OCR " 글돌이 " 의 계층구조

o
이 균 하
인하대학교 전자계산학과

A Hierarchical Structure on OCR "Guldori"

Kyoon Ha Lee
Department of Computer Science, Inha University

요약

한글을 포함하는 문서의 인식을 위한 시스템의 설계, 유지, 보수 및 확장을 체계적이며 용이하도록 하기 위한 문서인식 시스템의 기능별 계층화 구조를 제안하고 실용화를 목표로 추진중인 OCR "글돌이"의 원형 시스템에의 적용 타당성을 조사하였다. 각 계층은 인접 계층과 인터페이스만으로 연결토록 하여 상호 독립적인 방식을 취하였으며 특히 문자인식 등과 같이 소프트웨어의 구조가 복잡한 계층은 하드웨어 및 firmware 의 형태로 구성을 하여 임의의 워크 스테이션 또는 임의의 스캐너와 쉽게 접속되도록 하였다.

I. 서 론

문서인식 시스템은 서구문에서는 일부 실용화가 이루어지고 있으며 한글이나 한자의 경우도 실용화 가능성을 향하여 많은 연구들이 발표 또는 추진중에 있다. 그 외에 주변산업 또는 주변장비라고 할 수 있는 스캐너, 워크 스테이션과 호스트 컴퓨터 등도 상당한 속도로 발전되어 가고 있다. 또한 문서인식 시스템 자체도 단순히 개개의 문자를 만족할만 한 수준으로 인식하는 것이 시급한 문제이긴 하지만 어휘의 타당성을 고려한 문서 인식 시스템 기술을 계속 발전시켜 문서의 인식률을 높일 필요가 있다. 더 나아가서 문장의 줄거리 파악까지를 포함한 자연어 처리 차원에서의 문서인식 시스템 구축까지를 고려하는 높은 차원까지 발전시킬 필요가 있다.

이러한 환경에서 모든 부분을 하나의 틀에 맞추어 넣는다는 것은 문제를 대단히 복잡하게 만들 우려가 다분히 있다. 본 보고서에서는 이러한 문제점들을 고려하여 문서인식 시스템의 각 요소들을 기능별로 분리하고 계층화하여 계층별로 독립적인 설계 및 유지. 보수.확장이 가능토록 구조화하는 제안을 하였다. 제안된 구조는 현재 인하대학교 전자계산학과에서 실용화를 목표로 추진중인 OCR "글돌이"의 원형시스템에 적용하고 있다.

II. OCR 관련기술에 대한 고찰

문서인식을 위한 OCR 시스템은 일반적으로 그림1에서 보여주는 바와 같이 여러가지 기능으로 나누어 질 수 있다. 이들 각 기능들과 관련된 기술들은 영상입력을 위한 스캐너를 비롯하여 입력된 영상으로부터 날자영상을 추출하는 기술, 추출된 날자영상으로부터 문자를 판독해내는 인식기술, 문자판독 과정에서의 모호성 및 오류를 해결하기 위한 어휘의 타당성 검사 및 자연어 처리차원에서의 관련기술등으로 나누어 생

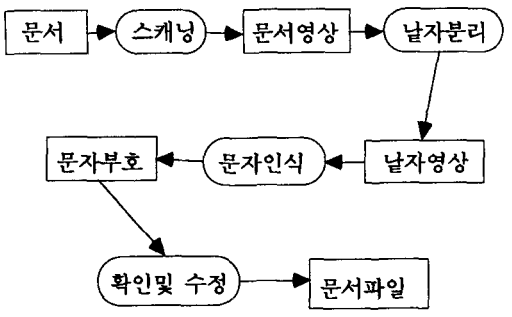


그림 1. 문서인식 시스템 블럭도

각할 수 있다. 위의 관련기술들을 구체적으로 조사, 분석하면 다음과 같이 열거할 수 있다.

영상의 입력을 위한 스캐너의 경우 목적에 따라 다음과 같은 여러가지 형태로 제공되고 있다.

- o 사용용도 --- 다목적 (TV 카메라)
  --- 문서전용 스캐너
- o 색의 구분능력 --- Monochrome
  --- Color
- o 명암처리 --- Gray level
  --- Binary level

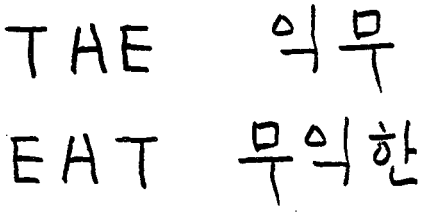
현재 문서전용의 밀착식과 같은 고정 초점방식의 스캐너가 문서의 인식에는 적합한 것으로 판단되어 널리 사용되고 있으나 장래에는 지면이 아닌 기타의 물체에 기록된 문자의 인식을 위하여 TV 카메라와 같은 다목적 영상입력장치도 사용 가능하여야 할것이다. 색의 구분능력에 있어서 흑백으로 충분하지만 그림과 글자가 혼재한 경우 칼라기능이 도움을 줄 수 있다. 명암처리 역시 이치값으로 충분하다고 생각하기 쉽지만 문서가 작성된 종이나 인쇄의 질이 좋지 않을 경우 일정한 값을 기준으로 "흑" 과 "백" 을 출력하는 단순한 방식보다는 16 등급 이상의 gray level 을 제공하는 방식이 높은 인식률을 위하여 크게 도움을 줄 수 있다.

일단 스캐너를 통하여 컴퓨터에 입력된 영상은 문자 또는 이와 유사한 크기의 단위로 분리되어야 인식이 가능해진다. 이와 관련된 기술분야로서는 신문의 영상으로부터 기사부분을 분리하는 기술[1]과 그림이 포함된 문서로부터 문장 부분만을 분리하는 기술등과 같은 문장 분리기술이 있다.

이렇게하여 얻어낸 문장의 영상은 행과 열로 나누어 문자단위로 분리하여야 한다. 일반적으로 사용되는 가로쓰기 방식에서는 행과 행 사이는 경계가 뚜렷하여 행의 구분에 문제가 없으나 열과 열 사이는 심한 접촉현상을 나타내고 있어서 이 문제를 해결하는 낱자 분리에 관한 연구 보고들도 찾아볼 수 있다[2,3]. 특히 한글을 비롯한 동양활자권에서는 가로쓰기 뿐 아니라 세로쓰기도 사용되고 있어서 이 분야의 계속적인 연구가 필요하다.

낱자 단위로 분리된 문자영상은 문자인식 소프트웨어에 의하여 분석되어 인식이 가능하게 된다. 문자를 인식하는 기술은 여러가지들이 발표되고 있으나 글자의 크기와 꼴의 종류에 관계없이 높은 인식률을 가지고 인식해낼 수 있는 기술을 위해서는 계속적인 연구를 필요로 하고 있다.

임의의 문서에 포함된 모든 낱자의 영상을 하나하나 완벽하게 정확히 인식하기에는 어렵다기보다는 불가능하다고 보아야 한다. 우선 인쇄의 질에 따라 인쇄잉크가 번진 경우의 획과 획 사이의 접촉현상이 있을수 있고 인쇄잉크의 부족으로 하나의 획이 연결성을 잃는수가 있다. 이러한 현상은 신문용지와 같은 저급의 종이질에서 특히 두드러지게 나타난다. 양호한 인쇄의 경우에도 영상 스캐너의 특성상 또는 부적절하게 주어진 명도값으로 획사이의 접촉 및 연결성의 상실이 있을수 있으며 이러한 이유때문에 낱자 단위의 완벽한 인식을 어렵게 하고 있다. 더 나아가서



(a) (b)
그림 2. 모호한 문자의 예
(a) "THE" 와 "EAT"
(b) "의무" 와 "무익한"

그림 2 에서 보이는 바와 같이 문자 자체가 갖는 이중해석 가능성(모호성) 때문에 인식오류는 항상 염두에 두어야 할 사항이다.

이와같이 여러가지 원인에 의하여 나타난 인식오류 또는 모호성은 앞 뒤 문맥에 따라 교정 또는 확정되어야 할 것이다. 좁은 시각에서 볼 때에는 인식결과를 어휘 단위에서 타당성 검토로 해결 할 수 있으며 장기적인 안목에서 볼 때에는 문장의 줄거리까지를 파악하여 해결하는 방법이 필요할 것으로 본다.

문자인식의 경우는 아니지만 인식결과가 모호한 경우에 통계적 방법이나 fuzzy 이론을 적용하여 해결하는 연구[4,5]도 있으며 인식결과가 대단히 중요한 내용일 경우에는 사용자가 직접 확인할 수 있도록 하는 방안도 필요하다.

이상에서 고찰한 바의 각종 OCR 관련기술들은 동시에 획일적으로 개선될 성질이 아니라 많은 시간을 두고 경험과 기술이 축적되고 발전되어 점진적으로 개선될 성질의 것으로 여겨진다. 본 보고서에서는 이에 대비하여 OCR 의 각 기능과 관련된 기술들이 독립적으로 용이하게 설계, 개선될 수 있는 하나의 계층구조를 제시하고 실용화를 목표로 한 OCR "글돌이"의 원형시스템에 적용하고자 한다.

III. 광학식 문서인식 시스템을 위한 계층구조

계층구조의 일반적인 개념에서는 상위계층에서 하위계층에 대한 호출만을 허용하고 하위계층으로부터 상위계층으로는 호출에 대한 응답만을 허용하고 있다. 이러한 개념에 따라 위에서 기술한 OCR 과 관련된 각종 기능들을 계층화 할 경우 일단은 그림 3 과 같은 구조를 생각할 수 있다.

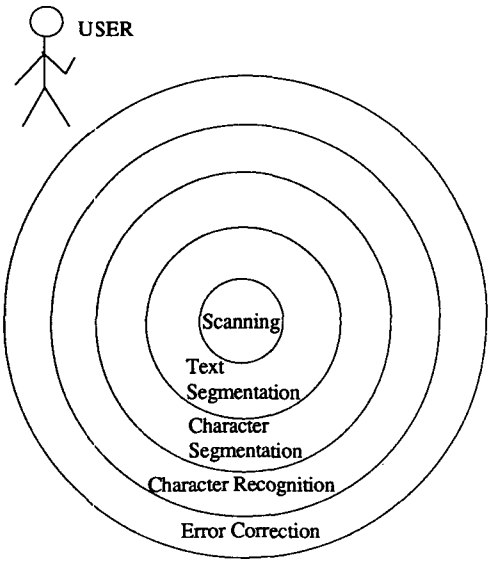


그림 3. 광학식 문서인식 시스템의 계층구조

여기서 Error correction 계층은 사용자의 요구에 따라 Character recognition 계층을 호출하여 오류가 포함되어 있을지도 모르는 문서인식 결과를 받아 들인 다음 가능한 범위에서 정확히 교정하여 사용자에게 제출하는 기능을 갖게 되며 Character recognition 계층은 Character segmentation 계층을 호출하여 낱자 하나의 영상을 받아 분석하여 인식해낸 결과를 부호화하여 Error correction 계층에 제공한다. Character segmentation 계층이 호출되면 문서영상을 저장하는 버퍼에 저장된 문단영상으로부터 행과 열을 구분하여 한개의 낱자영상을 출력하게 된다. 만일 문단영상의 모든 행과 열이 인식 완료되면 하위 계층인 Text segmentation 계층을 호출하여 다음에 처리해야 할 새로운 문단의 영상을 전달받게 된다.

Text segmentation 계층은 인식해야 할 모든 문단들을 문서영상으로부터 분리하여 상위계층에 제공하며 보유하고 있는 문서영상으로부터 더 이상 처리할 문단이 없는 경우 Scanning 계층을 호출하여 새로운 문서의 영상을 제공받게 된다.

이상에서 설명한 각 계층의 구조 및 알고리즘은 다양한 형태로 구현될 수 있겠으나 각 계층간의 인터페이스만을 정교하게 그리고 융통성있게 마련한다면 문서인식 시스템의 목적에 따라서 또는 개발 환경에 따라서 얼마든지 독립적으로 운영될 수 있을 것이다.

IV. OCR "글돌이"의 계층구조 설계

앞에서 제시된 계층구조의 각 계층에 대한 구현방법에는 여러가지가 있을 수 있겠지만 본 보고서에서는 현 시점에서의 상황에 적절하다고 판단되는 방법으로 OCR "글돌이"의 원형시스템에 구현하였다.

Scanning 계층은 상위 계층의 호출에 의하여 한 면의 문서영상 입력을 담당하는 계층으로 하드웨어인 스캐너와 이를 구동하는 구동모듈이 필수적이며 스캐너의 종류에 따라 달리 구현되어야 할 계층이다. 따라서 이 계층을 보다 구체적으로 분석한다면 그림 4 와 같이 구체화 될 수 있다. 스캐너는 하드웨어 장치이지만 경우에 따라서는 컴퓨터 디스크에 저장된 영상파일이 될 수도 있으며 이 경우에는 구동모듈도 당연히 영상파일을 다루기에 적합한 것으로 대치되어야 한다.

"글돌이"에서는 밀착식 이미지 스캐너를 선택하여

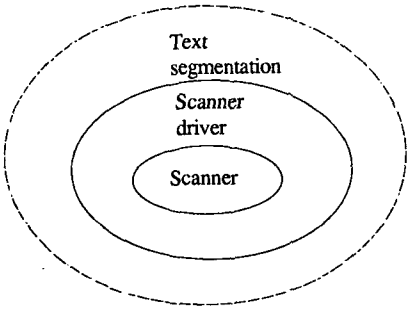


그림 4. Scanning 계층의 구체화

연결하였으며 영상의 해상도, 입력영역의 좌표설정, 조명램프의 제어 등 스캐너의 구동을 위하여 필요한 20가지 이상의 제어명령들이 있다. 이들 제어명령은 구동모듈에 의하여 자동적으로 발생되기도 하며 필요한 경우에는 대화형식 모듈에 의하여 수동제어가 가능하게 하였다. 스캐너의 기능은 컴퓨터 디스크에 수록된 영상파일로 대치될 수도 있도록 마련하여 필요하다면 문서영상의 오프라인 입력도 가능하다. 이 방식은 최근 급격히 보급되고 있는 광 파일 시스템과 관련을 지을 수 있으며 앞으로 널리 보급될 광대역 통신망과 연결된다면 원격 입력도 가능할 것이다.

가변초점방식의 스캐너인 경우에는 밀착식 경우의 제어명령 이외에도 초점조절, 방향조절, 줌 기능 등 다양한 제어명령이 추가될 것으로 예측되며 초기에는 대화형식에 의한 수동제어에서 소프트웨어의 발달에 따라 상위계층으로부터의 자동제어도 기대 되어진다.

Scanning 계층과 Text segmentation 계층사이에는 제어와 결과보고를 위한 여러가지 정보가 통과될 수 있겠으나 가장 대표적인 것은 문서영상을 담고있는 영상버퍼라 할 수 있다. 300 DPI 의 해상도에 이치영상 신호로 A4 크기의 문서영상을 입력할 경우 약 1.1 Mbyte정도의 영상버퍼가 필요하며 gray 레벨을 64 등급으로하고 색 신호도 포함하는 경우에는 버퍼의 크기를 9 Mbyte 이상 충분히 고려해야 하기 때문에 현재 구현 대상에서는 보류되었으나 앞으로 CPU 속도의 발전과 기억용량의 증대에 따라 기대해 볼만 한 사항이다. 이치 영상에서의 1.1 Mbyte 역시 작은 용량이 아니기 때문에 기억장소의 절약을 위한 영상부호의 압축이 바람직하며 문서 전용의 영상을 대상으로 실험한 결과에 의하면 Modified Huffman 부호가 100 Kbyte 미만으로의 압축을 가능하게 하고 있으나 부호화와 복호화를 위한 CPU 부담의 증가가 수반되었다. "글돌이"에서는 CPU 부담 경감과 영상부호 압축의 절충안으로서 단순한 run length 부호로 문서영상을 영상버퍼에 구축하였다.

Text segmentation 계층은 스캐닝에 의하여 받아들인 문서영상 으로부터 문장의 한 단에 해당하는 영상만을 분리해내는 기능을 갖는다. 백지에 문서만 한 단으로 질서 정연하게 작성된 경우에는 문제가 없지만 그림, 선분 기타의 도형이 포함된 경우에는 문장부분만을 단 단위로 구분하는 것이 필요하다. 신문의 영상을 분석하는 연구보고도 있으나 CPU 의 부담도 문제가 될 뿐 아니라 현 시점에서 활용이 용이하지 않을것으로 판단되며 더욱이 여러가지 종류의 문서에 대하여 획일적으로 적용할 수 있는 알고리즘은 쉽지 않을것으로 보인다. 따라서 "글돌이" 의 원형 시스템에서는 현 시점에서 적절하다고 판단되는 man-machine 협동에 의한 해결을 위하여 고 해상도 그래픽 모니터와 마우스 및 키보드를 사용하여 대화형식으로 구현하였다.

이 때 사용자는 모니터에 나타난 문서 영상을 관찰하여 인식을 하고자 하는 문단부분을 마우스를 이용하여 설정하는 것으로 문단 분리작업은 끝난다. 이 방법은 사용자에게 주어진 문서영상 중에서 불필요한 부분을 피하고 필요한 부분만을 쉽게 선택할 수 있다는 점에서 바람직하다고도 할 수 있다. 또 한 사용자가 모니터를 통하여 문서영상의 질을 쉽게 파악할 수 있으므로 영상의 질이 불량할 경우 쉽게 스캐너의 제어 파라메타를 수정하여 재 입력을 시도할 수 있다. 물론 문서자체가 기울어져 영상이 회전된 경우에도 모니터를 통하여 즉시 파악이 가능하므로 올바른 각도로 재 입력을 할 수 있다.

문단으로 분리된 영상은 상위계층으로 전달되어야 한다. 이 때 문단의 영상을 별도의 영상버퍼에 적재할 수도 있으나 문서영상버퍼를 공유하고 버퍼상의 좌표값만을 전달하도록 하여 문단영상 저장용 버퍼를 절약하였다.

Character segmentation 계층은 row segmentation (행 분리) 과 column segmentation (열 분리) 으로 세분화된다. 인쇄나 종이의 질이 현저하게 불량한 경우를 제외하고는 문서전용 스캐너를 통하여 입력되는 영상에 잡음이 적은 편이므로 가로 및 세로 투영방법을 이용하여 행과 열을 분리하였다. 가로쓰기에서 행간의 확실히 구분되어 있어서 분리과정에 문제될 것이 없으나 열간의 간격은 문자의 심미성과 가독성을 높이기 위

하여 서로 어긋나게 배열되거나 접촉하는 수가 흔하게 있어서 본지에 별도로 발표하는 낱자분리 알고리즘을 이용하여 분리하였다. 세로쓰기의 분리를 위해서는 열의 분리가 용이하게 되는 반면 행의 분리가 쉽지 않게 되므로 별도의 설계가 추가되어야 할 것이다.

문자인식계층은 문서인식 시스템의 핵심이 되는 부분으로 본 연구실에서 이미 개발하여 발표한 바 있는[6,7] 모듈을 별도의 하드웨어에 이식하여 시험 사용하고 계속 개선중에 있다. 이 부분은 문자를 구성하는 획들의 기하학적 속성들과 상관구조를 분석하여 문자인식을 하며 따라서 글의 꼴이나 크기에 영향을 비교적 작게 받도록 되었다. 또 한 이 부분은 설계가 용이하지 않고 복잡하며 문서인식 시스템의 성공여부의 비중이 높은 부분이므로 동일한 알고리즘을 여러가지 CPU 또는 서로 다른 조건의 하드웨어에 이식시킬 경우 번거로움과 위험부담이 따를것으로 판단된다. "글돌이"에서는 이러한 요소들을 고려하여 별도의 CPU 를 갖는 하드웨어에 firmware 및 다운로드 모듈로 구현하고 고속 인터페이스로 호스트와 연결이 가능토록 하여 문서인식의 핵심부분이 각종 워크 스테이션 또는 호스트 시스템에서 호환성을 갖도록 구현하였다.

낱자 단위로 문자인식을 하는 문자인식 계층은 앞에서 지적한 바와 같은 인쇄의 질, 종이의 질 등의 여러가지 이유로 완벽한 인식이 곤란한 경우가 있을 수 있다. 확실하지 않은 인식결과를 상위계층에 아무런 정보없이 전달하는 것은 보다 더 높은 인식율을 위하여 바람직스럽지 못하다. 문자의 인식 결과는 확실하다고 판단되는 경우(1)와 확실치 않은 인식결과에서 이중해석 가능성이 있는 모호한 경우(2)와 모호성 없이 인식은 되었으나 표준이 되는 문자구조와는 약간의 거리감이 있어서 인식결과가 확실하지 않고 미심적은 점이 있는 경우(3) 및 인식이 불능한 경우(4)가 있을 수 있으며 이러한 내용들이 인식된 낱자의 부호와 함께 속성(attribute)으로 부여되어 출력 되도록 하였다. 특히 (2) 의 이중해석 가능성의 경우는 가능하다고 판단되는 문자들의 부호와 함께 각 부호에 대한 확률값을 함께 속성으로 부여하여 상위 계층에서의 판단에 도움이 되도록 하였다.

Error correction 계층은 문자인식 계층으로부터 속성과 함께 전달받은 문자부호들을 문맥에 따라서 또는 문서 원본의 내용에 따라서 수정하는 일을 해야 한다. 앞의 설명중 (1) 에 해당되는 확실한 경우라도 인식 오류가 포함될 가능성은 있을 수 있으며 이 경우는 어휘조사 또는 문장의 줄거리를 파악하는 자연어처리기술 까지도 필요할 수 있다. (2) 의 모호한 경우는 가능성이 있는 각 문자부호의 가능성 확률값에 일반 문장에서의 (또는 주어진 문장에서의) 각 문자부호가 갖는 조건확률 (conditional probability) 값들을 적용하여 해결할 수 있다. (3) 과 (4) 의 경우에는 문서원본과 대조하여 결정하는 것이 확실한 방법이지만 오류의 포함을 상당수준 허용하는 경우라면 어휘의 타당성 검토에 의하여 자동으로 수정을 할 수도 있다. 그러나 분명한 사실은 자동으로 수정을 할 경우에는 모든 자료사전의 구성 및 통계값 및 문맥분석이 완전해야만 하며 현재로서는 모든 경우에 만족될 수 있는 기술 수준이 못된다고 볼 수 있다.

따라서 "글돌이"의 구현에서는 고 해상도의 그래픽 모니터를 사용하여 문자 인식에 문제가 되는 주위 부분의 문서영상을 비트 이미지로 출력하고 인식 결과의 문자부호를 폰트로 작성 출력하여 사용자가 "글돌이"와 대화형식으로 용이하게 오류를 수정 및 모호한 부분의 결정을 할 수 있도록 마련하였다. 이러한 구현 방식에 의하여 현 수준에서의 소프트웨어에 의한 어휘 타당성 검토 결과와 함께 사용자의 문맥에 따른 판단 및 시각에 의한 사용자의 판독 결과를 동시에 수용할 수 있기 때문에 실용화에 큰 도움을 줄 것으로 기대된다. 한편 계속적인 문자인식 및 Error correction 계층의 발전은 사용자의 오류수정 부담을 점진적으로 줄여줄 수 있으며 약간의 오류를 허용하는 일반문서의 경우는 완전 자동처리까지도 가능하리라고 본다.

이상에서 설명한 대화형식의 오류수정이 바람직한 방법일 수 있지만 다른 한편으로 볼 때 대화형식은 사용자가 모니터 앞에 항상 대기해야 하는 부담을 주게 된다. 따라서 필요한 경우에는 man-machine 간의 대화를 중단하고 문자인식 계층에서 제공한 속성치와 어휘 타당성 조사결과 등을 문서 인식 결과의 파일에 그대로 포함시켜 출력함으로서 후일 문서원본을 확인하면서 워드 프로세서를 이용하여 배치(batch)형태의 수정을 행할 수도 있도록 구조화 하였다.

V. 맺는 말

전체적인 입장에서 볼 때 단순하지 않고 대단히

복잡하다고 볼 수 있는 문서인식 시스템과 관련되는 각종 기술 및 기능들을 고찰하였으며 각종 기술 또는 기능들은 관련 기술들의 발전 또는 하드웨어 및 환경의 발전에 따라 각 각 독립적으로 발전될 가능성이 높다는 지적을 하였다. 독립적으로 발전되는 각종 기술 또는 기능들을 문서인식 시스템에 손 쉽게 적용할 수 있도록 하기 위한 기능별 계층화 설계의 타당성을 확인하기 위하여 실용화를 목표로 추진중인 OCR "글돌이"의 설계에 적용을 하였다. 각 계층은 현재의 기술 수준과 환경을 고려하여 바람직하다고 판단되는 형태로 구현을 하였으며 계속적인 각 계층들의 발전은 타 계층에 큰 영향을 주거나 받지 않으면서 이루어질 수 있다고 판단된다. 또한 각 계층의 독립적이고 지속적인 발전은 실용적인 한글문서 인식 시스템을 가능하게 할것으로 믿으며 더 나아가서는 능률 높은 문서 인식 시스템의 구축이 가능하리라 생각된다.

[참 고 문 헌]

[1] K. Inagaki, T. Kato, T. Hiroshima and T. Sakai, "MACSYM: A Hierarchical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents," Pattern Recognition vol. 17, no. 1, pp. 85-108, 1984.

[2] R.G. Casewy, G. Nagy, "Recursive Segmentation and Classification of Composite Character Pattern," IEEE Conf. on PR&IP, pp. 1023-1026, 1982.

[3] Yea Shuan Huang, Ken Wen Lin and Yihung Chen, "Field Segmentation and Character Isoration Method in Free-Format Chinese Printed Document," on 89' International Conf. on Computer Processing of Chinese and Oriental Languages, pp. 151-155, 1988.

[4] H. C. Lee and K. S. Fu , "A Stochastic Syntax Analysis Procedure and its Application to Pattern Classification," IEEE Trans. on Computers, vol. C-21, no. 7, pp660-666, July, 1972.

[5] M. G. Thomason, "Finite Fuzzy Automata, Regular Fuzzy Languages and Pattern Recognition," Pattern Recognition, vol. 5, pp. 383-390,1973.

[6] Kyoon Ha Lee, Kie Bum Eom and R. L. Kashyap, "Character Recognition using Attribute Grammar," on IEEE Proc. CV&PR, pp. 418-423, 1988.

[7] 이 균하, "ADRPG를 이용한 한글 문서 인식," 인하대학교 기초과학연구소 논문집 제 9 호, PP. 67 - 71, 1988.