특집

# 한글 문서 분석 및 인식 기술의 최근 연구 동향

김 두 식*, 김 상 엽**, 이 성 환*,**
*고려대학교 컴퓨터학과
**고려대학교 영상정보처리학과

## I. 서 론

한글 인식에 대한 연구가 이제 어느덧 삼십여년의 역사를 갖게 되었다. 문자 인식 분야는 실용적으로는 컴퓨터와 인간의 보다 원활한 인터페이스를 추구하려는 목적에서 출발하였으며, 학문적으로는 인간의 우수한 능력 중의 하나인 패턴 인식의 능력을 컴퓨터에게 부여하여 인간처럼 사고하고 판단할 수 있는 인공지능 컴퓨터의 실현에 목적이 있다.

특히, 근래에는 휴대가 가능한 초소형 컴퓨터의 등장과 함께 작고 간편한 명령 입력 도구의 개발을 위한 온라인 문자 인식 기술에 관한 연구가 크게 주목을 받게 되었다면, 최근에는 대용량 멀티미디어 시대의 도래와 함께 기존의 방대한 문서 정보를 일괄적으로 전산화할 수 있는 오프라인 문자 인식 기술에 관한 연구가 시급히 요구되고 있다.

문자 인식의 방법은 가장 단순한 원형 정합 방법에서 시작하여 수학적인 해석에 바탕을 둔 확률 통계적 방법, 문자를 구성하는 기본 요소들의 구조적 연관성을 이용하는 구조적인 방법, 인간의 뇌에 대한 생물학적 모델링에 기반한 인공 신경망을 이용하는 방법에 이르기까지 다양한 접근 방법들이 동서양의 구분을 떠나 폭넓게 연구되고 있다[1].

그러나, 동양의 문자 체계는 서양의 그것과는 달리 문자의 수가 방대하고 문자의 구조적 특성이 상이하기 때문에 서양에서 연구되어 왔던 문자 인식 방법들을 그대로 적용할 수가 없다. 동양권 문자 체계는 크게 한자, 한글, 일어로 구분될 수 있는데, 이중 일어는 문자의 수가 50여자에 불과하므로 서양권의 문자 인식 기술을 쉽게 적용할 수가 있는 잇점이 있다. 한자의 경우에는 문자의 수가 너무 방대하기 때문에 아직까지 실용적인 해결 방법이 제시되고 있지 못하는 실정이다. 반면에 한글의 경우에는 비록 문자의 수가 11,172자에 이르지만, 자주 사용되는 KSC 5601 완성형 2,350자만을 고려하거나 한글의 구조적 특징을 이용하여 자소별로 인식하는 등의 해결책을 모색하고 있다.

최근에는 하나의 언어를 대상으로 하는 인식 문

제에만 국한하지 않고 다중 언어 문자 인식에 관한 연구가 활발히 진행되고 있는데, 이는 실생활에 주로 사용되는 문서들이 다양한 언어들을 함께 포함하고 있다는 현실에 기인한다. 다중 언어 문자 인식에 관한 문제는 서로 다른 특성을 갖는 다중 언어 문자들을 하나의 통합된 인식기로 인식하려는 접근 방법과 각 언어의 특성에 적합한 언어별 인식기들을 적절히 통합하려는 접근 방법으로 나눌 수 있다[2].

문자는 입력된 문자의 생성 형태에 따라 인쇄체 문자와 필기체 문자로 구분된다. 인쇄체 문자 인식 분야에서는 다중 활자체에 대한 인식의 문제가 비교적 오랫동안 연구되어 왔으며, 활자체에 의한 변형을 최소화하는 특징 추출기의 설계와 활자체의 특성에 구애받지 않는 인식기의 구현을 연구하고 있다.

필기체 문자는 인쇄체 문자와는 달리 필기자의 필기 특성에 따른 문자의 변형이 심하고 동일 필기자라고 하더라도 필기시마다 문자의 형태가 다양하기 때문에 전처리 및 특징 추출 단계에서 문자의 변형을 최소화하고, 획의 변형에 대한 적응적 정합 기능을 갖는 인식기의 개발을 연구하고 있다.

문자 인식 기술이 실용화되기 위해서는 문서 인식 기술을 수반하여야 한다. 문자 인식 기술은 개별적인 문자에 대한 인식만을 고려하지만, 문서 인식 기술은 스캐너와 같은 다양한 입력 장치를 통해 입력된 문서 영상으로부터 잡영을 제거하고 기울어짐이나 휘어짐 등의 다양한 형태 변형을 보상하는 영상 처리 기법을 포함할 뿐만 아니라, 문서 내의 구조적인 특징과 논리적인 정보를 추출하여 단과 문장을 구분하고, 문자 분할 기술을 이용하여 개별 문자들을 분리하는 등의 전반적인 문서 처리 기술들을 일컫는다.

본 논문에서는 최근의 연구 동향을 중심으로 다양한 오프라인 한글 인식 기술들을 소개하고, 현 한글 인식 기술에 관한 연구의 방향을 전망한다.

II장에서는 인쇄체 한글 인식에 대하여, III장에서는 필기체 한글 인식에 대하여 각각 논한다. IV장에서는 문서 구조 분석 및 이해에 관한 연구 동향을 살펴보고, V장에서 결론을 맺는다.

## II. 인쇄체 한글 인식

인쇄체 한글 인식의 문제는 현재까지 많이 연구되어왔으며 인식 성능에도 많은 진전이 있었다. 초기에는 다양한 특징 추출 기술들 중에서 인쇄체 한글 인식에 적합한 특징을 찾고자 하는 비교 분석적인 측면에서의 연구가 활발하였으며, 방대한 문자수를 갖는 한글의 인식 문제를 효과적으로 극복하기 위하여 인식기의 구조를 계층적으로 구성하는 인식기의 설계에 관한 연구가 꾸준히 계속되고 있다. 최근에는 다중 언어 및 다양한 활자체를 고려하는 인쇄체 한글 인식에 관한 연구가 활발히 연구되고 있으며, 통신 기술의 활성화와 함께 등장한 팩스 문서의 인식을 위한 저해상도 문자 인식 기술도 연구되고 있다.

### 1. 특징 추출

패턴 인식 문제에서의 특징 추출은 입력 패턴으로부터 불필요한 정보들을 제거하고 인식에 필요한 정보들만을 추출함으로서 인식에 필요한 정보의 최소화하고 인식기의 성능을 향상시키는 중요한 역할을 한다.

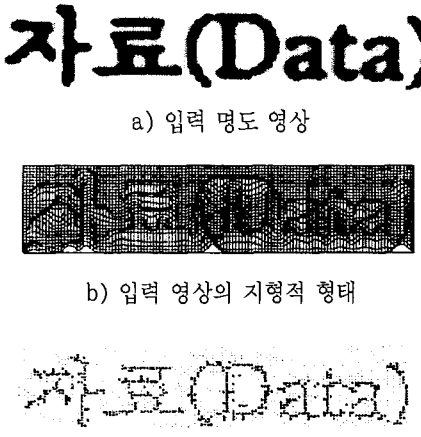
특히, 언어의 특성에 따라 인식에 적합한 특징이 서로 다르기 때문에 언어에 적합한 특징 추출 기술이 요구된다. 따라서, 특징 추출에 관한 연구는 인쇄체 한글 인식을 위한 가장 기본적이며 필수적인 분야라고 할 수 있다.

다양한 한글 활자체에 대하여 정보량과 엔트로피의 분포를 분석한 연구를 통해 대부분의 정보량이 문자의 가장 자리 부분에 위치하고 있음이 확인되었다. 이러한 연구 결과에 근거하여 수평, 수직, 대각선 방향으로 압축된 문자 영상에 Fourier 기술자를 이용한 특징 추출 방법이 등장하였고, 실험을 통하여 다양한 활자체와 크기를 갖는 대용량의 한글 인식 문제에 효과적임이 알려졌다[3].

일반적으로 문자 인식에 사용되는 특징으로는 기울기 특징, 그물눈 특징, 0 교차 특징 등이 있으며, 이들을 혼용 및 변환한 복합적인 특징들도 있다. 최근에는 통계적인 방법이나 신경망 이론에 근

거한 특징 선택 방법들[4]도 연구되고 있다.

한편, 문자 영상의 이진화로 인한 정보의 손실을 방지하기 위하여 명도 영상으로부터 직접 문자의 지형적 특징을 추출하는 방법[5]이 제안되었다. 이 방법은 그림 1(a)와 같은 명도 영상에 대하여 그림 1(b)와 같은 3차원적인 지형적 형태를 분석함으로서 그림 1(c)와 같이 피크(peak), 능선(ridge), 사면(hillside), 안장점(saddle), 협곡(ravine), 평지(flat), 분지(pit) 등과 같은 지형적 특징을 추출하였다. 또한, 이러한 지형적 특징들을 접촉된 문자의 분할 문제에 적용함으로서 이진 영상에서의 문자 분할 방법보다 우수한 성능을 보이는 문자 분할 기술이 제안되었다[6].




a) 입력 명도 영상

b) 입력 영상의 지형적 형태

c) 추출된 지형적 특징

〈그림 1〉 지형적 특징 추출

한글에서는 작은 획의 유무가 전혀 다른 문자를 의미하기도 하므로, 인쇄 상태가 나쁜 경우에는 한글의 인식 성능이 크게 저하되는 경우가 많다. 따라서, 인쇄 상태가 나쁜 상황에서의 한글 인식을 위한 효과적인 특징 추출에 관한 연구가 있다[7]. 여기에서는 그물 눈 특징, 교차 수 특징, 윤곽선 길이 특징, 연결 화소 특징, 압축 패턴의 윤곽선에 대한 Fourier 기술자를 사용한 특징들을 고려하여 인쇄 상태가 나쁜 상황에서의 한글 인식을 위한 다양한 특징 추출 방법을 비교 실험하였다.

## 2. 자소 기반 한글 인식과 음절 기반 한글 인식

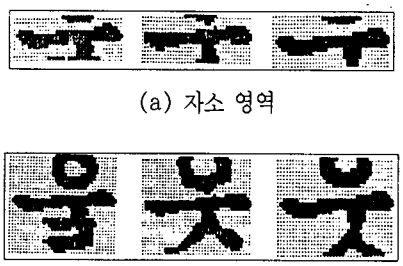
영어와는 달리 한글 인식의 문제에서는 11,172자나 되는 대용량의 문자에 적합한 인식 기술을 고려하여야 한다. 일상적으로 사용되는 문자들만 하더라도 2,350자나 되기 때문에, 영어 인식에 적용하였던 방법들을 그대로 적용하기가 곤란하다. 즉, 전통적인 구조적 방법으로는 대용량의 문자를 인식할 수 있는 시스템을 설계하기가 어렵고, 패턴 정합 방법을 적용할 경우에는 문자수가 많아질수록 인식 속도가 저하되며, 신경망 방법은 대용량 문자에 대한 학습 자체가 불가능하다.

따라서, 대용량 한글 인식에서의 이러한 문제점을 극복하기 위해 여러개의 인식기들을 계층적으로 결합함으로서 복잡한 하나의 문제를 여러개의 단순한 문제로 나누어 해결하려는 접근 방법이 일찍이 시도되었으며, 이를 크게 자소 기반 인식 방법과 음절 기반 인식 방법으로 나누어 볼 수 있다.

자소 기반 인식 방법은 한글의 모음 형태에 기반하여 한글을 6가지 유형으로 분류하고, 각 유형에 대한 자소의 위치가 일정하다는 특성을 이용하여 자소별로 문자를 인식하는 방법이다. 이러한 형태의 한글 인식기는 하나의 유형 분류기의 결과에 의해 여러개의 자소 인식기 중 적절한 자소 인식기들이 선택되는 두 층의 인식기 결합 구조를 갖으며, 여러개의 작은 신경망 인식기의 계층적인 결합으로 한글을 효과적으로 인식할 수 있음을 보이는 대표적인 예라고 할 수 있다.

특히, 이러한 자소 기반 인식 방법에서 자소 인식을 위해 추출되는 자소의 영역이 부정확하기 때문에 다른 획의 일부가 잡영 효과로 작용하여 신경망의 학습을 어렵게 한다는 것을 발견하고, 자소 영역을 확대함으로서 인식 성능을 향상시키는 연구가 있다[8]. 여기에서는 그림 2(a)와 같이 모음 'ㅜ', 'ㅡ', 'ㅜ'는 그림 2(b)와 같이 자소 영역을 확대함으로서 문자 구별의 모호성을 없앨 수 있었다.

자소 기반 인식 방법이 한글의 모음 형태에 따라 한글을 6가지 유형으로 인위적으로 분류하는데 반하여, 음절 기반 인식 방법은 문자 영상에서 추출된 특징들의 통계적인 분석에 근거하여 군집화



(a) 자소 영역

(b) 확대된 자소 영역

〈그림 2〉 확대된 영역에 대한 자소 인식

를 수행한다. 따라서 음절 기반 인식 방법은 대용량의 문자 집합을 여러개의 소집단으로 분류하는 대분류 단계와 대분류된 소집단 내에서 문자를 인식하는 상세 분류 단계로 구성된다[3].

이러한 음절 기반 인식 방법은 한글만의 고유한 특성에 국한하지 않고 문자의 통계적인 특성에 의하여 대분류기의 군집화를 수행하므로 한글이 아닌 영어나 한자에도 쉽게 적용이 가능하며 다중 언어의 인식을 위한 문제에도 적용할 수 있다[9].

음절 기반 인식 방법에 관한 내용은 다중 활자체 및 다중 언어 문자 인식과 관련하여 다음 절에서 구체적으로 논한다.

한편, 팩스 영상처럼 해상도가 낮고 변형이 심한 문자의 인식 문제에 대하여 자소 기반 인식 방법과 음절 기반 인식 방법의 장단점과 특징을 다양한 실험을 통하여 분석한 연구도 있다[10].

### 3. 다중 활자체 및 다중 언어 문자 인식

이질적인 환경에서 만들어진 다양한 문서들은 서로 다른 고유한 활자체를 갖으며, 동일 문서라 하더라도 문서의 가독성을 향상시키기 위해 다양한 활자체가 복합적으로 사용되는 경우가 많다. 따라서, 실용적인 문자 인식기의 개발을 위해서는 다중 활자체에 대한 문자 인식 기술이 요구된다.

초기에는 활자체의 변형을 흡수하려는 특징 추출 기술에 관한 연구나 인식기에 대한 학습의 일반화 능력을 개선하여 다양한 활자체에 대해서도 우수한 인식 성능을 갖는 인식 기술에 관한 연구가 있었으며, 최근에는 각각의 활자체에 대하여 독립적으로 학습된 개별적인 인식기를 유기적으로 통합하려는 연구가 있다. 일례로 숫자 인식의 경우에 인식기를 병렬적으로 결합하여 인식 성능이 향상될 수 있음을 보인 연구 사례가 있다[11].

전자의 예로 다중 활자체의 한글 인식을 위하여 학습의 일반화 능력이 우수한 신경망을 이용한 방법[12]에서는 문자 추출 신경망, 유형 분류 신경망, 문자 인식 신경망이라는 세 종류의 신경망을 이용하여 문자열에서의 문자 추출, 모음 형태에 따른 한글의 6가지 유형 분류, 자모의 인식 기능을 각각 담당한다. 특히, 이 방법은 전통적인 다층 신경망이 대분류의 기능을 효과적으로 수행함을 보였으며, 전통적인 자소 기반 인식 방법의 대표적인 예이기도 하다.

다양한 활자체의 대용량 한글을 고속으로 인식하기 위하여 최적 트리 분류기를 제안한 방법[3]은 앞에서 언급한 음절 단위 문자 인식 방법에 해당된다. 특히, 최적 트리 분류기의 효율적인 설계를 위하여 ISOETRP 군집화 알고리즘[13]을 도입함으로서 특징들의 통계적인 특성을 기반으로 트리 분류기의 단계적 분별 정확도를 최대화하고자 했다.

다중 언어 문자의 인식을 위해서는 언어마다 고유하게 갖고 있는 다양한 특성들을 효과적으로 활용할 수 있도록 인식기 구조가 설계되어야 한다.

일례로, 다중 언어, 다중 활자체의 대용량 문자의 인식을 위하여 계층적 신경망을 이용한 방법[9]에서는 다양한 활자체 및 크기의 문자로 인한 형태 변형을 보상하기 위하여 점밀도를 이용한 비선형 형태 정규화 방법을 적용하였고, 다중 언어의 대용량 문자에서 나타나는 서로 다른 위상 구조를 효과적으로 표현하기 위하여 계층적 특징 추출 방법을 고안하였다. 이러한 계층적 분류기는 적응적 SOFM 대분류기, LVQ4 언어 분류기, LVQ4 상세 분류기로 구성되며, 언어 분류기에 의해 분류된 각각의 언어들은 언어마다 개별적으로 학습된 상세 분류기에 의해 최종적인 문자 인식이 수행된다. 이렇게 인식기의 언어 종속성을 최소화한 모듈화 구조는 새로운 언어에 대한 확장성을 용이하게 하는 장점이 있다.

### 4. 문자 분할

문자 인식 기술에 대한 연구가 진전됨에 따라 이를 실생활에 적용하려는 노력이 진행되면서, 문자 분할의 문제가 제기되었다. 영문서의 경우에는 문자의 접촉이 대부분이지만 한글 문서에서는 문자간의 접촉뿐만 아니라 하나의 문자가 분리되는 경우가 흔히 존재한다. 특히, 실생활에 존재하는 한영 혼용 문서를 처리하기 위해서는 영어의 접촉 유형과 한글의 접촉 및 분리 유형을 동시에 고려해야 하기 때문에 보다 복잡한 형태의 문자 분할 기술이 요구된다.

영어와 한글이 혼용된 일반적인 한글 문서에서의 문자 추출을 위하여 신경망으로 구현된 문자 분리기를 이용하는 방법[14]이 제안되었다. 이 방법은 입력된 문자열 상에서 문자간의 접촉이 발생하였을 때, 문자 분리 신경망이 출력한 여러 개의 분리 후보점에서 인식을 수행한 후, 인식 결과를 이용하여 정확한 분리 위치를 선택한다. 이러한 문자 분리 신경망은 다양한 종류의 문자 접촉 유형을 학습함으로써 분리 후보점을 찾는데 사용된다.

문자열 영상을 이진화할 경우 접촉되거나 겹치는 문자의 경계 부분에서 문자 분할에 유용한 정보들이 많이 손실되므로, 이진화를 하지 않고 명도 영상에서 직접 문자 분할을 하는 연구도 있다[6]. 이 방법은 명도 영상의 문자 경계 부분에서 나타나는 지형적 특징과 명도값 변화를 고려하여 문자 분할 후보 영역을 결정하고, 다단계 그래프 탐색 기법을 이용하여 명도값을 추적함으로써 비선형 문자 경계를 찾은 후, 인식 결과를 이용하여 문자를 분할한다.

### 5. 인쇄체 한글 인식의 문제점 및 향후 연구 방향

이상과 같이 인쇄체 한글 인식에 대하여 전반적으로 살펴보았다. 과거의 인쇄체 한글 인식에 관한 연구가 제한된 조건에서 문자 인식기의 성능을 향상시키고자 하는 실험적인 연구에서 출발하였다면, 최근에는 문자 인식 기술의 실용화를 앞두고 다중 언어와 다중 활자체, 잡영과 변형이 심한 저해상도 문자, 복잡한 배경 영상을 갖는 문자를 고려한 다차원적 접근 방법이 다양하게 연구되고 있으며, 문자 인식 기술을 문자 분할 기술과 통합하려는 연구도 활발히 이루어지고 있다.

## III. 필기체 한글 인식

필기체는 인쇄체에 비하여 동일 문자에 대한 변형이 보다 복잡하고 글자간의 접촉 패턴이 다양하다. 따라서, 인쇄체 한글 인식에 관한 연구가 비록 초보적인 수준이지만 비교적 일찍 실용화 단계에 접어들고 있는 반면, 필기체 한글 인식에 관한 연구 분야는 필기 문자에 있는 형태 변형을 효과적으로 흡수하려는 형태 정규화 및 특징 추출에 관한 연구와 복잡한 패턴을 효과적으로 인식하는 분류기의 개발을 위한 연구를 중심으로 다양한 연구와 실험이 꾸준히 진행되고 있다[15].

본 장에서는 필기체 한글의 인식을 위해 사용하는 특징 추출 기법, 형태 정규화 기법, 낱자 단위 문자 분할 및 인식 기법에 대하여 소개한다.

### 1. 특징 추출

인식기의 성능은 인식 알고리즘 뿐만 아니라, 사용하는 특징에 크게 좌우되므로 효과적인 특징의 선택이 중요한 관건이 된다. 특히, 필기체는 인쇄체에 비하여 문자 간의 변형이 심하므로 이러한 변형을 효과적으로 보상할 수 있는 특징 추출 방법의 개발이 중요하다.

명도 영상에서 직접 특징을 추출하는 방법[5]은 문서의 이진화[16]에 의한 유용한 인식 정보의 손실을 피하기 위하여 이진화 과정을 거치지 않고 문자 인식에 사용될 특징을 추출한다.

변형이 심한 필기체 한글의 인식을 위하여 적응적 획정합 방법을 이용한 연구[17]에서는 세선화된 문자 영상의 체인 코드 정보를 기반으로 획의 구조적 특징점과 획을 추출하였고, 비선형 형태 정규화 방법을 이용한 연구[18]에서는 내접원에 바탕을 둔 획 밀도를 이용하여 비선형 형태 정규화를 수행한 후 획의 방향 성분을 특징으로 사용하였다.

은닉 마르코프 모델에 기반한 문자 인식 방법을

제안한 연구[19]에서는 입력 문자에 대하여 영역 투영 외곽선 변환을 수행하여 4 종류의 영역 투영 외곽선을 추출한 다음, 이들 외곽선에 대한 방향 성분을 특징으로 사용하였다.

필기체 한글은 주로 획으로 구성되어 있으므로, 문자 영상에서 추출된 방향 성분을 인식의 특징으로 사용하려는 연구가 있다. 이 방법은 문자의 통계적 특징에 의하여 생성된 참조 패턴들과의 거리 비교를 수행하여 후보 문자 집합을 생성한 후, 신경망을 이용한 상세 분류를 수행한다. 이때 대분류 단계에서는 DSF(directional segment feature)를 사용하고, 상세 분류 단계에서는 DCF(directional contributivity feature)를 사용하였다[20].

네오코그니트론을 인식기로 사용한 연구[21]에서는 입력 문자 영상에서 특징 패턴을 직접 추출하기 때문에 별도의 특징 추출 과정이 없다.

### 2. 형태 정규화 기법

형태 정규화는 다양한 크기와 모양을 갖는 문자 영상을 일정한 크기와 모양을 갖는 영상으로 변환시키는 과정이다. 특히 필기체 문자와 같이 형태의 변형이 매우 심한 경우에는 이러한 형태 정규화 과정을 통해 형태 변형을 보상함으로서 특징 추출 및 인식의 효율을 향상시킬 수 있다.

필기체 문자에서 발생하는 형태 왜곡을 효과적으로 보상하기 위한 다양한 비선형 형태 정규화 기법들에 대한 체계적인 비교 연구[22]가 있으며, 이러한 정규화 기법들을 이진 영상이 아닌 명도 영상에 직접 적용하여 문자 영상의 정보 손실을 피하는 방법도 제안되었다[23].

### 3. 한글 낱자 인식 기법

패턴 정합 방법에 기초한 한글 인식 방법은 대부분 획 정합 방법을 사용한다. 일례로, 문자 영상에 대한 외곽선 추적과 세선화 과정을 거쳐 외곽선의 방향 성분 분포, 골격선의 방향 성분 분포, 구조적 특징점 분포 등의 특징을 추출한 후, 추출된 획의 방향과 길이에 대한 중점 분포 특징을 이용하여 획 정합을 수행하는 방법이 있다[17].

문자의 전역적인 변형을 흡수하기 위하여 이단계의 패턴 정합 방법을 수행하는 방법[18]도 있다. 이 방법은 비선형 형태 정규화를 수행한 문자 영상에 대하여 국부적인 변형을 흡수하기 위해 입력 패턴과 표준 패턴 사이의 비선형 패턴 정합을 수행한 후 상세 분류를 통해 최종 결과를 결정한다.

패턴 정합 방법은 입력 패턴과 참조 모델의 비교에 의해 문자를 인식하므로 참조 모델의 설계가 인식 성능에 큰 영향을 준다. 따라서, 대용량 필기체 문자의 인식에 적합한 최적의 참조 모델을 설계하기 위해 LVQ3 알고리즘을 개선하고, 개선된 LVQ3 알고리즘을 시뮬레이티드 어닐링과 결합하여 LVQ의 단점을 보완한 연구[24]가 있다.

은닉 마르코프 모델을 이용하여 필기체 한글을 인식하는 방법[19]이 있다. 은닉 마르코프 모델은 학습 과정을 통해 대상 패턴의 특징들을 확률값으로 표현하는 모델로서 학습 과정과 분석 과정이 수학적으로 명료하고 신뢰성 있는 확률의 추정이 용이하므로 문자 인식뿐만 아니라 음성 인식 분야에서도 널리 응용되고 있다. 제안된 방법은 하나의 입력 문자 패턴에 대해 영역 투영 외곽선 변환을 적용하여 4종류의 영역 투영 외곽선을 추출한 후, 이들 외곽선에 대한 방향 성분을 이용하여 4개의 은닉 마르코프 모델을 구성한다. 또한 효율적인 인식 시스템의 구성을 위해 은닉 마르코프 모델의 매개변수에 몇 가지 제약을 가함으로서 불필요한 매개 변수의 추정을 피하였으며, 퍼지 트리 분류기를 사용하여 전반적인 처리 속도를 개선하였다.

네오코그니트론을 이용한 필기체 한글 인식 방법[21]은 별도의 자소 추출 과정 없이 네오코그니트론의 선택적 주의 기능을 이용하여 자소를 인식하였다.

### 4. 대용량 필기체 한글 데이터베이스

인쇄체 한글과는 달리 사람이 직접 필기한 한글 글씨는 동일한 문자라 하더라도 필기자의 필기 유형에 따른 다양한 변형을 가지고 있다. 따라서 필기체 한글 인식에 관한 체계적인 연구를 위해서는 다양한 변형을 포함하는 대용량의 한글 글씨 데이타베이스가 필요하다.

포항공과대학교에서 구축한 PE92 필기체 한글

영상 데이터베이스[25]는 KSC 완성형 2,350자 100벌로 구성되어 있으며, 각 글자는 100x100 크기의 256 명도 영상으로 되어 있다.

고려대학교에서 구축중인 필기체 한글 데이터베이스, KU-1[26]은 KSC 완성형 한글 중에서 사용 빈도순 상위 1,500자 1,000벌로 구성되어 있으며, 각 글자는 100x100 크기 256이내의 명도 영상으로 되어 있다. 그리고 다양한 데이터의 수집을 위하여 각종 수집 용지와 필기 도구를 사용해서 여러 분야에서의 필기자를 대상으로 데이터를 수집하였다.

### 5. 필기체 한글 인식의 문제점 및 향후 연구 방향

지금까지 필기체 한글 인식에 관한 최근의 연구 동향을 살펴보았다. 필기체 문자는 인쇄체와는 달리 동일 문자에 대한 변형이 복잡하고 글자간의 접촉 패턴이 매우 다양하다. 따라서 필기체 문자의 다양한 형태 변형을 보상하기 위한 특징 추출 및 형태 정규화에 관한 연구가 꾸준히 진행되고 있으며, 변형에 강한 인식 방법에 관한 연구도 활발히 이루어지고 있다.

특히, 최근 시행되고 있는 필기체 한글 데이터베이스의 구축은 필기체 한글 인식에 관한 체계적인 연구를 촉진시킬 것으로 예상되며, 다양한 접촉 패턴을 갖는 필기체 문자열의 낱자 분할에 관한 과학적인 연구를 위하여 문자열 데이터베이스의 구축도 필요할 것이다.

## IV. 문서 구조 분석 및 이해

기존의 종이 문서에 보관된 대량의 정보를 컴퓨터에 자동으로 입력하고자 하는 문서 인식에 관한 오랜 연구는 멀티미디어 시대의 도래와 함께 한층 발전된 형태의 문제 해결을 모색하고 있다. 과거의 초보적인 문서 인식 기술에서 문서 내에 포함된 그림이나 도표는 문자를 정확히 추출하는데 하나의 걸림돌에 불과하였으나, 최근에는 도표와 그래프의 인식을 위한 연구가 꾸준히 이루어지고 있으며, 문자는 물론 문서의 구조 정보를 최대한 보존하기 위하여 문서 인식 결과를 단순한 텍스트 문서가 아닌 멀티미디어 전자 문서의 형태로 저장하려는 연구가 활발히 진행되고 있다.

일례로, 일상 생활에서 쏟아지는 대량의 종이 문서들을 바탕으로 생산되는 멀티미디어 전자 백과사전이나 전자 신문 등의 출판 산업은 문서에 포함된 문자뿐만 아니라 다양한 그림과 도표를 인식할 수 있는 진보된 문서 인식 기술에 의하여 실질적인 생산 비용의 절감 효과를 기대할 수 있다.

한편, 문서 인식을 위한 문자 추출의 전단계로만 여겨졌던 문서 구조 분석 기술은 최근들어 독자적인 응용 분야를 확보하게 될만큼 그 역할의 중요성이 더욱 커지고 있다.

일례로, 대량의 문서 정보를 영상의 형태로 저장하고 있는 광화일 문서에서 필요한 정보를 신속하게 검색하기 위한 목적으로 문자 인식 기술의 도움없이 해당 문서의 제목이나 주요 핵심 문장을 고속으로 추출하는 기술이나, 스캐너와 팩스 등을 통하여 입력된 문서 영상에서 잡영을 제거하고 기울어짐과 삐뚤어짐 등의 문서 왜곡을 교정하여 문서의 가독성을 높이는 문서 영상 향상 기술 등은 문자 인식 기술과 결합되지 않고도 문서 처리 및 분석 기술이 매우 다양한 분야에서 직접적으로 응용될 수 있음을 보여주는 좋은 예라고 할 수 있다.

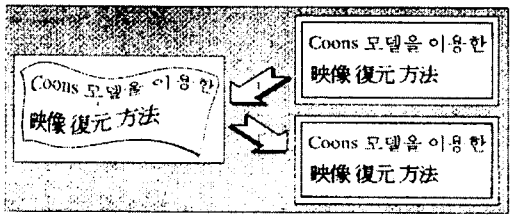
일반적인 문서의 구조 분석은 문서 영상에 포함된 잡영을 제거하고 문서의 입력 과정에서 발생하는 다양한 형태적인 변형을 보상하는 전처리 단계, 문서 영상의 물리적 및 통계적 특성을 바탕으로 각 영역을 분할하고 그림, 도표, 문단 등을 분별하는 영역 분할 단계로 나누어진다.

### 1. 문서 영상에서의 변형 복원

스캐너나 팩스와 같은 문서 입력 장치를 통해 입력된 문서 영상은 광학 센서나 롤러의 기계적인 결함으로 인하여 복잡한 영상 왜곡이 발생할 수 있으며, 디지털 카메라 등으로 입력된 문서 영상은 종이 자체의 3차원적인 휘어짐으로 인하여 비선형적인 변형이 발생하기 쉽다.

이러한 비선형적인 변형 영상의 복원 문제를 해

결하기 위하여 두가지 접근 방법을 제시한 연구가 있다[27]. 첫 번째 방법은 영상을 복원하기 위한 역변환 함수의 근사 함수를 찾기 위하여 Coons 변환 함수를 이용하는 방법으로서 그림 3에서와 같이 비선형 변형된 영상에 대한 변환 함수 T의 근사 함수를 찾기 위하여 문서의 경계에서 B-Spline 곡선 근사법을 적용하였고, 이를 Coons 변환에 적용하여 근사 함수 T*를 구하였다. 두 번째 방법은 분할-통합 방법을 이용하여 복잡한 비선형 방정식을 풀지 않고도 변형된 영상을 빠르게 복원할 수 있음을 보였다.




〈그림 3〉 Coons 모델을 이용한 비선형 영상 복원

## 2. 일반적인 문서 영상의 구조 분석

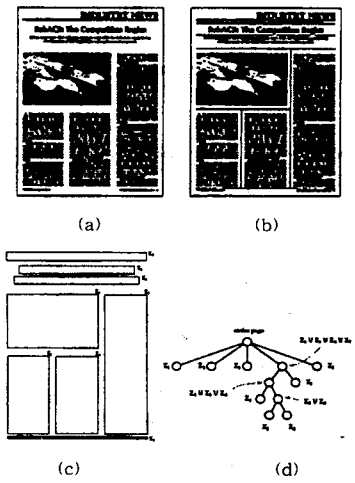
문서의 영역 분할 방법은 크게 상향식 방법과 하향식 방법으로 나누어진다. 상향식 방법은 문서의 기본 요소에서 시작하여 동일 특성을 갖는 요소들을 병합하면서 점점 영역을 확장하는 방법으로, 대표적으로 연결 요소 분석 방법을 사용한다.

하향식 방법은 전체 문서에서부터 단계적으로 분할하면서 각 세부 영역을 구분하는 방법이다.

대표적인 하향식 방법은 그림 4와 같이 문서의 단과 행을 고속으로 분할하기 위하여 하향식 문서 분할 기법으로 알려진 재귀적 X-Y 분할 방법을 사용하는 방법이 있으며, 문서의 구조 분석 속도를 향상시키기 위하여 화소 단위가 아닌 연결 요소 단위의 투영을 수행하는 방법[28,29]도 있다.

## 3. 서식 문서 영상에 대한 구조 분석 및 이해

다양하고 복잡한 형태를 갖는 일반 문서와는 달리 서식 문서는 은행 전표나 영수증 등과 같이 일




a) 입력 문서 영상, b) cut의 위치,
c) recursive X-Y cut을 이용한 영역 분할,
d) 입력 문서 영상에 대한 X-Y 트리

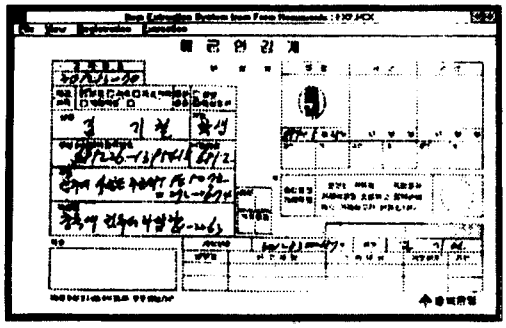
〈그림 4〉 X-Y cut을 이용한 문서 구조 분석

정한 형태를 갖고 있으며 대량 처리가 요구되는 기업 환경에서 주로 사용되기 때문에 오래전부터 서식 문서 인식에 관한 연구가 활발히 이루어졌다.

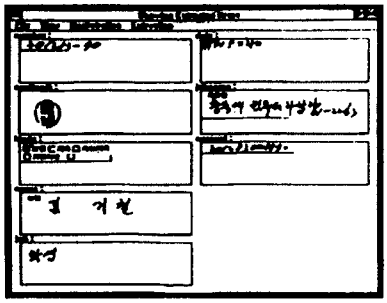
특히, 서식 문서의 인식은 주어진 환경에서 얻을 수 있는 특정 정보를 활용하는 후처리가 가능하기 때문에 보편적인 문서의 인식보다 문제가 비교적 단순하다고 할 수 있다.

일반적인 서식 문서의 인식 과정은 전처리 단계, 서식의 판별 단계, 항목의 추출 단계, 문자 인식 단계로 나누어진다. 전처리 단계에서는 서식 문서의 입력 중에 발생하는 영상의 기울어짐을 교정하고 잡영을 제거하며 서식의 크기와 형태를 정규화한다. 서식의 판별은 두 종류 이상의 서식을 처리해야 하는 경우에 수행된다. 항목의 추출 단계에서는 빈 서식과의 비교 방법이나 도표 분석 방법, 고정 영역 추출 방법 등을 통하여 사람이 기입한 항목들을 추출한다. 마지막의 문자 인식 단계에서는 사람이 기입한 필기 문자를 인식하는 단계로서 인식 대상 항목의 문맥 정보를 활용하여 인식 성능을 향상시킬 수가 있다. 일례로 해당 항목이 주소

나 성명 또는 숫자 중 특정값으로만 입력될 수 있다는 사전 지식이 있다면 이러한 정보를 이용하여 문자 인식기의 오류를 줄일 수가 있다.



(a) 입력 서식 문서 영상



(b) 자동 추출된 항목 영상

〈그림 5〉 서식 문서에서의 항목 추출

대표적인 예로 그림 5는 서식 문서로부터 항목을 자동으로 추출하는 과정을 보여주고 있다[30].

이 방법은 백화소 연결 요소 추출 방법을 이용하여 빈 서식과 채워진 서식에서의 각 항목을 자동으로 추출한 다음, 등록된 각 항목에 대응하는 후보 항목을 대상으로 최소 거리 척도를 이용하여 최종적인 항목을 결정한다.

### 4. 문서 구조 분석 기술의 문제점 및 향후 연구 방향

그림이나 도표 등이 포함되어 있는 일반 문서들을 효과적으로 분석하고, 문단의 순서와 구조적인 정보를 이해하는 문서 구조 분석 및 이해에 관한 연구는 아직까지 많은 연구 과제를 안고 있다.

특히, 이러한 문서 구조 분석 및 이해 기술은 앞에서 소개한 문자 인식 기술의 보편적인 실용화를 위하여 반드시 해결해야할 과제이기도 하다.

## V. 결 론

본 논문에서는 한글 문서의 분석 및 인식에 관한 최근의 연구 동향을 살펴보았다.

인쇄체 한글의 인식에 관한 연구는 다중 활자체 및 다중 언어에 대한 인식의 문제로 연구 범위가 확대되고 있으며, 이에 따라 문자 인식 기술의 적용 범위도 다양해질 것으로 예상된다.

필기체 한글의 인식에 관한 연구는 표준화된 한글 글씨 데이터베이스의 구축을 계기로 보다 활발하고 체계적인 연구가 이루어질 수 있을 것으로 기대되며, 필기체의 변형을 효과적으로 보상할 수 있는 형태 정규화와 특징 추출 방법의 개발 및 인식 기술의 발전이 꾸준히 이루어져야 할 것이다.

한편, 최근 해외에서 꾸준히 발표되고 있는 다양한 문서 구조 분석 및 이해에 관한 연구 결과들은 국내의 문서 인식에 관한 연구를 더욱 촉진시키는 계기가 될 것으로 기대되며, 문서 구조 분석 정보를 표현하는 양식의 표준으로서 자리를 잡고 있는 DAFS(Document Attribute format Specification)[31]의 등장과 인터넷 문서 사용량의 증가는 문서 인식 기술의 응용 범위를 넓히는 데 커다란 역할을 할 것으로 전망된다.

### 부 록

문자 인식이나 문서 구조 분석에 관한 보다 자세한 자료는 다음과 같은 관련 분야의 권위있는 국제 학술지나 학술대회, 워크샵 프로시딩 등을 참고할 수 있다. 이들은 각기 인쇄체 문자 인식, 필기체 문자 인식, 문서 구조 분석, 지도 및 도안의 인식 등과 같이 문서 인식에 관한 여러 가지 주제

들을 다루고 있다.

· IJDAR(International Journal of Document Analysis and Recognition)

1998년 초부터 Springer 출판사를 통하여 발간되는 문서 분석 및 인식에 관한 국제 학술지이다.

*http : //documents.cfar.umd.edu/IJDAR/*

· ICDAR(International Conference on Document Analysis and Recognition)

문서 분석 및 인식에 관한 가장 큰 규모의 국제 학술 대회이며, 제 5회 학술 대회가 1999년 10월에 인도의 Bangalore에서 개최될 예정이다.

*http : //www.cedar.buffalo.edu/icdar99/*

· IWDAS(International Workshop on Document Analysis Systems)

문서 인식에 관한 국제 워크샵으로 IAPR (International Association for Pattern Recognition)의 후원을 받고 있으며, 제 3회 워크샵이 1998년 11월에 일본의 Nagano에서 개최될 예정이다.

*http : //www.korea.ac.kr/DAS98/*

· IWFHR(International Workshop on Frontiers in Handwriting Recognition)

필기체 문자 인식에 관한 국제 워크샵이며, 제 6회 워크샵이 1998년 8월에 대전에서 개최될 예정이다.

*http : //ai.kaist.ac.kr/iwfhr98/*

## 감사의 말씀

본 논문은 시스템공학연구소 자연어정보처리연구부의 연구비 지원을 받았음.

## 참 고 문 헌

〔1〕 이 성환, 문자 인식 : 이론과 실제, I, II권, 홍릉과학출판사, 1994년 4월.

〔2〕 진 성일, 백 영목, 임 길택, "다국어 문서 이해에 대한 최근 연구 동향," 제 9회 영상 처리 및 이해에 관한 워크샵 발표 논문집, 1997년 1월, pp. 101－107.

〔3〕 이 성환, "다양한 활자체 및 크기를 갖는 대용량 한글의 고속 인식을 위한 최적 트리 분류기," 한국정보과학회 논문지, 제 20권 제 8호, 1993년 8월, pp. 1083－1092.

〔4〕 D. Zongker and A. Jain, 'Algorithms for Feature Selection － An Evaluation," Proc. Int. Conf. on Pattern Recognition, Vol. 2, 1996, pp. 18－22.

〔5〕 S.－W. Lee and Y.－J. Kim, 'Direct Extraction of Topographic Features from Gray Scale Character Images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 17, No. 7, 1995, pp. 724－729.

〔6〕 S.－W. Lee, D.－J. Lee and H.－S. Park, 'A New Methodology for Gray－Scale Character Segmentation and Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 18, No. 10, 1996, pp. 1045－1050.

〔7〕 이 성환, "퍼지 트리 분류기를 이용한 인쇄 상태가 나쁜 한글의 인식," 한국정보과학회 논문지, 제 20권 제 11호, 1993년 11월, pp. 1609－1618.

〔8〕 이 진수, 권 오준, 방 승양, "개선된 자소 인식 방법을 통한 고인식률 인쇄체 한글 인식," 한국정보과학회 논문지, 제 23권 8호, 1996년 8월, pp. 841－851.

〔9〕 S.－W. Lee and J.－S. Kim, 'Multi－lingual, Multi－font and Multi－size Large－set Character Recognition Using Self－Organizing Neural Network," Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 28－33.

〔10〕 김 두식, 이 성환, "저해상도 인쇄체 한글 인식을 위한 자소 기반 방법과 음절 기반 방법의 성능 비교," 한국정보과학회 가을 학술발표논문집, 용인, 제 23권 제 2호, 1996년 10월, pp. 587－590.

〔11〕 김 은주, 백 종현, 변 혜란, 이 일병, "유전자 알고리듬에 의한 다중인식기의 결합," 제 5회 인공지능, 신경망 및 퍼지시스템 종합학술대회 발표 논문집, 서울, 1996년 10월, pp. 143－146.

〔12〕 권 재욱, 조 성배, 김 진형, "계층적 신경망을 이용한 다중 크기의 다중활자체 한글문서 인식," 한국정보과학회 논문지, 제 19권 제 1호, 1992년 1월, pp. 69－78.

〔13〕 C. Y. Suen and Q. R. Wang, 'ISOETRP : an interactive clustering algorithm with new objects," Pattern Recognition, Vol. 7, No. 4, 1984, pp. 211－219.

〔14〕 배 진학, 박 세현, 김 항준, "영 숫자 한글 문서에서 문자 분리 및 인식," 한국정보과학회 논문지(B), 제 23권 제 9호, 1996년 9월, pp. 941－949.

〔15〕 이 성환, "오프라인 필기 인식 기술의 연구 현황," 제 2회 문자 인식 워크샵 발표 논문집, 1994년 9월, pp. 3－37.

〔16〕 오 균, 조 성배, 이 일병, "문서영상 이진화 알고리즘에 대한 체계적인 평가," 제 8회 영상처리 및 이해에 관한 워크샵 발표논문집, 부산, 1996년 1월, pp. 115－120.

〔17〕 김 기철, 이 성환, "필기체 한글의 오프라인 인식을 위한 획 정합 방법," 대한전자공학회 논문지, 제 30권 B편 제 6호, 1993년 6월, pp. 604－613.

〔18〕 박 정선, 이 성환, "필기체 한글의 오프라인 인식을 위한 효과적인 두 단계 패턴 정합 방법," 대한전자공학회 논문지, 제 31권 B편 4호, 1994년 4월, pp. 351－358.

〔19〕 박 희선, 이 성환, "은닉 마르코프 모델을 이용한 필기체 한글의 오프라인 인식," 한국정보과학회 논문지, 제 20권 제 6호, 1993년 6월, pp. 890－902.

〔20〕 S. H. Kim and J.－I. Doh, 'Off－line Recognition of Korean Scripts Using Distance Matching and Neural Network Classifiers," Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 34－37.

〔21〕 김 태우, 김 승종, 김 재훈, 민 병석, 김 한우, 최 병욱, "개선된 네오코그니트론을 이용한 필기체 한글 인식 앨고리듬," 제 2회 문자 인식 워크샵 발표 논문집, 1994년 9월, pp. 119－132.

〔22〕 S.－W. Lee and J.－S. Park, 'Quantitative Evaluation of Nonlinear Shape Normalization Methods for Large－set Handwritten Character Recognition," Pattern Recognition, Vol. 27, No. 7, 1994, pp. 895－902.

〔23〕 S.－Y. Kim and S.－W. Lee, 'Nonlinear Shape Normalization Methods for Gray－Scale Handwritten Character Recognition," Proc. Int. Conf. on Document Analysis and Recognition, Ulm, Germany, August 1997, pp. 479－482.

〔24〕 H.－H. Song and S.－W. Lee, 'LVQ Combined with Simulated Annealing for Optimal Design of Large－set Reference Models," Neural Networks, Vol. 9, No. 2, 1996, pp. 392－336.

〔25〕 D. H. Kim et al., 'Handwritten Korean Character Image Database PE92," Proc. 2nd Int. Conf. on Document Analysis and Recognition, Tsukuba, Japan, Oct. 1993, pp. 470－473.

〔26〕 김 대인, 김 상엽, 이 성환, "대용량 오프라인 한글 글씨 영상 데이터베이스, KU－1의 설계 및 구축," 제 9회 한글 및 한국어 정보처리 학술발표 논문집, 부산, 1997년 10월(게재 예정).

[27] S.-W. Lee, E.-S. Kim and Yuan Y. Tang, 'Nonlinear Shape Restoration of Distorted Images with Coons Transformation," Pattern Recognition, Vol. 29, No. 2, 1996, pp. 217-229.
[28] J. Ha et al., 'Recursive X-Y Cut using Bounding Boxes of Connected Components," Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 952-955.
[29] J. Ha et al., 'Document Page Decomposition by the Bounding-Box Projection Technique," Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 1119-1122.
[30] 김 기철, 이 성환, "서식 문서 영상의 구조 분석," 한국정보과학회 논문지, 제 22권 제 1호, 1995년 1월, pp. 182-192.
[31] J. Liang et al., 'The Prototype of a Complete Document Image Understanding System," Proc. Int. Workshop. on Document Analysis ystems, Malvern, USA, October 1996, pp. 131-154.

## 저 자 소 개

**金 斗 植**

1971年 10月 16日生
1995年 8月 고려대학교 전산과학과 학사
1997年 8月 고려대학교 전산과학과 석사
1997年 9月 고려대학교 대학원 컴퓨터학과 박사과정 재학중

관심분야 : 영상처리, 패턴인식, 신경망 등

**金 相 燁**

1971年 3月 20日生
1994年 고려대학교 정보공학과 학사
1996年 고려대학교 정보공학과 석사
1997年~현재 고려대학교 대학원 영상정보처리학과 박사과정 재학중

관심분야 : 문자인식, 신경망, 패턴인식, 영상처리 등

**李 晟 瑍**

1962年 6月 2日生
1984年 서울대학교 계산통계학과 학사
1986年 한국과학기술원 전산학과 석사
1989年 한국과학기술원 전산학과 박사

1989年～1994年 충북대학교 컴퓨터과학과 조교수
1995年～현재 고려대학교 컴퓨터학과, 영상정보처리학과 부교수

관심분야 : 패턴 인식, 컴퓨터 비젼, 신경망 등