자소 클래스 인식에 의한 off-line 필기체 한글 문자 분할

황 순 자† · 김 문 현†

요 약

문자 분할은 필기체 문서 서식의 자동 인식 과정에서 중요한 부분이다. 본 연구는 Off-Line 필기체 한글로부터 문자를 분할하기 위한 방법을 제안한다. 제안한 방법은 한글의 구조적 특성에 기반을 두고 있다. 먼저 투영에 의하여 입력 단어로부터 분할을 위한 특징과 연결 화소, 획을 추출한다. 두 번째 단계에서 획의 모양과 위치, 획과 획과의 관계를 이용하여 한글의 기본 자소 클래스 영역을 찾는다. 세 번째 단계는 분할 과정으로 WRC (White Run Column) 다음에 초성이나 수평 모음이 오는 경우 이 WRC에서 수직으로 분할하며, 분할된 세그먼트의 길이가 임계값 이상이면 자소 클래스와 문자의 칼럼에 대한 특징을 이용하여 예상 분할 영역을 찾고, 이 영역에 있는 획을 따라 요철 형태로 분할한다.

Consonant-Vowel Classification Based Segmentation Technique for Handwritten Off-Line Hangul

Soon Ja Hwang† · Moon Hyun Kim†

ABSTRACT

The segmentation of characters is an important step in the automatic recognition of handwritten text. This paper proposes the segmenting method of off-line handwritten Hangul. The suggested approach is based on the structural characteristics of Hangul. The first step extracts the local features, connected component and strokes from the input word. In the second step we identify the class of strokes. The third segmenting step specifies WRC(White Run Column) before consonant or horizontal vowel. If the segment is longer than threshold, the system estimates segmenting columns using the consonant-vowel information and column features, and then finds a cornered boundary along the strokes within the estimated segmenting columns.

1. 서 론

인간의 정보 전달 방법 중 가장 큰 비중을 차지하는 것은 문서 형태에 의한 것이다. 최근에 와서 컴퓨터를 이용한 정보 전달이 이루어지고 있으나 아직도 서면을 이용한 방법이 많이 사용된다. 그러나 서면으로 작성된 문서는 사람이 컴퓨터에 입력하는 과정을 거치게 되며 많은 노력을 요구하게 된다. 이에 따라 자동 문서 인식 시스템의 필요성이 대두되고 있으며 최근 들어 자동 문서 인식에 관한 연구가 국내외적으로 활발히 진행되어 왔다[1][2]. 그러나 지금까지 대부분의 국내 연구는 인쇄체 한글의 인식에 관한 연구였으며, 최근 off-line 필기체 인식에 관한 연구가 시작되고 있다. 그러나 대부분의 연구가 낱자 인식에 집중

† 정 회 원:성균관 대학교 정보공학과
논문접수:1996년 2월 29일, 심사완료:1996년 5월 1일

되고 있다. off-line 필기체 문자 인식에서 문자열로부터 개별 문자로 분할하는 문자 분할 과정은 반드시 필요하게 되며, 이는 필기체 한글 자동 인식 시스템의 실용화를 위한 중요한 과제이다. 필기체 영문자나 숫자의 분할에 관한 연구는 많으나, 한글의 경우 인쇄체 분할에 관한 연구가 발표되었을 뿐이다.

한글 분할을 위하여 흔히 쓰이는 투영 방법이나 연결 화소의 분석에 의한 방법은 모두 문자 폭에 크게 의존하기 때문에 문자 폭의 변화가 큰 필기체에 직접 적용하기가 어렵다. 또한 필기체 문자열에서 문자들은 흔히 겹치거나 접촉되기 때문에 기존의 수직 분할 방법으로는 이들 문자들을 정확하게 분할할 수가 없다. 필기체 영문자나 숫자의 분할에 많이 쓰이는 방법으로 외곽선 추적에 의한 문자 분할 방법과 구조적 특징을 이용한 방법이 있다. 그러나 한글은 자소의 모아쓰기 형태를 갖고 있어 문자 내에 여백이 존재할 뿐만 아니라 문자와 문자 사이의 접촉에서 나타나는 특징이 자소내 또는 자소 결합에서 흔히 발생하기 때문에 영 숫자에 사용된 방법의 직접 적용이 어렵다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 자소 클래스 인식에 기반한 분할 알고리즘을 제안하였다. 특히 한 번의 수직 투영으로 칼럼에 대한 특징들을 추출하며, 동시에 연결 화소와 자소를 구성하는 부분획을 추출한 후 이러한 특징들로부터 자소 클래스 영역을 추정하였다. 분할 단계에서는 한글의 구성 법칙과 경험적 지식을 이용한다. 우선 WRC 다음 초성이나 수평 모음이 오면 문자 사이의 간격으로 판단하여 수직으로 분할을 한다. 다음 겹치거나 접촉된 문자들에 대하여 자소 클래스 영역에 대한 정보를 이용하여 예상 분할 영역을 추정하고, 예상 분할 영역 내에 있는 획들의 최소 외접 사각형을 따라 분할함으로써 문자의 형태에 따른 분할을 수행한다. 따라서 겹치거나 접촉된 문자들을 분할하기 위한 외곽선 추적 등의 과정이 별도로 요구되지 않는다. 제안한 알고리즘은 획의 자소 클래스에 기반하기 때문에 문자의 크기, 필기 형태, 잡음 등에 강하다.

제 2장에서 기존의 관련된 연구들을 살펴보고, 3장에서 제안한 off-line 필기체 문자 분할 시스템에 관하여 자세히 설명한다. 4장에서 실험 결과를 기술하고, 끝으로 5장에서 결론과 앞으로의 연구 과제를 제시하고자 한다.

2. 관련 연구

문자 분할 방법은 크게 직접 분할 기법과 분할-인식 기법으로 나눈다.

2.1 직접 분할 기법

직접 분할 방법은 인식기와 상관없이 문자들을 분할하는 방법으로 속도가 빠른 장점을 가지고 있지만 접촉된 문자들에 대하여 정확한 분할점을 찾는 것이 어려우며, 오분할된 경우 인식률의 저하를 가져온다는 단점이 있다. 크게 수직 투영에 기반한 방법, 외곽선 추적에 의한 방법, 구조적 특징을 이용한 방법 등이 있다.

2.1.1 수직 투영에 기반한 문자 분할

수직 투영에 의하여 얻은 투영값 V(x)가 일반적으로 문자 내에서보다 문자 사이에서 작다는 점을 이용한다[3)4)5)]. 인쇄체 문자 분할에 많이 쓰이는 방법으로, 필기체 영문자나 숫자의 분할을 위하여 다른 방법과 병행하여 많이 쓰인다. (그림 1)은 권 재욱[6)] 등에 의한 인쇄체 문자 분할 알고리즘을 보여 준다. 수직 투영에 의하여 문자열을 백런 영역과 흑런 영역으로 나눈 후, 문자폭을 기반으로한 문자 결합 신경망과 문자 분리 신경망에 의하여 분리된 문자를 결합시키고 결합된 문자들을 분리시킴으로써 개별 문자를 추출한다.[7)]

(그림 1) 수직 투영에 의해 WR과 BR로 나뉘어진 문자열
(Fig. 1) String of WR and BR by Vertical Projection

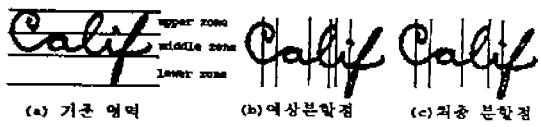
2.1.2 외곽선 추적에 의한 문자 분할

외곽선 추적에 의하여 얻은 국부적 최소치(local minimum) 또는 국부적 최대치(local maximum)에 의하여 예상 분할점을 찾고, 예상 문자 폭에 의하여 최종 분할점을 결정하는 방법이다.

(그림 2)는 Bozinovic와 Srihari의 외곽선 추적에 의한 문자 분할 방법을 보여 준다[8)]. 문자열을 상부, 중앙부, 하부 3개의 영역으로 나누고, 중앙부에서 lower

contour의 외곽선을 따라 국부적 최소치를 구하여, 수직 투영으로부터 얻은 투영값을 고려하여 예상 분할점을 찾는다. 다음 예상 문자폭을 추정하여 필요 없는 점들을 묶어 최종 분할점을 결정한다.

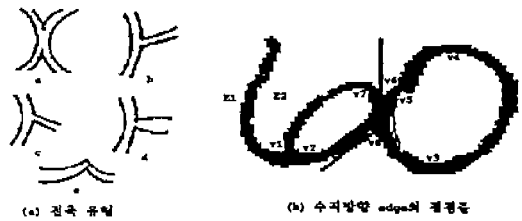
Beglou[9] 등은 단어에 대한 기준선을 정하고, upper contour와 lower contour 의 외곽선을 추적하는 방법을 제안하였다.



(그림 2) 외곽선 추적에 의한 영문자 분할
(Fig. 2) Segmentation of Characters by Contour Analysis

2.1.3 구조적 특징에 기반한 문자 분할

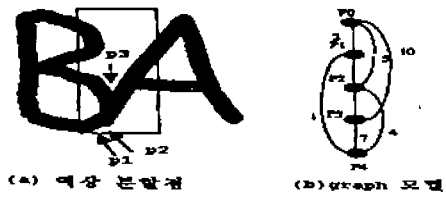
숫자나 문자 의 접촉 형태를 몇 가지 특징으로 분류하고의 접촉 유형을 찾아 분할하는 방법이다.(그림 3)에서와 같이 Westall[10] 등은 숫자의 접촉 형태를 5가지로 분류하고 그에 따른 분할 알고리즘을 제안하였다. Sethi[11] 등은 대략적인 분할 영역을 구하기 위해 각 알파벳을 구조적 특징에 따라 6가지 유형으로 구분한 후, 단어 내에 그러한 특징이 존재하면 각 알파벳이 존재하는 것으로 판단하여 분할 영역을 구한다. 전역적 정보에 의하여 단어 내의 문자수와 문자의 폭을 추정하여 분할하는 방법을 제안하였다. Lecolinet[12][13] 등은 문자 분할에 중요한 역할을 하는 문자의 일부분을 인식함으로써 문자의 위치를 추정하는 방법을 제안하였다. 그 이외에 연결 화소의 분석에 의하여 예상 분할점들을 찾은 후 예상 문자폭을 이용하여 분할하는 방법[16] 등이 발표되었다.



(그림 3) 구조적 특징에 기반한 문자분할
(Fig. 3) Character Segmentation by Structural Features

2.2 분할 인식 기법

분할-인식 기법은 투영값 및 윤곽선의 기하학적 특성을 이용하여 모든 분할 예상점을 찾고 일정한 폭을 갖는 사전 분할점들을 묶어 분할 및 인식을 반복적으로 수행하여 인식 결과를 보이는 방식으로 가능한 영역들을 모두 인식해 보아야 하기 때문에 시간이 많이 걸리는 단점이 있다. (그림 4)는 Casey와 Horn의 알고리즘을 보여준다[8]. 외곽선 추적에 의하여 국부적 오목점(locally concave point) p1, p2, p3을 찾아 예상 분할점으로 지정하고 인식기와 결합하여 분할 인식 과정을 되풀이한다.



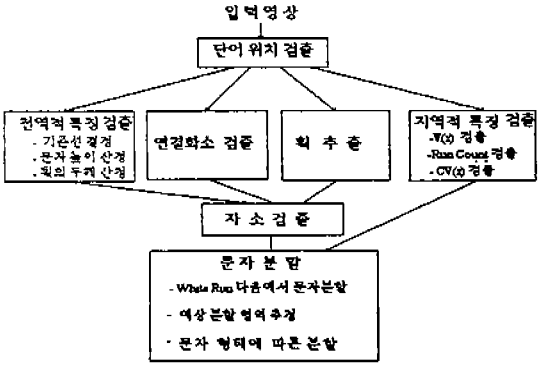
(그림 4) 분할 인식 기법
(Fig. 4) Segmentation-Recognition Method

김 두식[17] 등은 영·숫자가 혼용된 인쇄체 한글 분할을 위하여 수직 투영 방법으로 사전 분할점을 찾고 인식기와 결합하여 분할_인식을 반복 수행한다. 그 이외에 연결 화소 분석에 의하여 사전 분할점을 찾는 연구가 장명욱[18] 등에 의하여 발표되었다.

3. 분할 시스템의 구성

제안한 분할 알고리즘은 먼저 투영을 기반으로 문자 분할에 필요한 특징 및 연결 화소, 획을 추출하는 특징 추출 단계, 두 번째 자소의 기본 클래스인 초성, 수직 모음, 수평 모음, 종성의 부분 인식을 통하여 이들의 영역을 추출하는 자소 클래스 영역 추출 단계와 세 번째 문자 분할 단계로 구성된다. 특징 추출 단계에서 구하는 특징은 크게 전역적 특징, 지역적 특징, 연결 화소, 획으로 나눌 수 있다. 자소 클래스 영역 추출 단계는 추출된 획의 자소 클래스를 인식하고 그 영역을 추출한다. 분할 단계에서는 한글의 구성 법칙과 경험적 지식을 이용한다. WRC 다음 오른쪽에 초성이나 수평 모음이 오면 이 런은 문자 사이의 간격

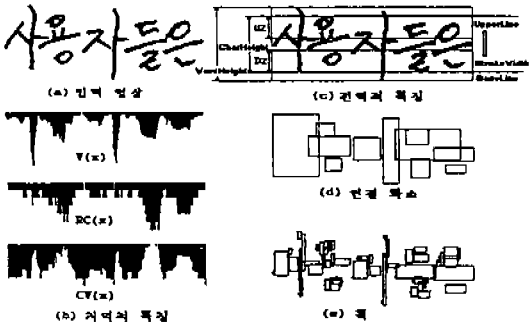
이므로 분할한다. 이 과정에서 겹치거나 접촉되지 않은 문자들이 검출된다. 다음은 세그먼트가 최소 문자폭*2보다 크면 1문자 이상을 포함할 수 있다고 가정하고 자소 클래스의 구성과 지역적 특징을 이용하여 예상 분할 영역을 찾고, 이 예상 분할 영역에 포함되는 획의 좌표를 이용하여 문자의 형태에 따른 요철형태의 분할을 수행한다. 필기체 문자 분할에서 인식과정 없이 정확한 분할점을 찾는 것은 그렇게 쉽지 않다. 특히 두 문자가 접촉된 경우 한 곳 이상에서 분할 가능한 경우가 많기 때문에 제안한 분할 시스템에서는 분할 확률이 가장 높은 점에서 분할하는 동시에 분할 후보점을 구한다. (그림 5)는 제안한 분할 시스템의 전체적인 구성을 나타낸다.



(그림 5) 시스템의 전체적인 구성
(Fig. 5) Diagram of Segmentation System

3.1 특징 추출

특징 추출 과정은 우선 수평 투영에 의하여 단어의 최소 사각형을 나타내는 좌표 (wSX, wSY, wEX, wEY)를 구하며 동시에 각 행에 대한 흑화소의 누적값 V(y)를 구한다. wSX와 wEX는 단어의 시작과 끝 열이며, wSY와 wEY는 처음과 마지막으로 흑화소가 나타나는 행의 좌표이다. 다음은 수직 투영으로 각 칼럼에 대한 특징을 구하며 동시에 연결 화소와 획을 구한다. 세선화 과정 없이 입력 영상으로부터 획을 추출함으로써 속도를 높일 수 있을 뿐만 아니라 분할에 필요한 정보의 손실이나 왜곡을 막을 수 있다. 특징 추출 단계에서 구하는 특징은 크게 지역적 특징, 전역적 특징, 연결 화소, 획으로 나눌 수 있다. (그림 6)은 수평 투영과 수직 투영으로부터 추출된 특징들이다.



(그림 6) 특징 추출 결과
(Fig. 6) Feature Extraction

3.1.1 전역적 특징

전역적 특징은 단어 전체에 대한 특징으로 자소 클래스의 분류를 위하여 사용된다. 필기체에서는 인쇄체와 달리 (그림 6a)에서와 같이 일부 획을 지나치게 높은 곳에서부터 시작하거나 아래까지 내려쓰는 경우가 많다. 이런 경우 단어의 시작행의 y좌표(wSY)로부터 단어의 끝행의 y좌표(wEY)를 기준으로 문자의 높이나 자소 클래스의 위치를 추정하기가 어렵다. 제안한 알고리즘에서는 수평 투영에서 얻은 흑화소의 누적값 V(y)를 이용하여 상한 기준선과 하한 기준선을 찾고, 문자의 기준 높이 및 상단 영역과 하단 영역을 추출하였다. 획의 두께는 획의 자소 클래스를 결정하기 위한 중요한 정보가 된다. 수직획이 기울어지지 않은 경우 획의 평균 런길이(RL)와 획의 높이(MaxY-MinY)가 같게 되므로 이 획의 폭(Width)이 획의 두께가 되어 기울어지지 않은 수직획들의 평균 폭으로부터 획의 두께를 구할 수 있다. 다음과 같이 전역적 특징들을 정의하며, (그림 6.c)는 이러한 특징들을 그림으로 나타낸 것으로 RC(x)는 런 횟수의 10배를 나타낸다.

상한 기준선(UpperLine): wSY로부터 V(y)가 최초로 MAXV(최대 흑화소의 누적값)*0.2 이상이 되는 행

하한 기준선(BaseLine) : wEY로부터 V(y)가 최초로 MAXV*0.2 이상이 되는 행

문자의 기준 높이(H_l) : BaseLine의 y좌표-UpperLine의 y좌표

상단 영역(UZ : UpperZone) : UpperLine부터 $0.4H_l$번째 되는 행까지의 영역

하단 영역(DZ : DownZone) : BaseLine부터 $0.4H_l$번째 되는 행까지의 영역

획 두께(StrokeWidth) : 높이가 $0.4H_l$이상이고 획의 RL이 $0.9H_l$ 이상인 수직획들의 평균폭

3.1.2 지역적 특징

지역적 특징은 각 열에 대한 특징으로 수직 투영에서 V(x)와 런 횟수 RC(x)를 구한다. 일반적으로 문자 내 열에서보다 문자 사이의 열에서 V(x)와 RC(x)는 작으므로 분할 가능성이 높다. 이런 특성을 이용하여 V(x)와 RC(x)로부터 추가로 분할 확률값 CV(x)를 계산하였다.

(그림 6b)는 추출된 지역적 특징들을 나타낸다.

V(x) : x열의 수직 투영에 의한 흑화소의 누적값

RC(x) : x열에서 백화소에서 흑화소로 바뀌는 부분의 수

$$CV(x) : 50 - \left\{ \frac{V(x)}{MAXV} \times 30 + \frac{RC(x)}{MAXRC} \times 40 \right\}$$

MAXV : 흑화소의 최대 누적값, MAXRC : 최대 런 횟수

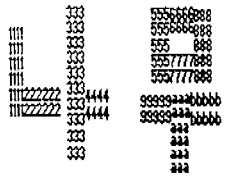
3.1.3 연결 화소 추출

김 의정의 수직 투영에 의한 연결 화소 검출 방법[7]을 이용하였다. 알고리즘은 영상의 왼쪽부터 열의 런을 추출하여 획을 구하는 과정에서 추출된 런의 왼쪽에 자신과 연결된 런이 있는지 검사하고 이전 런과 같은 연결 화소 인덱스(cIndex)를 가져야 하는지 또는 새로운 인덱스가 생성되는지 판단한다. 연결된 런이 두개 이상인 경우 가장 작은 인덱스 값을 부여하고 다른 연결된 런의 인덱스를 가지는 획의 연결 화소 인덱스를 현재의 인덱스로 바꾼다.

마지막 칼럼까지 투영이 끝나면 같은 연결 화소 인덱스를 가지는 획들로부터 각 연결 화소를 포함하는 최소 사각형을 구하여 그 정보를 저장한다. (그림 6d)는 투영에 의해 추출된 연결 화소의 블록을 나타낸다.

3.1.4 획 추출

수직 투영에서 지역적 특징과 연결 화소를 추출하며 동시에 비슷한 길이를 가지는 흑런을 하나의 블록으로 합성함으로써 자소를 구성하는 부분획을 추출할 수 있으며 본 논문에서는 이 부분획을 획이라고 정의한다. (그림 7)에서와 같이 같은 연결 화소인 두 런의 길이 변화가 임계값 이하이면 하나의 획으로 합성할 수 있으며 급격한 변화가 있으면 새로운 획을 생성한다. 그 결과 굴곡점이나 분기점에서 런길이가 급격히 변하여 새로운 획으로 검출되며 획에 굴곡점이나 분기점이 없는 경우 하나의 획으로 검출된다.



(그림 7) 런 길이 변화에 의한 획 추출
(Fig. 7) Stroke Extraction by Runlength

(그림 6e)는 런길이 변화를 이용한 획 추출 결과이다. 획에는 모든 런에 대한 정보를 저장하는 대신 메모리를 줄이기 위해 획의 위치를 최소사각형으로 나타낼 수 있는 최대 최소 x, y 좌표와 이웃 획과의 접촉을 검사하기 위한 첫 번째 런과 마지막 런의 최대 최소 y 좌표, 획의 폭과 평균 런길이를 저장한다.

3.2 자소 클래스의 영역 추출

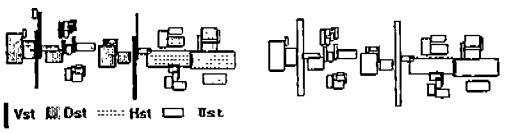
본 논문에서는 자소 클래스를 초성, 수직 모음, 수평 모음, 종성 4개의 클래스로 구분하고 이들에 기반한 분할 알고리즘을 개발하였다. 자소 클래스 인식을 위하여 방해가 되는 획들을 제거하거나 기울어짐에 따라 두 개의 획으로 나뉜 획들을 합성하였다. 획의 형태와 위치, 다른 획과의 관계를 고려하여 획의 자소 클래스를 인식하고 이들의 영역을 추출하였다.

3.2.1 작은 획의 제거 및 합성

작은 획들은 문자 분할 영역 추정에 큰 정보를 주지 못할 뿐만 아니라, 획의 자소 클래스 인식을 어렵게 하기 때문에 제거하거나 이웃 획에 포함시켰다. 이러한 획들의 제거를 위하여 임계값 이하인 작은 획의 한쪽에 다른 획이 연결되지 않은 경우 이 획은 제거하며, 양쪽에 다른 획이 연결된 경우 획의 길이의 차가 적은 쪽인 획에 합성시킨다.

획의 성분은 획의 자소 클래스를 결정하는 데 중요한 정보가 된다. 제안한 알고리즘에서는 획의 성분을 폭과 높이에 따라 수직획(Vst), 사선획(Dst), 수평획(Hst)으로 분류하였다. (그림 8a)는 입력 영상에 대한 획 추출 결과인 (그림 6e)로부터 중요하지 않은 작은 획들을 제거하거나 합성시킨 후 획의 성분을 분류한 결과이다.

문자 분할에서 수직 모음은 특히 중요한 역할을 한다. 그러나 (그림 8a)에서와 같이 필기자에 따라 기울여 쓰거나 휘어진 경우 하나의 수직 모음이 두 개의 접촉된 획으로 검출되는 경우가 많다. 경필 쓰기체에서는 수직 모음 상단에서 검출된 획이 어느 정도 크기 때문에 앞의 획 제거 알고리즘에서 제거되지 못한다. 이러한 현상들은 수직 모음의 인식을 어렵게 하므로 수직획에 연결된 획의 길이, 폭, 위치 등을 고려하여 합성하거나 제거하였다. (그림 8b)는 수직획에 연결된 획의 제거 및 합성 알고리즘을 적용한 결과이다.

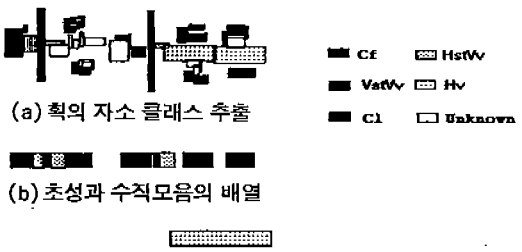


(a) 획의 분류 (b) 수직획에 연결된 획의 제거 및 합성

(그림 8) 획의 분류, 제거 및 합성
(Fig. 8) Classification, Exclusion and Composition of Strokes

3.2.2 자소 클래스의 인식

본 논문에서는 자소 클래스를 초성(Cf), 수직 모음(Vv), 수평 모음(Hv), 종성(Cl)으로 나누었다. 문자 분할에서는 인식에서와 달리 자소를 이루는 모든 획을 인식할 필요 없이 자소 클래스가 위치한 범위만 알면 된다. 제안한 알고리즘에서는 자소를 이루는 중요 획을 인식함으로써 각 자소 클래스의 위치를 찾는다. (그림 9)는 획의 자소 클래스 인식과 영역 추출 결과이다.



(a) 획의 자소 클래스 추출

(b) 초성과 수직모음의 배열

(c) 수평모음의 배열

(그림 9) 자소 클래스 분류와 영역 추출
(Fig. 9) Classification of Consonant and Vowel

1) 수직 모음 클래스

수직 모음의 수직획(VstVv)은 일반적으로 문자의 상단 영역에서 시작하며, 그 위에 다른 클래스의 획이 오지 않으며 어느 정도 긴 수직획을 포함하고, 초성의 경우 'ㄴ', 'ㅂ', 'ㅅ'을 제외하고는 상단 영역에 수평획이거나 높이가 임계값 이하의 획을 포함한다. 초성의 'ㄴ'과 'ㅂ'의 경우 필기자에 따라 다르지만 보통 획의 끝에서 수평획을 포함하게 된다. 이러한 특성을 이용하여 획이 상단 영역에서 시작하고 그 위에 어떤 획도 존재하지 않으며 런길이가 $0.8H_l$보다 큰 수직획인 경우 수직 모음의 수직획으로 판단하고, $0.6H_l$보다 큰 경우 상 하 좌 우측에 연결된 획의 존재 여부에 따라 수직 모음의 수직획을 찾는다. 어떤 연결된 획도 없으면 이 획은 수직 모음의 수직획으로 판단하며, 획의 높이가 $0.6H_l$보다 크고 상단에 연결된 획이 없으며 하단에 연결된 획이 있는 경우 모호한 획(AmbST: Ambiguous STroke)으로 판단하고 모든 획에 대한 자소 클래스 인식이 끝난 후, AmbST 아래에 수평 모음이 오면 그 획은 초성이며 종성이 오면 그 획은 수직 모음의 수직획으로 판단한다. 수직 모음의 수직획의 임계 영역에 작은 수평획이 연결된 경우 이 수평획을 수직 모음의 수평획(HstVv)으로 판단한다.

2) 초성 클래스

초성은 보통 상단 영역에서 시작하며 그 위에 다른 클래스의 획이 오지 않는다. 그 이외에 가장 윗부분

에 수평획이 존재하는 경우가 많으며 수직획인 경우 임계값 이하의 높이를 가지고 수직획 끝 부분에 임계값 이상의 수평획이 연결된 것이 보통이다. 그 이외에 수직 모음 검출에서 임계값 이하의 수직획 상 하 좌 우에 긴 수평획을 가지는 경우 이 획은 초성으로 판단하며, AmbST 아래에 수평 모음이 존재하면 이 획은 초성으로 판단한다.

3)수평 모음 클래스

수평 모음이 수직획을 포함하는 경우(ㅗ, ㅛ, ㅜ, ㅠ) 두 개 이상의 획으로 검출된다. 따라서 상단 영역 아래에 위치하는 수평획에 대하여 연결된 획의 좌표를 고려하여 확장된 획의 폭을 임시 저장한다. 상단 영역 아래에 존재하며 그 위에 초성이나 AmbST 가 존재하고, 획의 폭이 0.6Hl보다 크거나 연결 화소내에 하나의 획이 존재하며 이 획의 폭이 0.3Hl보다 크면 수평 모음으로 판단한다.

4)종성 클래스

종성은 수평 모음이나 수직 모음과 접촉되는 경우가 많아 정확한 검출이 어렵기 때문에 자소 클래스 영역 추출에 포함되지는 않지만 AmbST가 수직 모음인지 또는 초성인지 구분하기 어려운 획의 자소 클래스 분류를 위하여 중요하다. 종성은 항상 그 위에 다른 자소 클래스가 오며, 그 아래에는 다른 자소 클래스가 오지 않는다. 따라서 자소 클래스가 결정되지 않은 획에 대하여 최하단 획(DownStroke)으로 획의 시작과 끝이 임계 영역 내에 있으면 종성으로 판단한다.

3.2.3 자소 클래스의 영역 추출

자소는 하나 또는 그 이상의 획들의 2차원적인 조합으로서, 칼럼에 대한 자소 클래스 존재 여부를 판단하기 위하여 각 획의 자소 클래스는 X축 방향의 1차원 배열로 합성된다. 이 때 종성은 예상 분할 영역을 위하여 이용하지 않기 때문에 제외시켰다. (그림 9.b)와 (그림 9.c)는 자소 클래스를 1차원 배열로 나타낸 것이다.

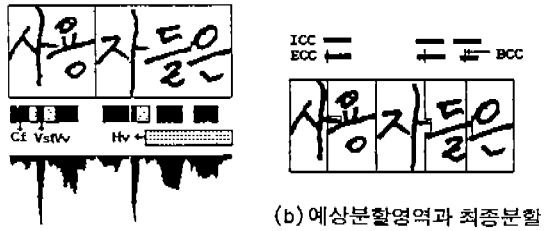
3.3 문자 분할

필기체에서 문자의 폭은 필기자에 따라 차이가 심하며, 같은 필기자라도 글자에 따라 폭이 달라지기 때문에 인쇄체 문자 분할에 많이 적용하는 평균 문자 폭이나 문자의 높이에 비례하여 추정한 문자폭의 적용이 어렵다. 이러한 문제점을 해결하기 위하여 본 연구에서는 자소 클래스에 기반한 분할 알고리즘을 개발하였으며, 평균 문자폭 대신 최소와 최대 문자폭을 적용하였다. 본 논문에서는 실험에 의하여 최소 문자 폭(W_{min})을 $0.5H_l$, 최대 문자폭(W_{max})을 $1.5H_l$로 정하였다.

분할 과정은 WRC 다음 초성이나 수평 모음이 오면 이 간격은 문자 사이의 간격으로 WRC의 시작 열에서 수직으로 분할을 한다. 이 과정에서 겹치거나 접촉되지 않은 문자들이 분할된다. 세그먼트 길이가 $2W_{min}$ 보다 크면 1 문자 이상이 포함될 수 있다고 가정하고 분할 예상 영역을 찾는다. 이 때 분할 예상 영역은 자소 클래스 영역과 지역적 특징을 이용한다. 분할 예상 영역이 존재하면 이 영역 내에 포함된 획들의 좌표를 이용하여 문자의 형태를 고려한 요철 형태의 분할을 수행한다.

3.3.1 WRC 다음에서의 분할 (분리된 문자의 분할)

WRC 다음에서의 분할은 $wSX + W_{min}$에서부터 오른쪽으로 WRC가 존재하는지 검사한다. WRC 다음에 초성이나 수평 모음이 오면 이 간격은 문자 사이의 간격으로 판단하고 이 열에서 수직으로 분할하고, 분할 칼럼 좌 우 $0.4H_l$ 영역을 분할 불가능 영역으로 지정한다. WRC 다음에 수직 모음의 수직획이 오면 이 칼럼은 문자내의 간격으로 분할되어서는 안되므로 분할 불가능 영역으로 지정하고 수직 모음의 수직획 끝 칼럼에서부터 분할 알고리즘을 계속한다. (그림 10a)는 WRC 다음 초성이나 수평 모음이 오는 경우 분할한 결과를 보이며, 겹쳐지거나 접촉되지 않은 문자들이 분리되는 것을 알 수 있다.



(a) WRC 다음에서의 분할 (b) 예상분할영역과 최종분할 결과

(그림 10) 분할 결과

(Fig. 10) Segmentation of Handprinted Hanguel

3.3.2 예상 분할 영역의 추정

$2W_{min}$보다 긴 세그먼트에 대하여 세그먼트의 시작 열(segSX)에서부터 끝 열(segEX)까지 사이에서 예상 분할 영역(ECC: Estimated Cutting Columns)을 추정한다. 예상 분할 영역은 우선 자소 클래스의 구성에 의하여 분할 가능한 모든 영역을 초기 예상 분할 영역(ICC: Initial Cutting Columns)으로 지정하고, 이 영역들에 대하여 W_{min}과 W_{max}, 지역적 특징을 이용하여 최종적으로 예상 분할 영역(ECC)을 결정한다. (그림 10b)에서 예상 분할 칼럼을 보여준다.

Step 1: 초기 예상 분할 영역의 추정

한글의 구성 법칙을 이용하여 초기 예상 분할 칼럼을 추정한다. 즉 수직 모음의 수직획 다음 임계 영역 내에 두 번째 수직 모음의 수직획이 없으면 수직 모음의 수직획 다음에서 초기 예상 분할 영역을 지정하고, 두 번째 수직 모음의 수직획이 존재하면 복모음으로 판단하고 두 번째 수직 모음의 수직획 다음에서 초기 예상 분할 영역을 지정한다. 초성이나 수평 모음 다음 다시 초성이나 수평 모음이 오면 이들 사이에서 초기 예상 분할 영역을 지정한다. 초기 예상 분할 영역이 추정되면 이 구간에서 분할 확률값이 가장 큰 영역을 최적 분할 열(BCC: Best Cutting Column)로 지정하고 예상 분할 영역 추정 단계로 간다.

Step 2: 예상 분할 영역의 추정

초기 예상 분할 영역이 실제 문자 사이의 구간인지를 결정한다. 검사 방법은 우선 지역적 특징을 이용하여 세그먼트의 시작 칼럼과 끝 칼럼으로부터 다음 조건이 만족될 때까지 칼럼을 제거한다.

$RC(x) \le 4$ and $CV(x) > 10$ and $V(x) < 0.5\ H_1$ (조건 1)

둘째 W_{min}과 W_{max}에 의한 검사다. BCC의 왼쪽과 오른쪽 세그먼트의 길이를 l_i과 l_{i+1}로 정의한다. l_{i+2}은 다음 오른쪽 세그먼트의 길이가 된다. 다음 조건을 만족하면 이 초기 예상 분할 영역을 제거하고 다음 예상 분할 영역으로 이동한다.

$(l_i < W_{min})$ or $(l_i > W_{min}$ and $l_i < W_{max}$ and $l_{i+1} < W_{min}$ and $l_i < l_{i+2})$ (조건 2)

셋째 지역적 특징에 의한 초기 예상 분할 영역의 삽입이다. l_i가 W_{min} 보다 크면 세그먼트의 시작 열에서부터 (조건 1)을 만족하는 영역을 초기 예상 분할 영역으로 검출하고 (조건 2)에 부합한지 검사한다.

(그림 10b)에서 입력 영상에 대하여 추정된 초기 예상 분할 영역과 예상 분할 영역, 최적 분할 영역을 보여준다.

3.3.3 문자 형태를 고려한 분할

특징 추출 단계에서 추출된 연결 화소들 중에서 예상 분할 영역 내에 있는 연결 화소들을 찾으며, 이 때 연결 화소가 어느 쪽 문자를 포함하는지 검사하여 CC_l, CC_b, 또는 CC_r에 각각 저장한다. 다음은 예상 분할 영역에 있는 연결 화소에 포함된 획들 중 예상 분할 영역 내에 있는 획들이 어느 쪽 문자에 연결된 획인지 결정하여 CH_l, CH_b, 또는 CH_r에 각각 저장하고 분할 과정을 수행한다.

분할은 최적 분할 칼럼의 시작점인 Pixel(Cx, wSY)에서 시작한다. Cx는 출발점의 x 좌표를 나타낸다. 분할점 P가 백화소이면 y값을 증가시키고, 흑화소이면 이 점부터 흑화소를 포함하는 획 Sk의 최소 외접 사각형을 따라 분할을 확장한다. 이 때 Sk가 CH_l에 포함되어 있으면 Sk의 최소 외접 사각형중 Cx의 오른쪽 부분을 따라 분할을 확장한다. 반대로 Sk가 CH_r에 포함되어 있으면 Sk의 왼쪽 부분을 따라 분할을 확장한다. Sk가 CH_b에 포함된 경우 분할은 1차 분할점에서 수행되며 후보 분할점을 저장한다. 이 때 Sk의 자소 클래스가 초성, 수평 모음 또는 종성이면 Sk의 시작 칼럼이 1차 분할점이 되며, Cx가 후보 분할점이 된다. 그렇지 않은 경우 Cx가 1차 분할점이 되며 Sk의 최소 x값이나 최대 x값이 예상 분할 영역 내에 있는 경우 이들 값에서 후보 분할점이 결정된다. (그림 10)은 (그림 6)의 입력영상에 대한 최종 분할 결과이다.

다음은 겹쳐지거나 접촉된 문자들에 대한 분할 알고리즘이다.

```
CurvedSegment(integer Cx, wSY)/* Algorithm for
Overlapped or Touching Characters */
  for Index = wSY to wEY do begin
    if Pixle(Cx, index) is black pixel then begin
      /* splitting is extended along the outline of
      s rectangle */
```

```
        if Pixel(Cx, index) belongs to CH_l then Extend cutting to Sk's rightmost column
        else if Pixel(Cx, index) belongs to CH_r then Extend cutting to Sk's leftmost column
        else if (Pixel(Cx, index) belongs to CH_b then
        begin
          Extend cutting to Primary Cutting Point
          Store the Alternative Cutting Point
        end
        Extend cutting to Pixel(Cx, Y_max[Sk] + 1)
        Change index to Y_max[Sk] + 1
    end
end
```

4. 실험 및 결과

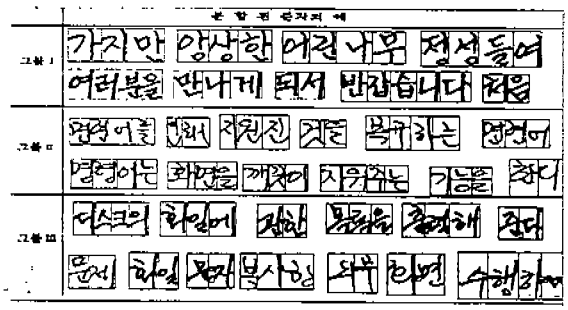
실험은 PENTIUM PC와 MS WINDOWS 3.1 환경하에서 Borland C++ 언어로 구현하였다. 실험 영상은 300 DPI 해상도로 획득하였다. 실험 데이터는 다음의 3 그룹으로 나누어 획득하였다.

A 그룹:61 단어, 198자 (10명의 학생이 필기한 기울어지지 않은 정서체 문장)

B 그룹:478 단어, 1,466자 (63명의 report로부터 획득한 영상에서 한 문장씩 발췌)

C 그룹:370 블록, 910자 (B 그룹의 영상으로부터 접촉된 문자 블럭 선택)

(그림 11)은 각 그룹의 실험 데이터에 대한 분할 결과의 일부분을 발췌한 것으로 필기자에 따라 문자의 크기, 높이, 폭이 다를 뿐만 아니라, 문자의 높이에 대한 폭의 비가 크게 다르다. 그 뿐만 아니라 동일한 필기자에서도 문자폭이 크게 다름을 알 수 있다. (그림 11)에서 보는 바와 같이 제안한 분할 알고리즘에서는 획이 어느 정도 기울어지거나, 휘어진 경우에도 정확한 분할을 수행할 수 있었으며, 겹치거나 접촉된 문자들이 획의 외접 사각형을 따라 분할됨으로써 문자의 형태에 따라 정교하게 분할되는 것을 볼 수 있다. 몇몇 문자들이 1차 분할점에서는 정확한 분할이 되지 않지만 후보 분할점에서 정확하게 분할되는 것을 알 수 있다. 또한 제안한 알고리즘은 잡음에 강한 것을 알 수 있다.



(그림 11) 실험 결과
(Fig. 11) Experimental Results

(표 1)은 각 그룹의 실험 데이터에 대한 분할 결과이다. 분할율은 후보 분할점에서 정확하게 분할된 경우까지 포함된 것이다.

〈표 1〉 문자 분할 정확도
〈Table 1〉 Accuracy Rate of Segmented Characters

그 룹	총 문자 수	정확하게 분할된 문자 수	분할 정확도
A	198	196	98.9%
B	1,466	1,392	94.8%
C	910	796	87.5%

문자의 기울어짐을 허용하지 않은 정서체(A 그룹)인 경우 98.9%의 높은 분할율을 알 수 있다. 학생들의 report로부터 발췌한 B그룹의 실험 데이타에서도 94.8%의 비교적 높은 분할율을 보이고 있다. 문자 분할에서 가장 어려운 것은 접촉된 문자들의 분할이다. 접촉된 문자들만을 별도로 실험한 결과 87.5%의 분할율을 보이고 있다. 이것은 제안한 알고리즘이 접촉된 문자들의 분할에도 성능이 굉장히 좋다는 것을 보여 준다. 위 결과와 비교할 off-line 한글 필기체 분할에 관하여 발표된 연구가 아직 없다. (표 2)는 영문자나 숫자의 분할에 관한 연구들의 실험 결과를 요약한 것이다. 실험 결과에서 보는 바와 같이 영문자나 숫자에서도 필기체 문자의 분할은 그 특성상 아직 그렇게 높은 분할율을 보이지 못하고 있다. 실험 데이터가 각 각 다르기 때문에 직접 비교할 수는 없지만 70%-90%의 분할 정확도를 보이고 있다.

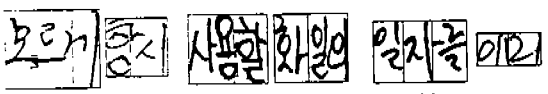
분할 오류 중 미분할된 문자는 25자로 약 1.7%에 해당한다. (그림 12)은 분할 오류의 예이다. 분할 오류

〈표 2〉 기존 연구의 실험 결과 비교
〈Table 2〉 Experimental Results of each handwritten character segmentation methods

연구자	분할 방법 및 특징	실험 대상	정확도
Han/Sethi	전역적 특징 + 구조적 특징	50개의 우편 봉투에서 발췌 : 1119자	85.7%
Westall	구조적 특징	1000개의 문서 : 390블럭의 연결된 숫자	73%
Seshdri	구조적 특징	296개의 연결된 숫자	80%
*Bozinovic	외곽선 추적 + 투영 방법	64 단어 : 기울어짐 허용 안함	77%
*Simon	구조적 특징	25 단어 : 25명의 필기자	91%

*온 인식기와 결합에 의한 성능 평가

의 원인은 크게 4가지로 나눌 수 있다. (a)는 획의 자소 클래스를 오인식하여 잘못 분할한 경우로 전체 오류 74자 중 35자로 약 반이 여기에 속한다. 오인식의 원인은 주로 그림에서 보는 바와 같이 수직 모음 'ㅏ'나 'ㅓ'의 수직획을 짧게 쓰고 수평획이 위나 아래쪽에 연결되며, 그 아래에 종성이 존재하지 않는 경우이다. 초성 'ㅅ'의 왼쪽 획이 경사가 없을 경우 수직 모음의 수직획으로 인식될 수 있으며, 'ㅂ'이나 'ㅁ'의 길이가 문자의 높이와 거의 같은 수준으로 쓰는 경우(c의 '비') 수직 모음으로 인식되어 분할 오류의 원인이 된다. (b)는 문자의 폭이 문자 기준높이의 0.5보다 작고 두 문자폭의 합이 문자 높이의 1.5 배보다 작기 때문에 '사'나 '일'의 수직 모음이 검출되었는데도 수직 모음 다음에서 분할이 되지 않은 경우이다. 미분할된 대부분의 원인이 여기에 속한다. (c)는 획이 문자에서 분리되어 있어 오분할된 경우로 '일자를'에서 '자'의 수평획이 분리되어 초성으로 인식되어 오분할되었으며, '이미'에서 'ㅁ'의 왼쪽 획이 분리되어 수직 모음으로 인식되어 오분할된 예이다.



(a) (b) (c)

(그림 12) 분할 오류 예
(Fig. 12) Examples of Segmentation Error

5. 결 론

필기체 한글의 분할을 어렵게 하는 요인으로 문자의 겹침이나 접촉, 문자폭의 심한 변화, 다양한 필기체, 문자의 기울어짐 등을 들 수 있다. 이러한 문제점을 해결하기 위하여 자소 클래스의 인식에 기반한 문자 분할 알고리즘을 제안하였다. 특히 한 번의 수직 투영으로 문자 분할을 위하여 중요한 특징으로 이용되는 열에 대한 정보, 즉 지역적 정보와 연결 화소, 획을 동시에 검출하였다. 투영에서 검출한 획으로부터 한글의 구조적 특성을 이용하여 자소 클래스 영역을 검출하였으며, 한글의 자소 구성 법칙과 경험적 지식을 기반으로 분할 알고리즘을 개발하였다. 제안한 알고리즘에서는 세선화 과정을 거치지 않고 투영에 의하여 모든 필요한 정보를 획득하기 때문에 세선화 과정에서 발생하는 분할을 위한 정보의 손실이나 왜곡을 막을 수 있었다.

겹치거나 접촉되지 않은 문자들을 WRC에서 수직으로 분할함으로써 빠른 속도로 분할이 가능하며, 세그먼트가 $2W_{min}$보다 큰 경우 예상 분할 칼럼을 추정하고 이 예상 분할 칼럼 내에 있는 획들의 최소 외접 사각형을 따라 분할함으로써 문자의 형태에 따른 분할을 수행할 수 있었다. 이 때 접촉된 문자의 경우 획이 양쪽 문자에 포함되므로 후보 분할점을 허용함으로써 필기체 문자 분할을 위한 직접 분할 방법의 문제점을 해결하고자 하였다.

실험에서 제안한 알고리즘이 겹침이나 접촉된 문자들을 포함하는 다양한 필기체의 문자 분할에 적합하다는 것을 입증하였다. 특히 문자 크기와 폭의 변화, 잡음에 강한 것을 보인다. 획이 어느 정도 기울어지거나 휘어진 경우에도 정확한 분할이 가능하였으며, 겹치거나 접촉된 문자들이 문자의 형태에 따라 정교하게 분할될 수 있었다.

그러나 실험 결과에서와 같이 자소 클래스의 오인식에 의하여 오분할되는 것을 막기 위하여 자소 클래

스 인식 방법을 개선하고, 최소 문자폭에 의하여 분할되지 못한 문자들의 검증 과정을 위한 연구가 수반되어져야 할 것으로 보이며, 획이 문자에서 분리되거나 획의 끊김을 해결할 수 있는 방법이 요구되어진다. 또한 문서 인식의 실용화를 위하여는 영 숫자가 혼용된 문자열에 대한 통합된 분할 알고리즘의 개발을 위한 연구가 이루어져야 할 것으로 보인다.

참 고 문 헌

[1] 이성환, "문자 인식", 홍능 과학 출판사, 1994

[2] S. Mori, C. Suen, "Historical Review of OCR Research and Development", proc. IEEE, Vol.80, No.7, Jul. 1992

[3] Y. Lu, "On the Segmentation of Touching Characters", The Second International Conference on Document Analysis and Recognition, 440-443, Oct. 1993

[4] 최봉희, 이인동, 김태균, "문자 영역 추출 과정에서의 오분리 교정", 한국정보과학회 논문지 제21권 1호, pp.86-93, 1994년 1월

[5] T. Bayer, U.Kreβel, "Cut Classification for Segmentation", Proc. of 2nd Int. Conf. on Document Analysis and Recognition, pp.565-568, Tsukuba Science city, japan, Oct. 1993

[6] 권재욱, 조성배, 김진형, "계층적 신경망을 이용한 다중 크기의 다중활자체 한글 문서 인식", 한국정보과학회 논문지 19권1호, pp.69-78, 1992년 1월

[7] 황순자, 김용경, 이경수, 김문현, "한글 인쇄체 문자의 2단계 분할 방식", 제1회 지능기술 공동 학술회의, 1995년 8월

[8] Y. Lu, M. Shridhar, "Character Segmentation in Handwritten Words-An overview", Pattern Recognition, pp.77-96, Vol.29, No.1, 1996

[9] M. M. Beglou, M.J.J. Holt, S. Datta, "Slant Independent Letter Segmentation for Cursive Script Recognition", IWFHR, pp.375-380, Sep. 1991

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

[11] K. Han, I.K. Sethi, "Off-Line Cursive Handwriting Segmentation", Proceedings of the 3th ICDAR, pp.894-897, Aug. 1995

[12] E. Lecolinent, J.P. Crettez, "A Grapheme-Based Segmentation Technique for Cursive Script Recognition, IWFHR, pp.740-748, Sep. 1991

[13] S. Seshadri, D. Sivakumar, "ATechnique for Segmenting Handwritten Digits", IWFHR, pp. 443-448, Sep, 1991

[14] M. M. Beglou, M.J.J. Holt, S. Datta, "Slant Independent Letter Segmentation for Cursive Script Recognition", IWFHR, pp.375-380, Sep. 1991

[15] R. Fenrich, "Segmentation of Automatically Located Handwritten Words", IWFHR, pp. 33-44, Sep. 1991

[16] 김의정, 김태균, "인쇄체 문서 인식을 위한 문자 추출에 관한 연구", 제2회 문자 인식 워크샵, 171-179, 1994년 9월

[17] 김두식, 이성환, "한글과 영·숫자가 혼용된 문서를 위한 효과적인 문자 분할 방법", 제8회 영상처리 및 이해와 관한 워크샵, pp.19-26, 1996년 2월

[18] 장명욱, 천대녕, 양현승, "연결 화소를 이용한 문서 영상의 분할 및 인식", 한국정보과학회 논문지 제20권 12호, pp.1741-1751, 1993년 12월

황 순 자

1976년 항공대학 항공경영과 졸업(이학사)

1982년 연세대학교 산업대학원 전자계산학과 졸업(공학석사)

1988년 독일 Konstanz Univ. 정보학과 졸업(이학석사)

1991년~현재 성균관대학교 대학원 정보공학과 박사과정

주관심분야: 패턴 인식, 인공 지능

김 문 현

1978년 서울대학교 전자공학과 졸업(공학사)
1980년 한국과학기술원 전기 및 전자공학 석사(공학석사)
1988년 Univ. of Southern California 컴퓨터공학과(공학박사)
1988년~현재 성균관대학교 정보공학과 부교수
주관심분야: 패턴인식, 인공지능