

컴퓨터가 책을 읽는다

OCR(Optical Character Recognition)

많은 분량의 텍스트를 타이핑해 본 사람이라면 텍스트를 자동으로 타이핑해주는 장비가 하나 있으면 하고 바란 적이 있을 것이다. 그러나 이제 그 바램은 현실화되고 있다. 인쇄된 문서를 전자 문서로 변환하는 기술, 바로 OCR(Optical Character Recognition) 이 등장했기 때문이다.

OCR은 말 그대로 광학적으로 인식할 수 있는 문자를 컴퓨터가 읽을 수 있도록 하는 기술을 말한다. 원리적으로 1자를 수십 개의 모눈으로 분할하여 특정한 모눈의 흑백 또는 자획형상의 특징에 의해 문자를 판독하는 것이다.

현재로는 특정형의 타이프 문자로 인지한 숫자 및 알파벳에 한하여 읽을 수 있고, 사람이 직접 쓴 글자까지는 읽지 못한다는 한계가 있다.

PDF 전자문서로 더욱 주목받는 OCR 기술

OCR 기술이 처음 개발된 것은 1929년 독일 G. Tauscek이라는 사람에 의해서였다. 그는 특정한 신호를 인식해 코드화할 수 있는 기술을 소개했고, 이를 통해 문자인식과 관련된 첫 특허를 획득했다. 그러나 초창기 G. Tauscek의 OCR 기계는 형판을 이용해서 빛을 검출해내는 장치에 불과했다. 형판을 통해 나온 빛이 통과하기 전의 것과 같은 것인지 판별하는 것이었다. 이는 원시적인 OCR이었지만, OCR의 기본 원리를 이용한 최초의 OCR 기계였다.

이후 1950년대 미국에서는 안보를 위해서 인쇄된 메시지를 컴퓨터 처리를 위한 기계 언어로 변경시키는 작

업을 진행했는데, 이런 작업을 사람을 거치지 않고 기계를 통하는 방법을 연구하게 됐다. 그렇게 개발된 시스템은 지금과 같은 모습의 OCR로 변화하게 되었다. 이런 OCR 기술이 처음 상업적으로 활용된 것은 1965년 영국 은행에서 청구서 지불 시스템이 도입되면서부터였다. 이는 컴퓨터가 인식할 수 있는 간단한 문자를 이용해서 처리 속도를 높인 것으로 업무 효율을 향상시켜 주었다. 이후에도 문자인식 기술은 시험지답안, 인구통계조사, 투표 검표 등에 활용되면서 꾸준히 발전해오고 있다. 그리고 최근 PDF가 기업, 개인 문서의 중심으로 등장하면서 고급OCR에 대한 연구가 집중되고 있다. 기존 하나나 둘의 문자를 인식하는 것이 아니라, 문장을 인식하고, 책을 읽어내는 기술이 개발 중인 것이다.

인쇄된 텍스트, 문서편집 가능한 텍스트로 변환

OCR 기술의 핵심은 종이문서를 전자문서로 만든다는 데 있다. 이는 종이라는 고정적인 한계를 극복하고, 전자문서로 다양한 활용이 가능한 상태로 만드는 것을 뜻한다. 그러나 아직까지는 문자 인식이 완벽하지만은 않다. 문자라는 것이 한글과 영어만으로만 이루어지지 않고, 다양한 전각, 반각 기호들이 추가되기 때문이다. 자

간이나, 행간에 따라서도 오류를 일으키기도 하는데, 이는 기술개발자들이 꾸준히 수정을 해오고 있다. 현재 영어는 90% 정도, 한국어의 경우에는 70~80% 정도 해독이 가능하다. 그러나 이 정도만 하더라도 분량이 많은 텍스트의 경우에는 상당히 유용한 수준이다. OCR을 이용해 문자를 편집가능한 텍스트로 변환하고 나면 몇몇 부분은 간단한 수정하면 되기 때문이다. 무엇보다도 OCR에 주목하는 것은 손쉽게 인쇄된 문서를 컴퓨터에서 사용할 수 있도록 텍스트화 할 수 있다는 점이다. 이는 앞으로 보관용 문서의 경우에 전자화해서 종이의 낭비를 없애고, 영구히 보존할 수 있도록 한다는 점에서 환영할 만한 일이다. 이는 실생활에 유용하게 사용될 수 있는데, 예를 들어 명함을 인식할 수 있거나, 기타문서들을 손쉽게 정리할 수 있다.

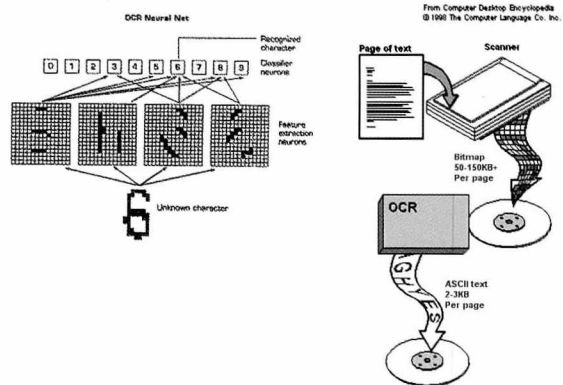
빛파동 이용 2차원 도형 1차원 신호로 개조

문자인식장치는 크게 광전변환장치, 기억장치, 인식처리장치, 출력장치로 구성된다. 광전변환장치는 지면의 문자를 전압의 일정한 파형으로 변환시키고, 2차원 도형인 문자의 형상은 파형을 분석해서 1차원 신호로 변환한다. 기억장치는 인식처리 장치를 제어하는 연산 제어 프로그램과, 인식한 형상을 특정한 형태로 표현한 데이터를 저장하게 되는데, 이를 통해서 기억장치 내의 표준 패턴과 비교해 가장 일치도가 높은 패턴에 해당하는 문자의 종류를 불러오고, 출력장치는 인식된 문자를 컴퓨터나 기타 정보처리장치에 전달하기 쉬운 형태로 전달한다.

문자 인식을 할 수 있는 기기로는 OCR(광학식 문자판독기, Optical Character Reader) 외에도 필기문자를 인식할 수 있는 태블릿(tablet), 특수문자를 인식할 수 있는 OMR(광학식 마크판독기 : Optical Mark Reader), MICR(바코드, 자기잉크판독기 : Magnetic Ink Character Reader) 등이 있다.

OCR기술의 어두운 그림자, 저작권 문제

이렇게 문서를 컴퓨터화하는 기술이 반드시 유익하다고만 할 수는 없다. 다양한 저작권 관련 문제들을 불러올 수 있다.



기존에는 책을 전자문서로 만들기에 어려움이 많았다. 그림파일로 저장을 하거나, 아니면 일일이 원하는 부분을 타이핑을 해서 이용해야 했다.

보이지 않는 매체간의 벽이 존재했던 것이다. 그러나 OCR 기술을 이용하면 책 한권이 몇 시간 만에 전자문서로 바뀐다. 이는 스캐너와 프로그램만 있으면 누구나 손쉽게 변환할 수 있기 때문에 책이 담고 있는 콘텐츠는 무방비에 노출되어 있다고 봐도 무리가 아니다. 그렇게 되면 저작권과 관련된 사항들이 더욱 모호해지게 된다. 출판계에서는 최근 서적 미리보기조차 크게 문제를 삼고 있는데, 이런 기술이 보편화되면 출판업계, 인쇄업계는 큰 타격을 입을 수도 있다.

이런 우려에 대해서 프로그램 제작업체들 사이에서는 자체적으로 보안을 강화하는 방법을 모색하고 있기도 하다.

김정삼 기자 sang@print.or.kr