셀들의 군집 정보를 이용한 한글 문자 인식률 향상 기법 연구*

신우준, 고윤식, 임영택, 윤영수, 박희완
한라대학교 정보통신방송공학부
email:cominbooks@naver.com, dbstlr1819@naver.com,
ekzmsdkdl12@naver.com, salang1282@hanmail.net, heewanpark@halla.ac.kr

Improving Korean Character Recognition Rate based on the Cell Clustering Information

Woojun Shin, Yoonsik Ko, Youngtaek Lim, Youngsu Yoon, Heewan Park
School of Info. & Comm., Broadcasting Engineering, Halla University.

요 약

문자인식 즉 OCR(Optical Character Recognition)기술은 광학적으로 인식할 수 있는 문자를 컴퓨터가 읽을 수 있도록 하는 기술을 뜻한다. 문자인식의 근간이 되는 방법은 스트링 매칭 기법이 사용되어 왔지만 한글의 경우 자음, 모음, 자음 조합으로 만 가지 유형이 넘고, 더욱이 상용한자와 영어를 섞어 쓰기 때문에 오인식되는 경우가 많다. 본 논문에서는 한글이 수직선, 수평선, 사선과 같이 방향성이 강한 선소들로 구성되어 있다는 점을 이용하여 한글의 인식률을 높이는 방법을 제안하였다.

1. 서론

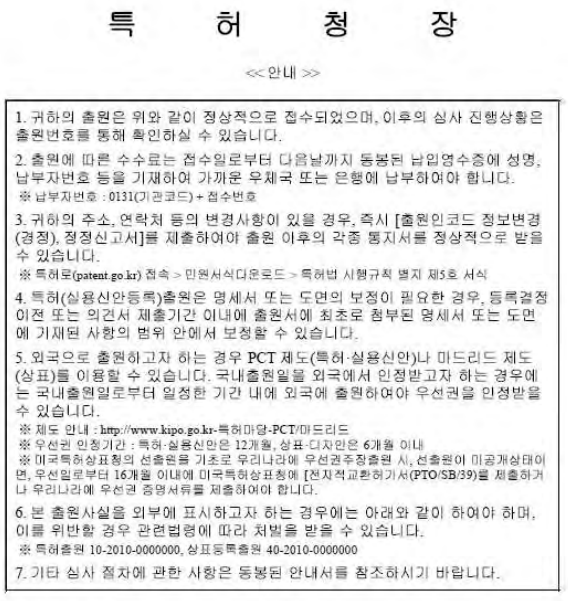
문자인식 즉 OCR(Optical Character Recognition)기술은 말 그대로 광학적으로 인식할 수 있는 문자를 컴퓨터가 읽을 수 있도록 하는 기술[1.2]을 뜻한다. 컴퓨터가 책이나 신문, 문서를 읽는 것이다.

문자인식 기술은 일상생활에서 쉽게 접할 수 있다. 예를 들어 책이나 신문에서 마음에 드는 글을 휴대폰으로 사진을 찍거나, 문서를 스캔한 후 생성된 이미지 파일로부터 텍스트를 추출해 내는 기술이다.

(그림 1)과 같은 출력된 문서를 스캔하여 이미지 파일로 저장할 수 있다. 그러나 이미지 파일로 문서를 저장할 경우에는 텍스트 정보가 없기 때문에 키워드 검색을 할 수 없으며, 용량이 커지는 문제가 있다. 즉, 이미지로 스캔한 문서는 보관된 데이터로서의 가치는 있으나 검색 가능한 정보로서의 활용 가치는 거의 없다.

그러나 이미지 안에 있는 문자들을 모두 텍스트로 바꿔주는 OCR 기술을 사용하면 원하는 내용을 검색할 수 있으며, 필요에 따른 단어나 문장들을 복사해서 유용하게 편집하고 활용할 수 있게 된다. 요약집을 따로 만들고 싶은 학생이나, 자료를 수정하거나 취합하는 사무직원, 논문을 쓰고 강의하는 교수 등 많은 개개인이 편익을 얻을 수 있다. 이것이 텍스트 중요성이고, 문자인식 기술의 필요성이다.

* 이 논문은 2014년 중소기업청에서 시행한 산학연협력 첫걸음기술개발사업의 결과물임.



특 허 청 장

<< 안내 >>

1. 귀하의 출원은 위와 같이 정상적으로 접수되었으며, 이후의 심사 진행상황은 출원번호를 통해 확인하실 수 있습니다.
2. 출원에 따른 수수료는 접수일로부터 다음날까지 동봉된 납입영수증에 성명, 납부자번호 등을 기재하여 가까운 우체국 또는 은행에 납부하여야 합니다.
※ 납부자번호 : 0131(기관코드) + 접수번호
3. 귀하의 주소, 연락처 등의 변경사항이 있을 경우, 즉시 [출원인코드 정보변경(경정), 정정신고서]를 제출하여야 출원 이후의 각종 통지서를 정상적으로 받을 수 있습니다.
※ 특허로(patent.go.kr) 접속 > 민원서식다운로드 > 특허법 시행규칙 별지 제5호 서식
4. 특허(실용신안등록)출원은 명세서 또는 도면의 보정이 필요한 경우, 등록결정 이전 또는 의견서 제출기간 이내에 출원서에 최초로 첨부된 명세서 또는 도면에 기재된 사항의 범위 안에서 보정할 수 있습니다.
5. 외국으로 출원하고자 하는 경우 PCT 제도(특허·실용신안)나 마드리드 제도(상표)를 이용할 수 있습니다. 국내출원일을 외국에서 인정받고자 하는 경우에는 국내출원일로부터 일정한 기간 내에 외국에 출원하여야 우선권을 인정받을 수 있습니다.
※ 제도 안내 : http://www.kipo.go.kr-특허마당-PCT/마드리드
※ 우선권 인정기간 : 특허·실용신안은 12개월, 상표·디자인은 6개월 이내
※ 미국특허상표청의 선출원을 기초로 우리나라에 우선권주장출원 시, 선출원이 미공개상태이면, 우선일로부터 16개월 이내에 미국특허상표청에 [전자적교환허가서(PTO/SB/39)]를 제출하거나 우리나라에 우선권 증명서류를 제출하여야 합니다.
6. 본 출원사실을 외부에 표시하고자 하는 경우에는 아래와 같이 하여야 하며, 이를 위반할 경우 관련법령에 따라 처벌을 받을 수 있습니다.
※ 특허출원 10-2010-0000000, 상표등록출원 40-2010-0000000
7. 기타 심사 절차에 관한 사항은 동봉된 안내서를 참조하시기 바랍니다.

(그림 1) 문서를 스캔한 이미지 파일

2. 관련연구

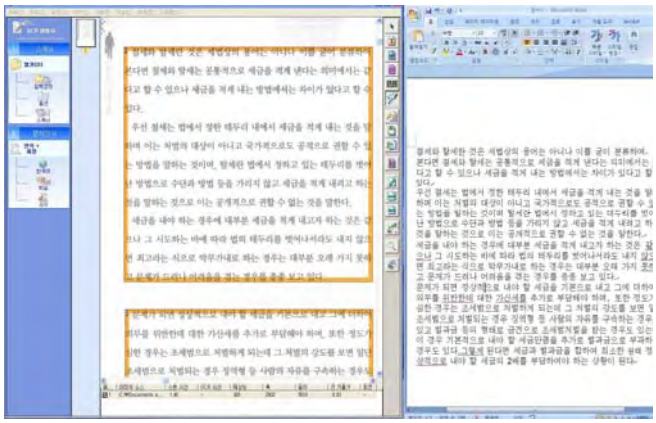
문자 인식 기술[3,4,5]은 우리 생활과 밀접한 분야에서 발전해왔다. (그림 2)는 직장인들이 자주 사용하는 명함을 인식하기 위해서 문자 인식이 적용된 명함 인식기이다.

직장인들의 경우 사회 활동을 하면서 여러 사람을 만나게 되고 명함을 주고받게 된다. 그런데 명함이 수십장, 수백 장이 넘어가게 되면 명함집을 사용하더라도 보관에 문제가 생길 수 있고, 매번 명함을 받을 때마다 집에 돌아

와 파일로 정리해놓지 않는다면 찾고 싶은 가게나 연락처를 바로 알고 싶을 때 명함집 안의 수많은 명함을 일일이 다 찾아봐야하는 문제가 생길 수 있다. 이런 불편함을 극복하기 위해 명함 인식기의 필요성이 대두되었으며, 지금은 스마트폰에서 카메라로 찍으면 바로 연락처로 저장이 되도록 발전되었다.

(그림 2) 명함 인식을 위해서 사용되고 있는 문자 인식 기술

문자인식 기술이 활용되는 대부분의 경우는 서론에서 언급한 바와 같이 휴대폰 카메라나 스캐너를 통해서 이미지 파일을 만든 후, 전문 문자인식 프로그램을 사용해서 해당 이미지 파일을 문자로 인식하는 것이다. (그림 3)과 같이 이미지 파일(jpg)을 불러와서 문자 인식 기능을 실행시키면 이미지 파일에 포함된 그림과 문자를 구별하여 인식하고 워드(.doc) 파일이나 PDF 파일, 엑셀(.xls) 파일, 또는 텍스트(.txt) 파일 등 여러 포맷으로 저장 가능하다.



(그림 3) 문자인식 프로그램

그러나 일반 개인의 경우 스캐너 장비가 없다면 스캔을 하기 쉽지 않고 스캔을 하더라도 문서 전체를 해야 한다. 사진을 찍을 경우 스캔한 결과물에 비해 인식률이 떨어질 수 있고 휴대폰의 사진을 PC로 옮겨야 하는 번거로운 절차가 있다.

이러한 번거로움을 줄이고자 (그림 4)와 같이 '펜의 형태'의 도구에 문자인식 기술을 적용하여 문서에 밑줄을 그어 필요한 내용만 파일로 만들도록 간결함과 편리성 방향으로 개발된 사례가 있다.

향후에는 펜을 넘어 안경에 문자 인식 기술이 접목되어 눈으로 보는 글씨를 문서로 변환하여 저장하는 것도 가능해질 것으로 기대된다.

(그림 4) 문자 인식 기능이 내장된 펜

그 외에 주차장에 들어오는 차의 번호판을 인식하여 무인 단속하는 용도 등으로도 활용되고 있다.

편리함도 중요하지만, 문자인식에 있어서 가장 중요한 것은 바로 인식률이다. 같은 내용이라도 오래된 문서보다 갓 출력한 문서를 스캔한 파일에서 인식률이 높다.

인식률에 영향을 미치는 대표적인 요인은 스캔 상태이다. 즉 문자 상태가 양호하도록 밝기, 명암, 해상도 등이 적절해야 하며, 문자가 구부러져 있거나 기울기가 있으면 인식률이 현저히 낮아진다[3].

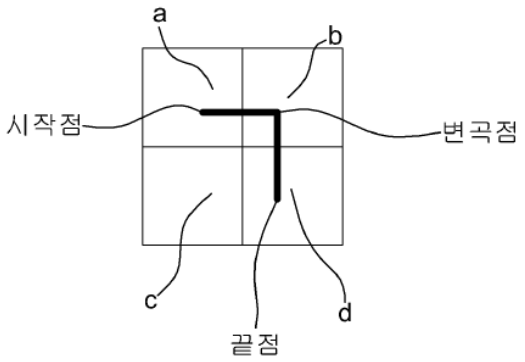
문자인식의 근간이 되는 방법은 스트링 매칭 기법으로, '가장 비슷하게 생긴 문자와 연결시키는 것'이다[4,5]. 영어의 경우 a부터 z까지 26유형의 패턴과 숫자의 경우 0부터 9까지 10가지 유형의 패턴이 있다. 패턴이 많지 않기 때문에 이들에 대한 인식률은 거의 100%다. 하지만 한글의 경우 자음, 모음, 자음 조합으로 만 가지 유형이 넘고, 더욱이 상용한자와 영어를 섞어 쓰기 때문에 오인식되는 경우가 많다. 여기에 서체라는 변수까지 생각해보면 한글 인식이 얼마나 어려운지 쉽게 알 수 있다. 2000년도에는 아르미, 글눈이 등 국내 문자인식 업체들이 활기를 띄었지만 2010년도에 와서는 국내 문자인식 회사들은 거의 사라진 상태다.

3. 셀들의 군집 정보를 이용한 한글 인식률 향상 기법

한글 인식률을 높이기 위해서는 기존의 기법에서 탈피한 새로운 개념의 기법이 필요하다. 그중 하나는 수학적 이론을 바탕으로 한 기법이다. 수직선, 수평선, 사선과 같이 방향성이 강한 선소들로 구성되어 있다는 점을 이용하여 한글을 인식보다는 구분하도록 하는 것이다.

이 방식은 기존의 문자인식 방식이 문서를 스캔해서 획득한 각 문자들이 구성하는 검은 색 셀들의 집합이 라이브러리에 저장된 다양한 문자형태의 집합과 비교하여 그 중 어느 문자와 가장 일치하는가를 판단하여 제시하는

스트링 매칭 기법을 이용하는 것과는 달리 획득한 셀들의 집합의 시작 군집과 변곡점, 그리고 끝 군집의 위치 좌표를 파악하여 문자를 인식하는 방식이다.



(그림 5) 'ㄱ'에 대한 셀들의 군집 정보

(그림 5)와 같이, 'ㄱ'의 경우, 획득한 'ㄱ'의 셀들의 집합을 4등분하고, 4등분한 좌상 구역을 1구역, 우상 구역을 2구역, 좌하 구역을 3구역, 우하 구역을 4구역으로 정하여 제 1 구역에 시작점이 있고, 제 2 구역에 변곡점이 있으며, 제 4 구역에 끝점이 있으면 'ㄱ'으로 인식한다. (그림 5)에서 a는 1구역, b는 2구역, c는 3구역, d는 4구역이다.

1구역에서 획득한 셀 집합 중 좌측에 위치한 지정된 오차범위 내의 수평 위치 정보를 갖고 있는 일부 셀 집합의 위치정보들을 시작군집으로 획득하고, 변곡점에 위치한 셀 집합 중 동일한 수평의 위치정보를 갖고 있는 우측 셀 집합과 정의된 구역 이내의 수직 위치정보를 갖고 있는 일부 셀 집합의 위치정보를 갖고 있는 하측 셀 집합들을 획득하고, 마지막으로 4구역에서 획득한 셀 집합 중 하단에 위치한 일부 셀 집합 중 정의된 구역 이내의 수직 위치 정보를 갖고 있는 선의 끝 부분에 위치한 셀 집합을 끝 군집으로 획득한다.

이를 x축과 y축을 이용한 함수관계로 표현하면, 1구역에 위치한 셀들의 위치 좌표 중, x=a, y=b, 이때 a는 최소값을 갖는 정의된 범위 내에 위치한 일부 셀들의 군집이고, b는 정의된 범위 이내의 동일한 y축 좌표를 갖고 있는 일부 셀들의 군집이며 이 셀들의 군집이 시작점이다. 2구역에 위치한 셀들의 위치 좌표 중, x=a, y=b, 이때 a는 최대값을 갖고 있는 정의된 범위 내의 셀들의 군집이며 b는 정의된 범위 이내의 동일한 y축 좌표를 갖고 있는 일부 셀들의 군집과 x=a,y=b, 이때 a는 정의된 범위 내에서 동일하며, b는 감소하는 값을 갖는 일부 셀들의 군집으로 변곡점으로 획득한다. 마지막으로 x=a, y=b, 이때 a는 정의된 범위 내에서 동일한 값을 갖고 있고, b는 최소값을 갖는 일부 셀들의 군집으로 끝점으로 획득한다. 이 세 지점을 1구역부터 순서대로 2구역, 3구역으로 연결하면 'ㄱ'이라는 글자로 인식하게 되는 것이다. 이러한 방식으로 'ㄴ'은 1구역의 시작점을 획득하고, 3구역의 변곡점을 획득하고, 4구역의 끝점을 획득하여 1, 3, 4구역의 점들을 연결하면 'ㄴ'이라는 글자로 인식한다. 나머지 자음들에게도 비슷하게 적용된다. 마찬가지로 1구역의 시작점, 3구역의 끝점, 1구역 또는 3구역의 변곡점, 2구역 또는 4구역에 끝점이 있으면 'ㅏ'로 인식할 수 있고, 나머지 모음에도 비슷하게 적용된다.

이러한 새로운 차원의 인식 방법으로 활자체 인식에 있어 거의 완벽에 가까운 인식률을 낼 수 있을 것으로 기대한다.

4. 결론

문자 인식 기술은 인식 기술에 있어 근간이 되는 기술이다. 문자 인식 응용 기술에는 우편물 자동 분류 및 자동 순로 구분기, 배기가스 과다 배출차량 자동 단속기, 버스 전용선 위반차량 자동 단속기, 과속차량 자동 단속기, 도난범죄차량 자동 검거기, 도로 교통세 무인 징수기, 금융기관 전표 자동 인식기, 생산라인 불량품 자동 분열기 등이 있다.

또한 머지않은 미래에는 각 도서관, 가정 마다 안내형 로봇이 있을 것이고, 책을 포함한 모든 정보 데이터가 그 로봇에 내재되어 있을 것이다. 기본적으로 문자인식기술은 음성인식기술이나 TTS(Text to Speech)기술과 결합하여 사람의 "뱀에게 물렸을 시 응급조치는?" 등의 음성을 인식하고 문자를 검색하여 대답을 해주는 시대가 열릴 것이다.

참고문헌

[1] I. D. Lee, "Character Recognition Technological", Korea Information Processing Society, Information Processing Society journal, vol.6, no.4, 1999.

[2] G. R. Lee, H. S. Jeong, M. W. Kim, "A Study on Character Recognition", Electronics and Telecommunications Research Institute, [ETRI]trend analysis of electronic communication, 1989.

[3] J. S. Kim, "A Computer Reads books-OCR", Printing Korea, vol.4, pp.156-157, 2007.

[4] J. K. Lee, "A Method for the Recognition of Printed Korean Characters", J.KIEE, vol.7, no.4, 1969.

[5] K. H. Lee, "A Case Study on Korean Character Recognition", Korean Institute of Information Scientists and Engineers, Journal of Computing Science and Engineering, vol.9, no.1, 1991.