韓國精密工學會 '97年度 秋季學術大會論文集 pp.427~430

조합형 문자구성을 이용한 문서 인식 알고리즘 개발

Development of an Algorithm for Korean Letter Recognition using Letter Component Analysis

김 영 재*, 이 호 재(서울시립대 대학원), 김 희 식(서울시립대)

Abstract

This paper proposes a new image processing algorithm to recognize korean documents. It take out the region of syllable area from input character image, then it makes recognition of a consonant and a vowel in the character. A precision segmentation is very important to recognize the input character. The input image has 8-bit gray scaled resolution. Not only the shape but also vertical and horizontal lines dispersion graph are used for segmentation.

The result shows a higher accuracy of character segmentation.

Keywords : OCR, Korean Recognition, Image Processing, Segmentation

1. 서 론

정보화 시대가 시작되면서부터 각종 정보, 문서들을 데이터 베이스화하여 저장하고, 이용하게 되었다. 그러면서, 기존의 문서를 컴퓨터에 입력시킬 필요성이 발생하고, 그 일을 주로 사람이 키보드로 입력하여 왔다. 또한 최근에는 스캐너를 이용하여 문서 영상을 입력받고, 그 영상에서 문자만을 추출하여 인식하는 연구가 활발하게 진행되어 있고, 일부 상용화되어 사용되고 있다.

기존의 영문, 숫자의 인식은 우리 나라 뿐만 아니라, 외국에서도 오랜 연구에 의해 상당한 기술적 진보가 있었지만, 한글 인식에 경우, 기존의 영문, 숫자 인식 알고리즘으로는 부족한 면이 있다. 영문, 숫자는 하나의 독립된 모양체이지만 한글의 경우 초성, 중성, 종성으로 조합된 문자체계이다. 따라서, 문자 모양의 종류가 다른 언어에 비해 상당히 많다. 완성형 한글 코드를 기준 하여도 2,350자에 이른다. 실제, 조합형 한글 코드의 경우는 수 만 가지의 모양을 가진다. 기존의 영문, 숫자 인식 방법을 그대로 적용하면, 상당한 인식처리 시간을 필요로 하고, 인식률도 떨어진다. 따라서, 한글 고유의 문서 영상 처리, 분리 및 인식 방법을 필요로 한다.

본 논문에서는 초성, 중성, 종성의 분리를 정확하고 효율적으로 하기 위해 종성의 유무 판별 알고리즘과, 중성의 위치 판별 알고리즘에 관한 연구를 다루었다.

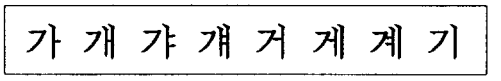
2. 한글 폰트의 구성

일반적인 문서 편집기의 한글 폰트는 한글 조합 원리에 따라 초성, 중성, 종성으로 구성된 각각의 폰트를 가지고 있고, 화면 표시 및 인쇄 과정에서 한글 코드에 의해 폰트를 조합하여 출력한다.

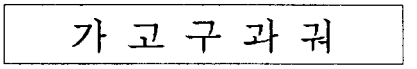
	중성의 종류	종성없음	종성있음
초성의 벌수	ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅖ ㅣ	0	5
	ㅗ ㅘ ㅙ ㅚ	1	6
	ㅜ ㅝ ㅞ ㅟ	2	6
	ㅛ ㅠ ㅡ ㅢ	3	7
	ㅕ ㅖ ㅢ	4	7
중성의 벌수	초성의 종류		
	ㄱ ㅋ	0	2
	ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ	1	3
	ㅆ ㅇ ㅈ ㅉ ㅊ ㅌ ㅍ ㅎ	1	3
종성의 벌수	중성의 종류		
	ㅏ ㅑ ㅘ	0	
	ㅓ ㅕ ㅚ ㅝ ㅟ ㅡ ㅣ	1	
	ㅐ ㅒ ㅔ ㅖ ㅙ ㅞ	2	
	ㅗ ㅘ ㅜ ㅠ ㅢ	3	

< 표 1. 한글 폰트 종류 >

폰트의 조합 과정에서 초성과 중성의 폰트는 종성의 유무에 따라 구성되는 모양이 다른 특징을 가지고 있다. 표 1에 의하면 일반적으로 초성일 경우에는 중성의 종류와 종성의 유무에 따라 8종류가 된다. 중성의 경우 초성의 종류와 종성의 유무에 따라 4 종류가 된다. 종성의 경우 중성의 종류에 따라 4 종류이다. 예를 들면, 'ㄱ'의 폰트 모양은 중성의 종류와 종성의 유무에 의해 그 종류가 다르다. 그림 1의 'ㄱ'은 서로 같은 모양의 폰트이지만, 그림 2의 'ㄱ'은 서로 다르다. 따라서, 한글 문자인식에 있어서 종성의 유무를 판별하는 일은 보다 정확한 문자인식을 위해 필요하다. 종성의 유무에 따라 문자인식 알고리듬이 다르게 적용될 수 있기 때문이다



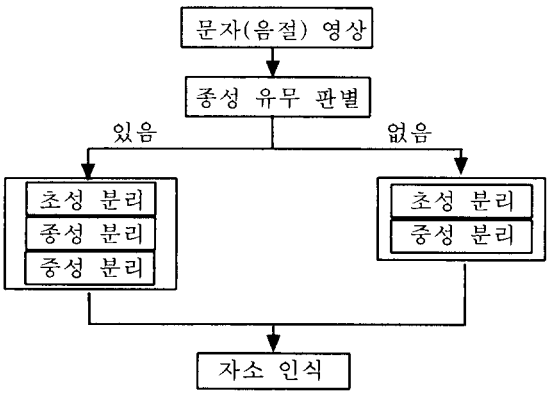
< 그림 1. 동일한 모양의 'ㄱ' 폰트 >



< 그림 2. 다른 모양의 'ㄱ' 폰트 >

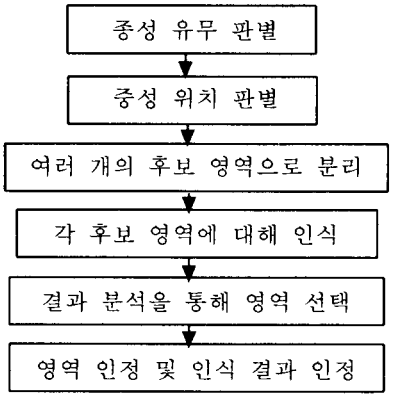
3. 한글 문자 인식의 과정

한글 문자는 자음과 모음의 조합형 문자여서 문자를 하나의 패턴으로 간주하여 인식하기는 문제점이 많다. 따라서, 본 연구에서는 그림 3과 같이 문자 인식 과정에 한글 조합 원리인 초성, 중성, 종성으로 분할하는 과정을 포함하고 있으며, 분할하는 과정에서 종성의 유무 판별은 중요한 역할을 한다. 자소(초성, 중성, 종성) 분할 이전에 종성의 유무를 판별하여 각각 다른 알고리즘을 적용하여 분할하였다.



< 그림 3. 초성, 중성, 종성의 분할 흐름도 >

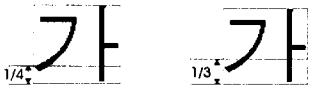
초성, 중성, 종성을 분리하는 과정에서 그림 4에서 알 수 있듯이 임의로 어떤 위치에 초성, 중성, 종성이 있다고 가정하고, 여러 종류의 영역들로 분리하여 후보 분리 영역을 설정한다. 이렇게 분리된 영역들에 대해 개별적인 인식을 수행한다. 그리고, 그 결과를 분석하여 후보 분리 영역들 중에 가장 정확하게 분리된 영역을 찾아내어, 그 영역에 의해 인식된 값을 취하였다. 이 때, 후보 영역 설정 과정에는 중성의 위치를 판별하는 알고리즘에 의해 그 후보 영역들의 수를 줄였다.



< 그림 4. 초성, 중성, 종성 분리 >

4. 종성의 유무 판별

음절에 종성이 있는 경우, 없는 경우는 서로 다른 몇 가지 특징을 가지고 있다. 일반적으로 종성이 있는 경우는 없는 경우에 비해 문자 영상의 하반부 모양이 복잡하여, 수직 성분 분포와 수평 성분 분포가 다른 특징을 가지고 있다. 여기서, 모양 판단 영역과 문자 영상의 수직 성분과 수평 성분의 추출 대상 영역은 종성 인식에 가장 효과적이라고 판단되는 부분을 선정하였다. 모양 판단의 경우 그림 5와 같이 문자 영상 높이의 1/4 부분을 대상으로 하였고, 수직 성분의 경우는 문서 전체 영상을 수평 성분의 경우는 그림 5와 같이 문서 영상 높이의 1/3 부분을 선정하였다.



< 그림 5. 모양 판단 대상 영역과 수평 성분 추출 영역 >

위의 영역에서 얻어진 각각의 특징들을 살펴보면 다음과 같다.

(1) 'ㅗ', 'ㅜ' 의 경우

종성이 없는 경우 표 2 와 같은 특징을 가진다.

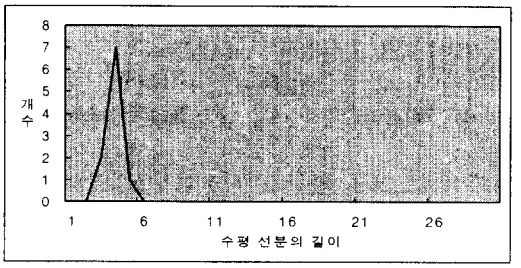
종류	모양	수평 성분	수직 성분
'ㅗ'	-	문자 폭 길이의 선이 영상 하단에 있다.	-
'ㅜ'	모양 판단 영역의 좌우에 공간이 있다.	-	-

< 표 2. 종성의 유무에 대한 문자 영상의 특징 >

(2) 'ㅏ', 'ㅓ', 'ㅣ' 의 경우
종성이 없는 경우 표 3과 같은 특징을 가진다.

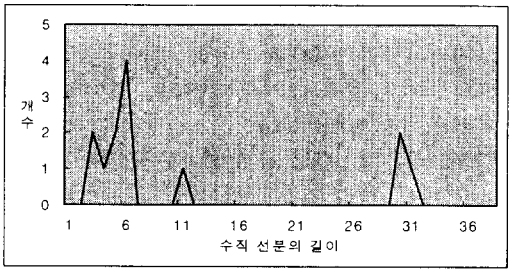
종류	모양	수평 성분	수직 성분
'ㅏ', 'ㅓ', 'ㅣ'	-	긴 선분이 거의 없다	문자 오른쪽에 문자 높이에 가까운 선이 있다.

< 표 3. 종성의 유무에 대한 문자 영상의 특징 >



< 그림 6. 문자 '지'의 수평 성분 >

예를 들어 문자 '지'의 수평 성분의 그래프를 보면 그림 6과 같이 긴 선분은 없고 짧은 선분만 있음을 알 수 있다.



< 그림 7. 문자 '지'의 수직 성분 >

수직 성분의 경우 그림 7과 같이 문자 높이에 가까운 선분이 있음을 알 수 있다.

(3) 중성이 있는 경우
중성이 있는 경우는 대체적으로 모양이 복잡하여 표 4와 같은 특징을 가진다.

종류	모양	수평 성분	수직 성분
'ㅁ', 'ㅂ', 'ㅇ', 'ㅍ', 'ㅎ'	문자 하단에 폐곡선이 존재한다.	중간 길이 정도의 선분이 많다.	중간 길이 정도의 선분이 많다.
그 외의 자음	-	중간 길이 정도의 선분이 많다.	중간 길이 정도의 선분이 많다.

< 표 4. 종성의 유무에 대한 문자 영상의 특징 >

5. 중성의 위치 판별

중성의 위치는 표 5와 같이 크게 '가' 와 같이 초성의 오른쪽에 있는 경우, '고' 와 같이 초성 아래에 있는 경우, '과' 와 같이 복모음으로써 오른쪽과 아래에 있는 경우의 3 가지가 있다.

기본 종류	종성(받침)이 있는 경우	위치
'가'	'각'	오른쪽
'고'	'곡'	아랫쪽
'과'	'곽'	오른쪽, 아랫쪽

< 표 5. 중성의 위치 >

따라서, 중성의 위치를 판별 할 수 있으면 초성, 중성, 종성 분리를 위한 영역 설정에 도움을 줄 수 있다.

본 연구에서는 문자 영상의 수평 성분 분석과 수직 성분 분석을 통해 종성 위치 판별하였다. 수평 성분 분석은 그림 8처럼 문자 영상의 2/3 영역을 대상으로 하여 수평한 선분의 개수 분포를 조사하였다. 중성이 초성의 아래에 있는 경우 수평으로 긴 성분의 선들이 있다.

< 그림 8. 수평 성분 분석 영역>

수직 성분 분석은 종성 유무 판별 때와 비슷하지만 대상 영역이 그림 9처럼 문자 영상의 1/2 만을 대상으로 하여 연산 시간을 줄이고, 초성 부분의 수직 성분에 의한 자료의 오류를 줄였다.

< 그림 9. 수직 성분 분석 영역 >

수평 성분 분석에 의해 수평 성분 중 문자 폭의 80 % 이상 되는 선이 있을 경우 중성은 초성의 아래에 있다고 인정하였다.

수직 성분 분석에 의해 수직 성분 중 문자 높이의 70 %이상 되는 선이 있는 경우 중성은 초성의 오른쪽에 있다고 인정하였다.

수평 성분과 수직 성분 분석에 의해 수평 성분 중 70%이상 되는 선과 수직 성분 중 70%이상 되는 선이 있을 경우 중성은 오른쪽과 아래에 있는 복모음이라고 인정하였다.

6. 실험 결과

(1) 종성 유무 판별 알고리즘의 적용 결과

문서 영상 종류	대상 개수	오류 개수	인식률
TEST1	1000	53	95%
TEST2	1000	61	94%
TEST3	1000	47	95%
합계	3000	161	95%

< 표 6. 종성 유무 인식 알고리즘 적용 결과 >

종성 유무 판별 알고리즘의 경우 실제 문자 영상 중에는 복잡한 복모음과 단순한 종성의 경우 분별하기가 힘들다. 실제로 문자 '식' 과 같은 문자 '쇠' 보다 단순하지만 종성을 포함하고 있다. 따라서, 이러한 경우 오류가 발생하였다.

(2) 중성 위치 판별 알고리즘의 적용 결과

문서 영상 종류	대상 개수	오류 개수	인식률
TEST1	500	9	98%
TEST2	500	12	97%
TEST3	500	7	98%
합계	1500	28	98%

< 표 7. 중성 위치 판별 알고리즘 적용 결과 >

중성의 위치 판별 알고리즘의 실제로 적용 결과, 문자 폭 길이의 수평한 선분을 찾는 과정에서의 오류 발생이 적어서 수직 선분을 찾는 알고리즘 적용 전에 중성 위치를 판별 할 수 있는 경우가 많았다. 따라서, 높은 인식률을 얻을 수 있었다.

7. 결 론

본 논문에서는 조합형 문자 분석을 이용하여 한글 문자 인식에 관한 연구를 하였다.

기존의 문자 인식은 패턴인식을 응용한 비선형 정합법이나 통계적 확률을 이용한 은닉 마르코프 모델 또는 신경회로망을 이용한 것이 대부분을 차지하였다. 하지만, 본 논문에서는 구조적 특징을 이용하여 문자의 모양 판단과 수평, 수직 성분을 측정하여 한글 고유의 문자 특징을 인식하는 알고리즘에 대해 연구하였고, 앞으로, 실제적인 한글 문자의 인식을 위해 초성, 중성, 종성의 분리 알고리즘을 보완하고, 분리된 자소 (자음, 모음)의 인식을 위한 알고리즘 개발이 이루어져야겠다.

8. 참고 문헌

[1] Craig A.Lindly 지음, 류성렬 옮김,"C 이미지 프로세싱", 1991 , 동일출판사

[2] Edward R. Dougherty, "Digital Image Processing Methods", 1994, Marcel Dekker pp. 43-102

[3] NHK 방송 기술 연구소 화상연구소, 국제 테크노정보 연구소 편역," C 언어에 의한 화상처리 실무", 1995, 국제 테크노정보 연구소

[4] Robert J.Schalkoff, "Digital Image Processing and Computer Vision", 1989, Jonh Wiley & Sons pp. 130-178

[5] 김두식, 이성환, "한글과 영.숫자가 혼용된 문서를 위한 효과적인 문자 분할방법", 1996.1, 8회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp. 19-26

[6] 김두식, 김상엽, 이성환, "한글 문서 분석 및 인식 기술의 최근 연구 동향", 1997.9, 전자공학회지 제24권 제9호 pp. 1058-1067

[7] 김희식, 최승영, 김영재, 박준호, "가스미터기 성능검사장치에서 계기판 숫자의 영상인식 시스템 개발 연구", 1994.12, 서울 시립대학교 산업기술연구소 논문집 2집 pp.113-126

[8] 남궁재찬, "화상공학의 기초", 1989, 기전연구사

[9] 이 성환, "문자인식 이론과 실제", 1994.3, 홍릉과학 출판사, 1권

[10] 일본공업기술센터편, "컴퓨터화상처리 입문", 1993, 기전연구사