자동입력으로 정보화 촉매역할하는 문자인식시스템

정보의 양은 날로 늘어 나고 있고 이를 효율적으로 관리하고 이용하기 위해서 다방면에 걸쳐 컴퓨터와 통신을 활용하고 있다. "구술이 서말이라도 꿰어야 보배"란 말이 있듯이 여기에는 반드시 정보의 입력이 선행되어야 한다. 〈편집자주〉

현 은 정
합산컴퓨터(주) 대표이사

기계적으로 손쉽게 처리

문자는 인간이 언어로 표현한 정보를 전달하기 위해 발명한 대표적인 부호 체계이다. 역사 초기에는 개개의 필기자가 문자를 기록하였으나 다량의 정보를 보다 효과적으로 기록하는 인쇄기술의 발달로 인쇄체 문자로 정보를 표현하게 되었다.

현대의 정보화사회에 접어 들면서 문자, 음성, 화상 등의 형태로 각종 매체를 통해 수없이 많은 정보를 접하게 된다. 그중의 절반 이상은 문자로 된 정보로 신문, 도서, 잡지, 보고서, 논문 등의 형태로 우리에게 다가온다.

이러한 정보의 양은 날로 늘어 나고 있고 이를 효율적으로 관리하고 이용하기 위해서 다방면에 걸쳐 컴퓨터와 통신을 활용하고 있다. "구술이 서말이라도 꿰어야 보배"란 말이 있듯이 여기에는 반드시 정보의 입력이 선행되어야 한다.

그러나 수집된 수많은 정보를 컴퓨터에 입력하기 위해 사람이 일일이 타이핑하는 경우 많은 시간과 노력이 요구된다.

이러한 문제를 해결하기 위해 기계적으로 신속, 정확하게 컴퓨터에 입력하기 위한 수많은 노력이 기울어져 왔는데 이것이 바로 문자인식시스템, 즉 OCR (Optical Character Recognition)이다.

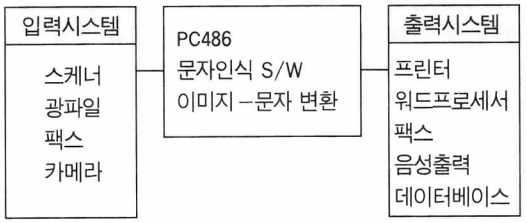
처리과정은 3단계

구성체계

문자인식 소프트웨어는 문서를 〈그림1〉과 같이 스캐너 등을 통해 입력해 이미지로 저장하고, 인식과정

을 거친 후 인식된 결과를 문자로 출력하는 시스템을 구성한다.

〈그림1〉 문자인식 시스템 구성체계



기본원리

문자인식은 〈그림2〉와 같이 3단계의 처리과정을 거쳐 수행된다.

○ 입력부문

먼저 처리하고자 하는 문서정보가 스캐너나 카메라 같은 이미지 입력장치를 통해 여러가지 형식, 즉 PCX, TIFF, BMP, GIF 등의 이미지로 컴퓨터에 보관된다.

○ 처리부문

문자인식 시스템은 입력된 이미지 데이터로부터 그림과 같은 비문자 영역과 문자 및 도표(선분과 글자) 등으로 이루어진 문자영역으로 분리한다.

비문자 영역은 그림의 형태로 별도 저장하여 차후 활용할 수 있으며, 문자영역은 개별문자나 자소로 분리한 다음에 문자 인식에 필요한 여러가지 정보를 추출해 낸다.

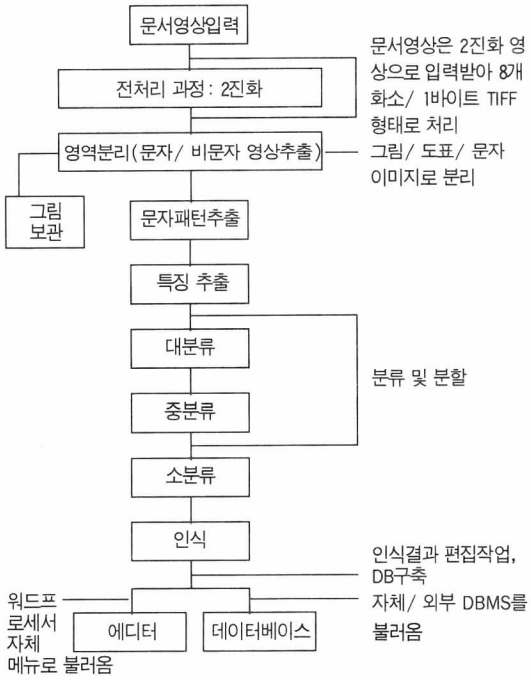
다음으로 추출된 정보를 이용하여 컴퓨터가 이해할 수 있는 문자코드로 변환하는 문자인식과정을 거친다. 여기서 인식된 결과는 이미지의 형태가 아닌 아스키 코드이다.

○ 출력부문

컴퓨터에 사용되는 아스키 코드로 변환된 정보는 문자인식 시스템에 내장된 자체편집기나 일반 워드프로세서를 통해 수정·편집할 수 있고 데이터베이스에 저장하거나 프린터 및 팩스 등을 통해 출력할 수 있는 정보가 된다.

문자인식의 원리와 과정은 간단하지만 사실상 문자

〈그림2〉 문서인식 처리과정



인식 시스템을 이루는 각 과정에 필요한 알고리즘은 상당히 복잡하고 미묘하다.

특히 한글은 구조적으로 다른 외국어에 비해 글자의 모호성(예:"의"와 익, "꺽"과 "객" 등)등 특이한 점이 많아 인식과정에 있어서 상당한 어려움이 따른다.

또한 한글은 세계에서 유래가 없을 정도로 독창적인 구조를 가지고 있으며, 자소의 결합으로 무궁무진한 변형을 만들어 새문자를 구성해 내기 때문에 적합한 인식시스템을 개발하는 데에 곤란한 점이 많을 뿐만 아니라 미국이나 일본 등의 선진기술을 도입하는 것도 거의 불가능하다.

온라인과 오프라인으로 분류

문자인식은 문자이미지 정보를 얻는 방식에 따라 〈그림3〉과 같이 온라인 인식과 오프라인(Off-Line) 인식으로 구분된다.

온라인 인식

타블릿(Tablet)이나 디지타이저 위에 사람이 전자펜을 이용하여 글자를 필기하면 즉시 그 결과를 인식해 활용하는 방식이다.

이 경우 입력장치로써 키보드가 필요하지 않기 때문에 문서결재나 펜 컴퓨터 및 초소형화를 요구하는 산업현장에서 주로 사용되고 있다.

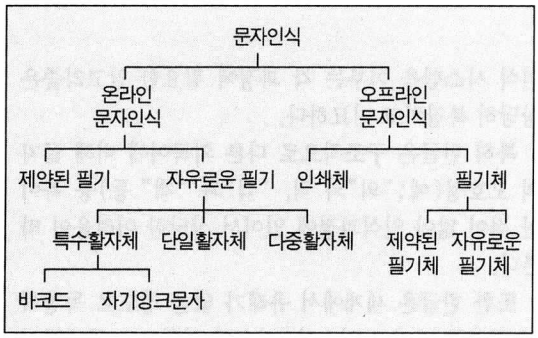
오프라인 인식

이미 작성된 인쇄체 문자나 문서화된 필기체 문자를 인식하는 방식이다.

일반적으로 문서인식 시스템은 오프라인 문자인식을 말하며 필기체 문자인식은 전세계적으로도 아직은 개발단계에 머물고 있다.

오프라인 인식은 온라인 인식과는 달리 시간적인 정보를 활용할 수는 없으나 다량의 인쇄 정보를 신속하게 처리할 수 있다.

〈그림3〉 문자인식의 분류



다양한 활용분야

기본적으로 문자인식 시스템은 〈표1〉과 같이 데이타의 입력이 필요한 모든 분야나 업종에 사용된다. 특히 다량의 자료입력이 요구되는 사무자동화, 데이터베이스 및 CD-ROM 구축, 광파일 시스템 구축, 전자출판 등에 유용하며 번역시스템, 시각장애자용 점자 및 음성출력, 각종 자료의 스크랩 등에도 활용할 수 있으며 화상정보의 문자정보 변환, 문자정보의 음성정보변환등 데이타의 멀티미디어화에도 활용한다.

한편으로는 실생활과도 밀접한 관계를 갖고 있다. 예를 들어 시각장애자들은 점자 프린터 또는 문자/음성 변환 프로그램만 갖추고 있다면 독서를 하기 위해 점자로 된 서적을 힘들게 구입할 필요가 없다. 문자인식 소프트웨어로 문서파일을 만든 다음 점자 프린트로 곧바로 출력하거나 음성으로 출력할 수 있기 때문이다.

더욱이 최근 시사정보나 전문적인 내용을 수록하고 있는 점자서적은 찾아 보기힘들기 때문에 많은 시간과 비용을 들인다 하더라도 구할 수 없다. 만일 이때 문자데이터를 음성으로 변환해 주는 프로그램과 스캔장비를 갖춘 컴퓨터가 있다면 문제는 쉽게 해결된다. 결과적으로 점자로 된 서적을 손으로 집어 가는 것보다 훨씬 효율적인 방법으로 독서할 수 있는 것이다.

또 다른 예로 보고서나 논문 작성시에 필요한 부분을 복사할 필요없이 문자인식 소프트웨어가 내장된 노트북 컴퓨터와 조그마한 핸디 스케너가 있다면 자신이 원하는 부분 또는 전체를 스캔한 후 노트북의 하드 디스크에 저장하면 모든 작업이 끝나게 된다.

〈표1〉 문자인식 시스템의 적용분야

관련업무 / 대상업체	업 무 내 용
일반기업체	인사관리, 경리, 판매관리, 수주전표, 입출고전표, 영업일지, 작업일지 등의 처리업무
금융업체	보험 가입신청서, 고객관리, 각종전표, 카드 등의 처리업무
공공기관	각종 영수증, 근태관리, 검침표, 전표, 양식관리 등의 데이터베이스 구축
출판업체	DTP, 자료편집 등 자동입력시스템
유통서비스	주문서, 카드신청서, 고객관리, 배달전표, 건강진단서, 각종 조사표, 경리전표 등의 처리업무
기타	문서 자동입력 시스템, 한글 전산화 작업, 설문조사 등 데이터 전산화를 필요로 하는 모든 분야

한글시스템도 곧 출시전망

문자인식 개발 현황

미국이나 일본 등의 선진국에서는 수십년 전부터 문자인식을 연구해 왔고 현재는 거의 100%에 가까운 인식율을 보이며 폭 넓게 상용화되고 있다.

미국은 연구기관과 학술기관에서 지속적인 연구를 진행해 오고 있고 기술적으로 특별한 훈련과정을 거치지 않고서도 거의 모든 텍스트를 효율적으로 인식할 수 있는 프로그램이 개발되어 있으며, 자동급지 장치가 부착된 스캐너의 보급이 확대되고 있다.

또한 사용하는 문자의 단순성에 힘입어 인식율 100%에 가까운 완벽한 성능을 자랑하고 있다. 일본의 경우 정부주도하에 PIPS라 불리는 프로젝트를 수립하고 산학협동으로 체계적인 연구가 이루어졌다.

일본의 문자는 비교적 복잡한 구조를 갖고 있어 우리나라 환경과 유사점이 많다. 문자인식 범위에 히라가나, 카타카나, 한문 등의 문자를 모두 포함시켜야 하므로 난이도가 높으며 약 99.5%의 인식율을 보이고 있다.

국내에서는 일부 대학과 연구소를 중심으로 연구가 추진되어 왔으나 프로그램 개발과 상용화가 미흡하며 94년부터 3~4개의 중소기업에서 제품을 출시하여 본격적으로 시장이 형성되고 있다.

최근 문자인식의 중요성이 부각되면서 관련 대기업 등에서도 개발이 활발이 진행되고 있고 많은 제품이 출시될 전망이다.

개발과제

첫째, 지속적인 연구와 투자가 선행되어야 한다.

문자인식 소프트웨어는 기본기술을 갖춘 후에도 약 3~5년간의 지속적인 개발기간과 자금이 소요되고 실패의 가능성도 상당히 높으며 기술력과 언어구문 해석력을 갖춘 특정의 전문인력이 요구된다. 이러한 점에서 중소기업형 산업이라 할 수 있으며 정부의 보다 많은 지원이 필요하다.

둘째, 기술력이 확보되어야 한다. 문자인식의 성패는 인식율과 처리속도에 달려 있는데 이것은 사람이 타이핑하는 것보다 훨씬 높은 효율성을 의미한다.

현재 국내 제품들의 인식율은 제품에 따라 90~98%(100글자당 10~20자의 오인식율)의 심한 편차를 보이는데 최소 99%이상이 되어야 하며 우리의 문서 특성상 통계나 도표도 인식되어야 한다.

처리속도는 문서의 스캔에서 해석·인식·저장 등 전체과정에서 소요되는 시간으로 현재 1분당 500~1,000글자를 처리하며 연속으로 1,024장까지 자동처리하는 제품도 있는데 처리과정의 자동화와 인식속도의 향상이 주요 개발과제로 부각된다.

정보화 촉매역할 담당

문자인식은 문자입력의 혁신을 초래할 것이며 워드프로세서와 같이 보편화되면서 정보화의 촉매역할을 담당할 것이다. 또한 멀티미디어로 대변되는 정보시스템과 연계하여 새로운 영역을 창출하고 음성인식분야의 실용화 등도 더욱 촉진할 전망이다.