지 제 7권, 제 11호, pp.3584-3591, 2000.

[2] 김병기, 김항준, "신경망 모델을 이용한 한글 문자의 유형 분류와 인식," 한국정보과학회 가을 학술발표 논문지, 제 16권, 제 2호, pp. 303-306, 1989.
[3] 권재욱, 조성배, 김진형, "계층적 신경망을 이용한 자중 크기의 다중활자체 한글문서 인식," 한국정보과학회 논문지, 제 19권, 제 1호, pp. 69-79, Jan. 1992.
[4] 조성배, 김진형, "인쇄체 한글 문자의 인식을 위한 계층적 신경망," 한국정보과학회 논문지, 제 17권, 제 3호, pp. 306-316, 1990.
[5] 장석진, 강선미, 김혁구, 노우식, 김덕진, "자소 인식 신경망을 이용한 한글 문자 인식에 관한 연구," 대한전자공학회 논문지, 제 31권, B편 제 1호, pp. 81-87, Jan. 1994.
[6] 김상우, 전윤호, 최종호, "복합 신경회로망을 이용한 인쇄체 한글 문자의 인식," 대한전자공학회 논문지, 제 27권, 제 4호, pp. 37-43, 1990.
[7] 김상우, 전윤호, 최종호, "신경회로망을 이용한 인쇄체 한글 문자의 인식," 대한전자공학회 논문지, 제 27권, 제 2호, pp. 228-234, Feb. 1990.
[8] 도정인, "인쇄체 한글 문자의 인식을 위한 자소 분리에 과한 연구," 한국정보과학회 가을 학술발표 논문지, 제 17권, 제 2호, pp. 175-178, 1990.
[9] 도정인, "한글 문서 인식 시스템의 개발," 정보과학회지, 제 9권, 베 1호, pp. 22-32, 1991.
[10] 김민석, 손항웅, 최완수, 김수원, "자소 추출 방법을 이용한 고속 한글 인식 시스템의 구현," 대한전자공학회 논문지, 제 29권 B편, 제 6호, pp. 416-424, 1992.
[11] 이성환, "다양한 활자체 및 크기의 한글 문자 영상에서의 정보량 및 엔트로피의 분포," 한국정보과학회논문지 제 19권, 제 2호, pp. 133-139, 1992.
[12] 박상철, 김수형 , "투영 프로파일의 간략화 방법을 이용한 인쇄체 한글 문서 영상에서의 문자 분할", 한국정보처리학회논문지B, Vol. 13-B, No. 2, pp. 89-96, 2006.
[13] 이창숙, "모음 특징을 이용한 한글 문자의 유형 분류," 신라대학교 전자계산학과 석사학위논문, 2000. 12
[14] 김우태, "신경회로망을 이용한 다중 크기 및 다중 글꼴의 한글과 영어 혼용문서 인식 시스템," 경북대학교 전자공학과 박사학위논문 1993.

저 자 소 개

김 병 기

1990년 경북대학교 전자계산기공학과 (공학석사)
1995년 경북대학교 컴퓨터공학과 (공학박사)
1995년 ~ 현재 : 신라대학교 컴퓨터정보공학부 교수