"DB구축, 이제 수작업시대는 끝!"

각종 서류나 문서를 스캐너로 읽어들인 후 컴퓨터 데이타로 전환시켜주는 문자인식소프트웨어는 DB, 출판, 광파일링 등 많은 한글데이터의 전산화가 필요한 분야에 유용하게 사용될 것으로 전망된다.

취재/이석기

●(주)삼홍시스템, NeuroOCR●

– 신경망 문서자동인식 시스템 –

문자 인식 기술은 영상처리 기술의 한 분야인 패턴인식의 한 분야로서 이미 오래전 부터 연구가 이루워져 왔다. 그러나 워낙 고난도 기술이라 상품화 되어 판매되고 있는 제품은 그리 많지 않는 것이 현실이다.

문자 인식에는 펜으로 문자를 패드위에 작성하면 즉시 인식하는 온라인 방식과 이미 인쇄된 문자의 자소를 분리하여 인식하는 오프라인 방식 등 크게 2가지로 분류할 수 있다. 이번에 소개되는 NeuroOCR은 오프라인 방식을 도입한 문자인식 소프트웨어로서 광학 스캐너를 사용하여 문자 데이터를 영상 이미지로 입력한 후 이를 컴퓨터가 이해할 수 있는 문자 데이터(ASCII의 TEXT 데이터)로 변환시켜 주는 것을 말한다.

선진국에서는 이미 문자인식 시스템을 개발하여 사무 자동화에 사용되고 있으며, 국내에서도 영문 문자인식 소프트웨어는 스캐너 판매 업체들이 스캐너와 함께 공급하여 많은 사람들이 활용

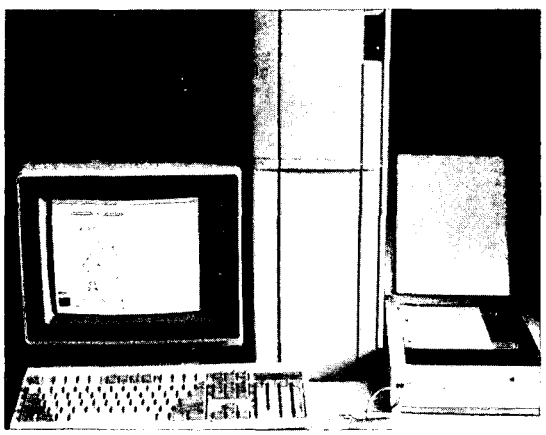


그림 1

하고 있다.

그러나 한글인 경우에는 영문보다 몇 배 더 어렵기 때문에 상용화가 늦어지고 있다. 그 이유는 매우 간단하다. 한글은 우리가 알다시피 글자의 수도 많고, 글자의 모양도 매우 다양하다. 한글은 대부분 초성, 중성, 종성으로 이루워져 있기 때문에 이를 분류하여 인식한다는 것이 여간 어려운 일이 아니다. 또한 불행히 과거의 국내 인쇄 기술이 발달하지 않아 인쇄 상태가 불량하기 때문

에 한글을 인식하는데 더 어렵게 만들고 있다. 그러나 다행히 국내의 인쇄 기술이 점점 발전하여 앞으로는 좋아질 것으로 예상된다. 인쇄상태에 자소가 정확히 분류되어 있고 문자 패턴이 손실되지만 않았다면 한글인식도 어려운 것만은 아니다.

■ NeuroOCR 신경 회로망 개념

NueroOCR은 신경 회로망 기법을 응용하여 문자를 인식하고 있다.

이는 다양한 크기와 다양한 활자체의 한글과 영어, 숫자, 특수문자를 인식하도록 하였다. 혼용문서를 인식하기 위한 방법으로 한글의 구조적 특성을 이용하여 먼저 유형분류 신경 회로망으로 한글을 6가지, 영어, 숫자, 특수문자를 1가지로 하여 총 7가지 유형으로 분류한후 각 유형별로 인식 신경 회로망을 사용하여 문자를 인식하는 계층적 구조로 구현하였다.

그림 2는 각 유형별 인식 신경회로망의 구성으로서, 한글은 초성, 중성, 종성으로 분할한 후 자소단위로 인식하였다. 사용된 다층 퍼셉트론 신경회로망은 Backpropagation 학습 알고리즘을 사용하였으며, 신경망 입력정보로는 글자영상에서 유용한 피쳐를 추출하여 사용하였다.

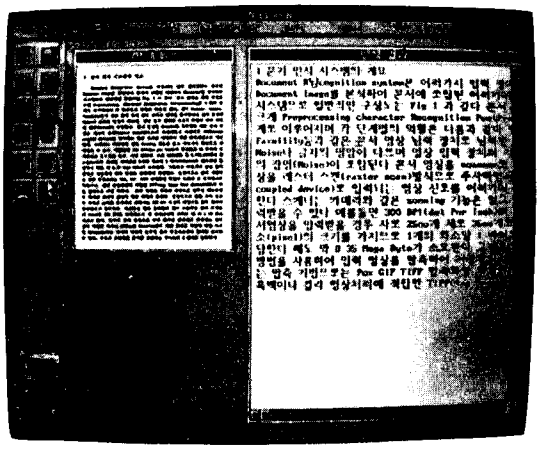
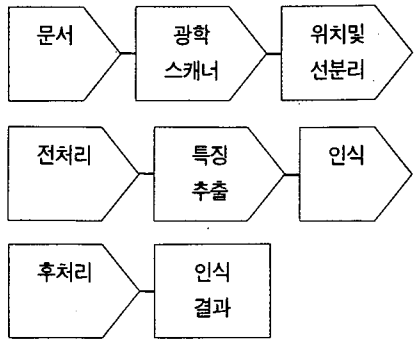


그림 2

NeuroOCR의 처리과정은 다음과 같다.



■ NeuroOCR의 특징

NeuroOCR은 오프라인 방식을 이용하였기 때문에 이미 종이에 인쇄된 한글이라면 인식할 수 있다. 문서를 편의로 구분한다면,

1) 한글 문자만 있는 문서
2) 한글 문자와 영문이 혼합된 문서
3) 한글 문자와 한문이 혼합된 문서
4) 문자(영문 포함)와 도표, 그림이 혼용된 문서

등으로 분류가 가능하다. 현재 판매되고 있는 1차버전은 주로 첫번째로 분류한 문서만 인식이 가능하도록 설계되어 있다. 특히 워드 프로세서에서 편집되고 명조체로 출력된 문서는 인식률이 더욱 좋다.

그러나 2차 버전에는 1), 2), 4)으로 분류한 문서의 인식도 가능하다고 한다. 또한 다단인식과 문서의 기울어짐 보정기능, 다양한 특수문자등 사용자가 일반적으로 입력하고자 하는 문서는 대부분 인식이 가능하도록 설계하고 있다. 주지해야할 사실은 NeruoOCR은 개인용 컴퓨터에서 운용이 가능하다는 것이다. 또한 기존에 사용하고 있는 DOS만 있어도 가능하기 때문에 다른 운용체계가 필요없고, PC에서 운용되기 때문에 초기 구매비용이 매우 저렴하다는 것이다.

그러나 별도의 하드웨어 없이 소프트웨어로만 운영되기 때문에 PC의 기능이 좋아야 원하는 인식속도를 구현할 수 있다. NeuroOCR은 인식속도가 초당 35자 이상(띄어쓰기 까지 합하면 초당 45자이상)인식할 수 있으며 인쇄상태에 따라 다

르기는 하지만 98%의 인식률을 가지고 있다. 예를 들어 A4용지위에 글자수만 1500자 정도 있을 때 스캐너의 구동시간과 인식까지 합하면 1분 30초정도의 시간이 소요된다.

또한 중요한 것은 입력 장치인 스캐너가 필수로 있어야 한다는 것이다. 스캐너에는 여러 종류가 있고 그 기능 또한 서로 다르다. NeuroOCR은 스캐너의 해상도는 300 DPI라야 하며, 흑백이어야 한다. 물론 칼라 스캐너도 흑백으로 입력할 수 있다. 스캐너는 여러 회사것이 있으나 NeuroOCR은 HP사의 제품(1차제품)이나 U-MAX, MICROTEK(2차제품)사의 제품은 직접 구동되고, 이외의 회사 제품은 스캐너에서 문서를 입력받아 저장한 후 이 파일(PCX 포맷으로 저장되어야 한다)을 NeuroOCR에서 파일을 불러와서 인식할 수 있다.

그러나 NeuroOCR은 필기체 인식은 안되며 인식된 결과를 다른 워드 프로세서에서 편집, 출력해야 한다. 이는 문자를 인식하는 역할을 충실히 하기 위해 그러한 기능을 포함시키지 않았다고 한다.

■ NeuroOCR을 사용할 수 있는곳

NeuroOCR은 많은 사람들에게 편리함과 시간을 절약해 준다. 이러한 서비스 제공은 기존의 업무 방식을 크게 개선 시켜주기 때문에 비용절감이나 시간절감을 가져올 수 있어서 경영에도 많은 도움을 준다.

사용할 곳을 간단하게 요약하면 NeuroOCR은 많은 양의 문자데이터를 컴퓨터에 수작업으로 입력시키고자 하는 고객에게는 아주 유용하다. 이 분야에 해당하는 고객으로서는 데이터베이스분야, 출판분야, 각종 서적입력 용역업체등에서 사용할 수 있다. 또한 광파일 구축업체, 맹인 독서 시스템이나 자동번역 시스템을 연구 또는 구축하는 업체는 매우 유용하게 사용할 수 있다. 또한 수시로 논문, 정부간행물, 기술서적, 잡지등을 컴퓨터에 입력하기 원하는 개인에게도 유용하다.

■ NeuroOCR의 시스템 구성

NeuroOCR은 그림 3과 같이 386이상의 개인용 컴퓨터(PC)를 사용하여 문자를 인식하는 시스템으로, 스캐너로 입력된 문자영상으로 부터 개별문자를 추출한 후 한글, 영어, 숫자, 특수문자를 인식한 다음 편집가능한 ASCII코드로 저장한다.

인식된 문자는 다른 워드 프로세서에서 사용가능하며, 음성합성 카드를 장착하면 음성으로도 출력이 가능하다. 또한 컴퓨터의 성능이 인식속도를 크게 좌우하므로 좋은 컴퓨터를 사용해야 한다.

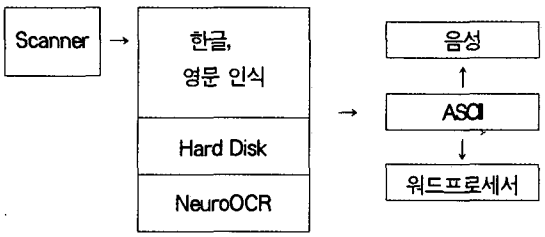


그림 3 NeuroOCR 시스템 구성

NeuroOCR에서 사용하는 컴퓨터는 486 컴퓨터(486-DX2 (66MHz)), 16MB 메모리, 20MB의 빈 하드디스크, 슈퍼 VGA Board, 마우스등이 필요하다.

■ 마지막 NeuroOCR의 기능을 간단하게 요약하면 아래와 같다.

1) 명조체 고딕체 인식 가능
2) ASCII 파일로 저장 가능
3) 초당 30자 이상 인식 가능
4) 97%의 인식률
5) 특수문자 인식 가능
6) 영문 인식 가능
7) 인식결과 음성합성 카드로 출력 가능
8) PCX 형태로 저장된 문서 영상 인식 가능
9) 다양한 글씨 크기 가능(9-21)
10) 각종 워드 프로세서에서 편집 및 출력 가능