□특집□

한글처리 - 문자 중심 인식 기술 고찰

안 창† 이 상 범††

◈ 목 차 ◈

1. 서 론

문자는 음성과 함께 사람이 언어로 표현하는 정보를 전달하기 위해 발명된 대표적인 부호 체계이다. 추상적인 정보로서의 문자 개념은 필기 도구와 사람의 필기 운동이라는 물리적 작용에 의해 구체적 형태로 기록되어, 이를 읽는 사람에게 정보를 전하는 것이다.

문자가 탄생한 이래 인간의 문명은 급속히 발전하게 되어 초기에는 필기에 의한 문자 표현이 주종을 이루었으나, 인쇄 기술의 발달에 힘입어 인쇄체 문자로 정보를 전달하는 경우가 일반화되었다.

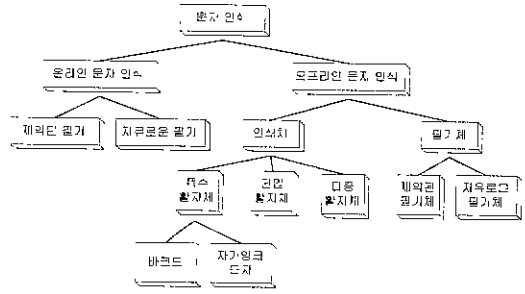
최근 고도의 정보화 사회로의 발전은 컴퓨터, 팩시밀리, 복사기 등의 이른바 사무자동화용 기기의 보급에 힘입은 바가 크다. 즉, 정보를 발생시키는 일은 훨씬 용이해진 반면에 인간의 지식을 얻는 주된 수단은 눈으로 사물을 보거나 귀로 소리를 들어 입력되는 패턴(pattern) 즉, 영상이나 소리를 인식하고 이를 지식으로 축적하는 연속적인 과정에 한정되어 있다.

따라서 정보의 홍수라 불릴 만큼 많은 정보 중에서 사용자가 원하는 정보만을 얻기 위해서는 작성된 서류 혹은 설계도면, 프로그램 및 데이터, 필기된 전표 등을 컴퓨터에 자동으로 입력할 수 있는 기술이 매우 필요하게 되었다. 문자 인식은 이와 같은 "입력의 자동화" 요구에 부응하는 기술로서 발달하게 되었고, 이 기술을 장치화한 제품이 바로 OCR(optical character reader)로서 흔히 광학 문자 판독기라 부르고 있다.

본 고에서는 컴퓨터에서 한글 처리를 위한 문자 중심 인식 기술의 동향과 문제점 등에 대해 개괄적으로 고찰한다.

2. 문자 인식 기술의 개요

문자 인식 기술은 인식 대상에 따라 그림 1과 같이 분류할 수 있다[1].

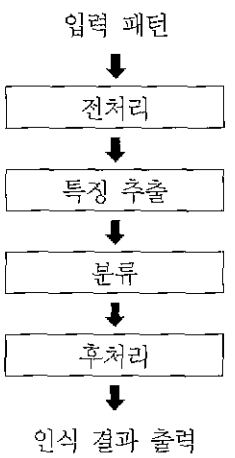


(그림 1) 문자 인식의 대상

† 정회원 . 단국대 전자공학과 박사과정
†† 정회원 : 단국대학교 컴퓨터공학과 교수

일반적으로 전자펜 혹은 태블릿(tablet)으로 문자를 입력하여 시간적, 공간적 정보를 갖게 되므로 획순, 획수, 필기 방향이나 속도 등의 부가 정보를 가질 수 있는 온라인(on-line) 정보 입력에 따른 인식 기법과 종이에 작성된 문서를 스캐너(scanner) 등에 의해 입력하여 공간적 밝기 정보를 가지는 오프라인(off-line) 정보 입력에 따른 인식 기법으로 나눌 수 있다. 또한 인쇄체 문자에 비하여 필기체 문자가 더 복잡한 변형을 포함하고 있으므로 필기체 문자를 인식하기 위해서는 복잡한 변형을 어떻게 흡수할 것인지가 가장 중요하다. 복잡한 변형을 흡수하기 위해서는 필기 문자의 본질을 정확히 파악할 필요가 있다. 필기 방법으로는 일정한 영역 안에서 자유롭게 필기하는 방법과 제약을 가하여 필기하는 방법이 있으며 이는 필기자가 문자를 쓸 때 입력 영역에 제한을 두어 필기에 의한 변형을 가급적 적게하는 것이다. 그러나 필기자에게 지나치게 노력을 강요하는 것은 불합리하므로 인식 시스템의 전체적인 효과를 고려하여 입력 방법을 정하는 것이 중요하다.

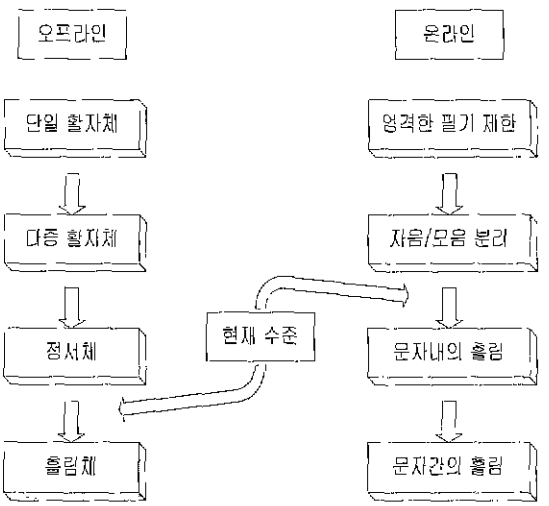
문자 인식은 그림 2와 같이 입력 데이터에서 특징을 추출하고, 미리 정해진 여러 모델 중의 한 모델로 분류하는 과정으로 구성된다.



(그림 2) 문자 인식 과정

초기의 문자 인식 기술에서는 주로 개별 문자의 인식에 연구의 초점을 맞추었으나, 다양한 문서 및 전표 등을 효과적으로 자동 입력하기 위해서 문서에서 필요한 정보만을 추출하여 적절히 처리하거나, 문서 자체의 구조를 분석하여 그 의미를 이해하는 진정한 정보 입력의 자동화를 실현하기 위한 연구로 발전하게 되었다[1].

현재까지 광학 문자 인식에 대한 연구는 상당한 수준까지 진척되어 이미 여러 시스템이 실용화되었으며, 최근에 와서는 필기체 문자 인식까지 연구 영역을 넓히고 있으나, 문서의 구조 분석 및 이해에 관한 연구 성과는 상대적으로 미흡한 형편이다. 국내에서는 신문이나 잡지, 교과서 등의 제한된 문서 영상을 대상으로 하여 부분적인 문서 구조 분석 및 이해에 관한 연구가 이루어지고 있으며, 문서 구조 분석을 통하여 기사를 추출하거나 문서 내에서 문자 이외의 영역을 분리 추출하는 정도의 연구가 대다수를 차지하고 있다. 따라서 문자 인식과 함께 문서 자체의 구조를 분석하여 그 의미를 완전히 이해하는 연구가 수행되어야 할 단계이며, 연구 수준을 도표화하면 그림 3과 같다.



(그림 3) 한글 문자 인식 기술의 수준

3. 문자 인식의 문제점

컴퓨터를 이용하여 문자를 인식하는 경우, 고려해야할 사항은 영상 입력 방법에 대한 문제와 입력된 영상 내에 혼재되어 있는 정보 중에서 문자 정보만을 추출하는 문제, 그리고 추출된 영역내의 질적인 문제, 이러한 기능을 실제로 구현할 때 발생할 수 있는 양적인 문제를 고려할 수 있다.

3.1 영상 입력 방법에 관한 문제

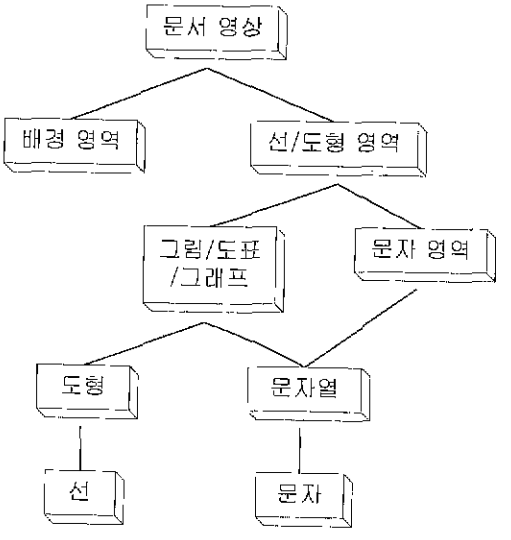
영상 내에서 문자 인식을 잘하기 위해서는 인식 대상 문자가 가장 잘 보이도록 입력하여야 한다. 즉, 입력 대상의 크기, 색상, 밝기 등의 속성에 따라 인식 대상 영역이 다른 불필요한 데이터와 구별이 가능하도록 설계하는 것이 중요하다. 예를 들어 입력 장치에 색 필터를 부착하여 영상의 전처리 과정에 대한 부담을 줄이는 방법, 혹은 대상 영역이 잘 드러나는 위치에서 입력하는 방법을 취하는 것이 인식 과정에 유리하다[2].

3.2 정보 분리에 관한 문제

오프라인으로 입력된 문서 영상의 경우 영상 내에 존재할 수 있는 정보의 구성 형태는 그림 4와 같이 표현할 수 있다[3].

문서 영상의 특징을 살펴보면 서로 다른 의미를 갖는 영역이 혼재 또는 중첩하여 존재하는 것을 알 수 있다. 따라서 문자를 인식하기 위해서는 이와 같은 여러 영역의 분류 및 제거 과정이 필요하며, 영역의 분류를 위해 문서 영상을 형성하는 각 영역의 특징 또는 성질을 정의할 수 있어야 한다[4].

일반적으로 사람이 문서를 보는 경우 문자로 읽혀지는 부분은 문자 영역, 그림으로 보이는 부분은 그림 영역으로 정의하고 있으나, 이와 같은 방법으로는 컴퓨터에서 처리가 불가능하므로 입력된 영상의 신호 레벨 즉, 256 밝기 단계를 갖는 농담 영상(grayscale image)의 경우 0~255까지의 신호 값에서 각 영역을 정의할 수 있어야 한다. 이를 "신호 레벨의 모델"이라 부르고 있다.



(그림 4) 문서 영상의 구성 요소

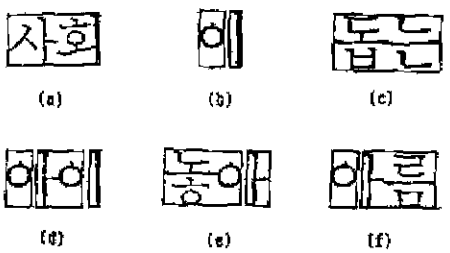
예를 들어 "경계(edge)가 명확히 존재하면 문자이고 국소적인 밝기의 변화가 낮은 부분은 그림 영역이다"라고 정의할 수 있을 것이지만, 이와 같은 간단한 정의만으로 모든 경우를 완전히 식별하거나 분류할 수 없다. 그러므로 영역 분류 결과의 좋고 나쁨에 따라 인식 결과의 정확도가 크게 좌우됨을 알 수 있다. 특히 밝기의 정도가 유사한 두 정보가 서로 중첩되어 존재하는 경우, 예를 들면 표가 작성되어 있는 문서에서 선과 문자가 서로 중첩(overlap)되어 존재하거나, 문자와 문자가 서로 접촉(touching)되어 존재하는 경우에 이를 효율적으로 분리하기 위한 연구가 활발하다[5-10].

한글 인쇄체 인식의 경우 문자 영역은 문자열로 구성되어 있으며, 문자열은 여러 문자의 집합으로 되어 있다. 따라서 문장에서 문자열을 추출하고 이를 낱개 문자로 분리하는 방법이 일반적

인 처리 과정으로 되어 있다. 이러한 처리 과정을 "신호 레벨의 모델"을 이용하여 처리한다면 추출 방법에 대한 조건의 몇 가지 정의가 다음과 같이 가능할 것이다.

① 한글 문자는 정방형을 이룬다
② 문자는 직선 상에 나열된다
③ 연속되는 문자는 비슷한 크기를 갖는다
④ 행간(行間)보다 자간(字間)이 좁다

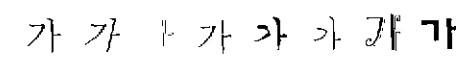
이외에도 여러 가지 조건을 부가하여 문장의 낱자 분리가 가능하겠지만, 한글 문장의 가독성(可讀性)과 심미성(審美性)을 높이기 위해 열과 열 사이를 근접시켜 인쇄하는 경우가 많아 그림 5와 같은 접촉 현상이 발생하기 쉽다.



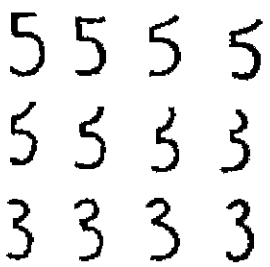
(그림 5) 인쇄체 한글의 접촉 형태

3.3 질적인 문제

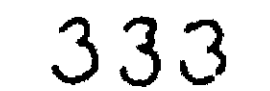
사람이 가지고 있는 뛰어난 문자 인식 기능은 직관과 경험에 바탕을 두고 있다. 따라서 컴퓨터를 사용하여 문자를 인식할 경우, 인식 과정을 객관화하고 정량화하여 알고리즘의 형태로 공식화하는 일반적인 방법이 완성되어 있지 않기 때문에 질적인 어려움이 발생한다. 예를 들어, 인쇄체 문자의 경우 다양한 활자체나 필기 문자의 크기 및 기울어짐의 불일치 등으로 인하여 인식에 어려움이 발생하게 된다. 그림 6에는 문자 인식시에 질적인 어려움을 야기시키는 문자의 예를 보였다.



(a) 다양한 형태의 문자 예



(b) 유사 문자의 예



(c) 일그러진 혹은 훼손된 문자 예

(그림 6) 질적인 어려움이 있는 문자 예

3.4 양적인 문제

문자 인식에서 질적인 어려움을 해결할 수 있다 해도 공학적인 측면에서 보면 양적으로 방대하여 기억 용량 면이나 인식 소요시간 면에서 실제 구현하기 어려운 경우도 발생한다. 예를 들면, 여러 가지 모양의 문자 패턴을 모두 컴퓨터에 기억시켜 놓고 문자 패턴을 비교하여 문자를 인식한다면, 각 문자를 20×20의 점으로 표현하고 각 점이 단순히 0 또는 1의 값만을 갖는다고 해도, 가능한 패턴의 수는 2400이나 된다. 이는 기억 용량 면에서 문제가 될뿐 만 아니라, 그렇게 많은 패턴을 비교할 수가 없다.

이러한 양적인 어려움은 밝기를 갖는 문자 영상, 즉 정보량이 많은 경우에 특히 심하며, 단일 활자체를 인식 대상으로 하는 경우, 숫자를 읽기 위해 작성된 알고리즘을 본질적인 수정 없이 한자(漢字)에도 적용한다면, 실제로 인식 장치를 구성할 때 글자의 종류가 증가하기 때문에 문자의 형상을 기억하는 기억장치가 증가하고, 다양한 가능성을 비교 판단하는 인식 처리 장치의 규모가 커지게 되어, 조건이 없는 문자 인식 장치를 실현

하기 곤란하다. 이 문제는 기술적이고 경제적인 문제이므로 비용의 제한만 없다면 해결할 수 있으나 경제성을 간과할 수는 없다.

4. 인식 기술

4.1 인식 방법

오프라인 한글 인식에 대한 연구는 고품질의 단일 인쇄체 한글부터 시작하여 저품질의 인쇄체 한글, 다중 활자체 한글을 거쳐 필기체 한글의 인식에 이르고 있다. 오프라인 문자 인식 방법은 크게 원형 정합 방법(template matching), 확률 통계적 방법(statistical approach), 구조적 방법(structural approach), 신경망을 이용한 방법(neural network based approach) 등으로 분류될 수 있다[1,11-18].

원형 정합 방법은 입력 문자 영상을 인식 대상이 되는 모든 문자 모델과의 정합을 통하여 유사도나 거리를 구하여 인식하는 방법이다. 원형 정합 방법은 인식 방법이 단순하고 하드웨어로 구현하기가 쉬우며, 단일 활자체와 같이 변형이 심하지 않은 문자 영상에는 적합하나, 필기체 문자와 같이 변형이 심한 경우에는 적용이 어렵다. 그러나 일본어, 중국어, 한글 등과 같이 문자의 형태가 복잡하고 인식해야할 문자 부류의 수가 많은 대용량의 필기체 문자 인식에 대한 연구가 활발히 진행되면서 필기체 문자에서 발생하는 변형을 효과적으로 흡수할 수 있는 비선형 형태 정규화 또는 비선형 형태 정합 방법 등이 개발됨에 따라 원형 정합 방법의 성능이 개선되었다.

확률 통계적 방법은 문자 영상의 통계적인 특징을 분석하는 방법으로 입력 문자 영상으로부터 특징 벡터를 추출한 다음, 판별 함수를 이용하여 특징 벡터가 특징 공간 상에서 어느 부류에 속하는가를 알아내어 입력 패턴을 분류하는 방법과, 표현하고자 하는 대상 패턴을 학습 과정을 통해 확률 값으로 표현하고 입력 패턴에 대해서 이를 주어진 모델로 생성해 낼 수 있는 확률 값을 계산하여 가장 높은 확률 값을 갖는 모델로 분류하는 은닉 마르코프 모델을 이용하는 방법 등이 있다. 이 방법들은 대체로 이론적으로 잘 정립되어 있으며, 본질적으로 많은 변형을 갖는 패턴들을 잘 처리할 수 있다는 장점을 갖는 반면, 한글과 같이 유사한 글자가 많고 복잡한 문자 집합에 대해서는 문자의 본질적인 구조적 특징을 이용하기가 쉽지 않다는 단점을 가지고 있다.

구조적 방법은 문자의 구성 원리에 입각하여 자획(stroke) 등과 같이 문자를 구성하는 기본 요소와 그들간의 연관성을 추출하여 문자를 인식하는 방법이다. 이 방법은 특히 복잡한 구조를 갖는 문자 집합의 인식이나 글자 모양의 변형이 심한 필기체 문자의 인식에 적합하지만, 인식 알고리즘의 기본이 되는 기본 요소의 추출 작업이 어렵고, 문자의 구조를 표현하는 규칙의 자동 생성 및 이러한 규칙에 기반한 문법적 추론 기법에 관한 연구가 아직까지 미약하다.

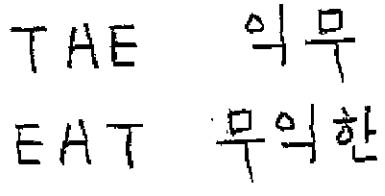
신경망을 이용한 방법은 인간의 두뇌 조직을 모델링한 것으로, 문자 인식을 위하여 단순한 기능을 수행하는 처리기들이 대규모 상호 연결된 병렬 분산 네트워크를 이용하는 방법이다. 신경망을 이용한 문자 인식은 학습에 의해 새로운 패턴에 적응할 수 있으며, 패턴의 국부적인 변형 및 잡영(noise)에 민감하지 않다는 장점을 갖는 반면, 패턴의 수가 큰 경우 학습에 걸리는 시간이 매우 길어지며, 인식 후보 대상의 개수가 많은 경우 성능이 저하된다는 단점을 갖고 있다.

4.2 후처리의 필요성

최근까지 진행되어 온 국내의 한글 인식에 대한 연구 방향을 고려할 때 대부분의 연구가 문자 인식 알고리즘에만 집중되어 왔다. 그러나 현실적으로 효과적인 문자 인식(혹은 패턴 분류) 알고리즘 만으로는 입력 장치의 특성 및 필기자의 다양

한 필기 형태로 인하여 사용자가 요구하는 수준의 높은 인식률과 처리 속도를 기대하기가 어렵다. 따라서 이러한 한계성을 극복하기 위해서는 한글 인식 알고리즘의 개발 뿐 만 아니라 입력 장치의 특성으로 인한 잡영의 분석, 입력 영상 처리시에 발생하는 정보의 손실, 그리고 필기체 문자에서 발생하는 변형의 양상 즉, 획 간의 접촉 변형, 획의 기울기 변형 등 필기자에 따라 다양하게 나타나는 문자 내의 형태 변형을 체계적으로 분석하는 연구가 필수적이다[19].

이와 같은 신호 레벨의 모델에 따른 결과 뿐 만 아니라 인식된 결과에 대한 검증이 필요하다. 인식 과정이나 결과 확신도에 있어서도 아무런 문제가 없을 것으로 보이는 그림 7의 예에서 알 수 있듯이 전후 문맥의 흐름을 이용하여 인식하는 연구가 필요하다.



(그림 7) 인식 결과의 보완

5. 결 론

최근 고도 정보화 사회로의 급진적 발전 추세에 따라 정보 산업이 새로운 차원에 들어서고 있다. 또한 산업 발전과 기술의 대형화, 고도화 등으로 인하여 매년 방대한 양의 정보가 처리되고 있다.

정보화를 이루기 위해서는 대부분 종이로 기록되어 전해오던 모든 정보를 컴퓨터에 저장하여, 이를 필요로 하는 사람이 적시 적소에 사용할 수 있어야 하며, 사무 자동화와 함께 보급된 개인용 컴퓨터의 빠른 확산으로 인하여 많은 양의 정보를 단시간에 처리할 수 있는 고도의 기술을 필요로 한다. 따라서 종이에 기록된 문자 데이터를 보다 효과적으로 컴퓨터에 입력하여야 한다. 문자 인식 기술의 발달로 인하여, 지금까지 대부분 수작업으로 이루어진 데이터의 입력을 자동으로 컴퓨터의 데이터 파일로 만들어 주는 것이 가능하게 되었다.

현재 문자 인식 기술은 우편물 자동 분류를 위한 우편 번호 인식, 산업 현장에서의 제품 검사나 분류, 문서 인식, 도면 인식, 팩스를 통해 전달받은 영상에서의 문자 인식, 워드프로세서 OCR, 금융기관에서의 전표 또는 수표의 자동 입력 등 여러 분야에 걸쳐 실용화되어 실생활에 효과적으로 사용되고 있다.

앞으로 컴퓨터에의 정보 입력 자동화에 대한 사회적 요구는 더욱 증대될 것이고, 우리 나라에서도 컴퓨터 이용의 고도화에 따라 한글 정보의 입출력 문제가 매우 중요하게 되었다. 따라서 앞으로 실용적인 온라인 필기 한글 입력 장치 뿐 만 아니라 인쇄체 한글 및 필기체 한글을 효과적으로 판독할 수 있는 OCR의 개발이 요망되고 있다. 그러므로 본 고에서 지적한 영상의 입력에 대한 문제, 정보의 분류 문제, 추출된 문자의 질적 문제와 이를 구현할 때 고려할 양적인 문제가 선결되어야 하며, 인식 결과를 효율적으로 검증하여 수정할 수 있는 시스템의 구축이 요구된다.

참고문헌

[1] 이성환, 문자인식 -이론과 실제, vol.1, 홍릉과학출판사, 1994

[2] 山中弘, 書像處理応用技術, 工業調査會, 1989

[3] 美濃導彦, "文書畵像處理の現狀と動向", 電子情報通信學會誌 Vol.76, No.5, pp.502-509, 1993年 5月

[4] L.A.Fletcher, R. Kasturi, "A Robust Algorithm

for Text String Separation from Mixed Text/ Graphics Images", IEEE Trans. on PAMI Vol.10, No.6, pp.910-918, Nov. 1988

[5] 장경식, 김재희, "지도에서 도로와 블록 인식", 정보처리학회논문지 4권 9호, pp.2289-2298, 1997년 9월

[6] 안창, 박찬정, 이상범,"대화식 클러스터링 기법을 이용한 칼라 지도의 문자 영역 추출에 관한 연구", 정보처리학회논문지 4권 1호, pp.270-279, 1997년 1월

[7] 황순자, 김문현, "자소 클래스 인식에 의한 off-line 필기체 한글 문자 분할", 정보처리학회논문지 3권 4호, pp.1002-1013, 1996년 7월

[8] 배창석, 민병우, "문자 영역의 분리와 기하학적 도면요소의 인식에 의한 도면 자동입력", 전자공학회논문지 31권 B편 6호, pp.91-103, 1994년 6월

[9] S. Liang, M. Shridhar, M. Ahmadi, "Segmentation of Touching Characters in Printed Document Recognition", Pattern Recognition, Vol.27, No.6, pp.825-840, 1994.

[10] J.M.Westall, M.S.Narasimha, "Vertex Directed Segmentation of Handwritten Numerals", Pattern Recognition, Vol.26, No.10, pp.1473-1486, 1993

[11] R.C.Gonzalez, R.E.Woods, Digital Image Processing, Addison-Wesley Pub. Co. Inc., 1992

[12] S. Bow, Pattern Recognition and Image Processing, Marcel Dekker Inc., 1992

[13] R.C.Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience Publication, 1973

[14] 장대수, 진용옥, "필터링에 의한 한글 자소의 특징점 추출과 해쉬 함수에 의한 자소 분별 알고리듬", 대한전자공학회논문지, 29권 B편 5호, pp.300-309, 1992년 5월

[15] 권재욱, 조성배, 김진형, "계층적 신경망을 이용한 다중 크기의 다중활자체 한글문서 인식", 한국정보과학회논문지, 19권 1호, pp.69-79, 1992년 1월

[16] A. Khotanzad & Y.H.Hong, "Rotation Invariant Image Recognition Using Features Selected via a Systematic Method", Pattern Recognition, Vol.23, No.10, pp.1089-1101, 1990

[17] 김태균, T.Agui and M.Nakajima, "Stroke 조합에 의한 필기체 한글의 표현과 인식", 대한전자공학회논문지25권 1호, pp.18-26, 1988년 1월

[18] L.Xu, A.Krzyzak and C.Y.Suen, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition", IEEE Trans. on SMC, Vol.22, No.3, pp.418-435, May 1992

[19] 이성환, 김은순, "주소 및 성명에서의 한글 인식을 위한 효율적인 오인식 교정 알고리즘", 한국정보과학회논문지, 20권 5호, pp.729-738, 1993년 5월

안 창

1987년 단국대학교 전자공학과 (공학사)
1989년 단국대학교 대학원 전지공학과 (공학석사)
1995년-현재 단국대학교 전자공학과 박사과정
1989년-1994년 기아정보시스템 연구소
관심분야 : 영상신호처리, GIS

이 상 범

1974년 연세대학교 전자공학과 (공학사)
1978년 서울대학교 대학원 전지공학과 (공학석사)
1986년 연세대학교 대학원 전자공학과 (공학박사)
1984년 미국 IOWA대학교 컴퓨터공학과 객원교수
1979년-1992년 단국대학교 전자공학과 교수
1993년-현재 단국대학교 컴퓨터공학과 교수
관심분야 : 컴퓨터구조, 영상신호처리, GIS