한글의 구조적 인식을 위한 자획 추출에 관한 연구

이 승호, 김 진형

한국과학기술원, 전산학과

A Study on Stroke Extraction for Structural Recognition of Korean Characters

Seungho Lee and Jin H Kim

KAIST, Department of Computer Science

요약

패턴 인식 시스템은 일반적으로 전처리 단계, 특성 추출 단계, 분류 단계 및 학습 단계 등의 네 가지 단계로 구성된다 문자 인식 시스템의 경우에 후처리 과정으로서 확인 단계가 첨가되어질 수 있다 그동안 한글 문자 인식에 대한 연구가 많이 되어 왔는데, 기존의 한글 인식 시스템에서는 전처리 단계와 특성 추출 단계에 대한 연구가 부족하였고 잡음 처리에 대한 연구도 거의 없었다. 또 기존의 연구는 주로 온라인 한글 인식과 인쇄체 한글 인식에 대한 방법에만 치중하였으므로 필기체 글자에 대해서는 적용하기가 어렵다는 문제점도 있다.

본 논문에서는 전처리 단계와 특성 추출 단계에 중점을 두고서, 한글 문자의 구조적 특성에 의한 자획 모델을 세운 후 이 모델을 이용하여 정확한 자획을 추출하는 새로운 방법을 제안하였다 본 논문에서 제안한 자획 추출 방법을 이용함으로써 잡음의 효과적인 처리가 가능하고, 한문에 대해서도 적용할 수 있어 구조적 한문 인식에도 사용이 가능하며, 필기체 한글 인식에도 적용할 수 있는 장점을 갖는다

I. 서론

지난 20 여 년 동안 패턴 인식과 컴퓨터 비전에 관한 관심이 고조되어 왔다 패턴 인식의 응용 분야는 매우 광범위한데, 그중에서도 특히 문자 인식에 대한 연구가 많이 있어 왔다 현대 사회에서 정보의 발생 및 전달의 양이 해마다 엄청난 속도로 증가되어 가고 있는데, 지금까지는 인력으로 이것들을 컴퓨터에 입력시키고 있으나, 보다 빠른 정보 처리를 위해서는 반드시 자동 입력 시스템을 필요로 하고 있다 이러한 자동 입력을 위해서는 먼저 컴퓨터에 의한 문자 인식이 가능해야만 한다

패턴 인식 시스템은 일반적으로 다음의 네 가지 단계로 구성된다 [Fu84] 첫째 단계는 전처리 단계로서 다음 단계에서의 특성 추출을 용이하게 하기 위하여 패턴 영상의 잡음 및 중복성을 제거하는 평골화 및 세선화 작업 등을 수행한다 둘째 단계는 특성 추출 단계로서 입력 패턴에 대하여 각각의 패턴들을 판별하는데 사용할 수 있는 특성 값들을 추출해낸다 세째 단계는 분류 단계로서 전 단계에서 구해진 특성 값들을 이용해서 입력 패턴이 어느 부류에 속하는가를 알아내거나, 그 패턴이 무엇인가를 규명해낸다 넷째 단계는 학습 단계로서 실험 데이타를 가지고서 테스트를 하면서 시스템을 개선시켜 나간다 문자 인식 시스템의 경우에 후처리 과정으로서 확인 단계가 첨가되어질 수 있다.

지난 10 여년 동안 한글 문자 인식에 대한 연구가 많이 있어 왔는데, 기존의 한글 인식 시스템들의 문제점들을 살펴보면 다음과 같다 첫째로 전처리 단계와 특성 추출 단계에 대한 연구가 부족한 점을 들 수가 있다. 구문론적인 인식 방법을 사용한 경우에 전처리 단계와 특성 추출 단계를 거쳐서 나온 결과인 패턴 그래프를 입력으로 가정한 상태에서 알고리즘을 전개시켜 나갔다 전처리 과정과 특성 추출에 대한 알고리즘이 좋을수록 다음 단계인 분류 과정이 용이해지기 때문에 이러한 저단계 처리 과정에 대한 연구도 매우 필요하다 둘째로 잡음 처리에 대한 연구가 거의 없었다는 점을 들 수가 있다. 잡음이 거의 없다고 가정하거나 가상되어진 잡음을 가지고서 문자 영상을 만들었는데 이러한 영상 데이타는 실제로 스캐너를 통해 얻어진 것과는 많은 차이가 있기 때문에 잡음 처리가 효과적일 수가 없었다 세째로는 기존의 연구가 주로 온라인 인식과 인쇄체 한글에 대한 방법에만 치중하였으므로 필기체 글자에 대해서는 적용하기가 어렵다는 점이다.

본 논문에서는 한글 문자의 구조 분석을 통하여 정확한 자획 모델을 세우고, 그 모델에 의한 자획 추출 알고리즘을 제안하였다 그리고 추출된 자획을 가지고서 구조적인 인식 방법을 사용하여 한글 문자를 인식하였다. 자획 추출 알고리즘은 한글 뿐만이 아니라 한문에 대해서도 적용할 수가 있으므로 구조적 한문 인식에도 사용되어질 수가 있고, 이러한 인식 방

법은 필기체 한글 인식에도 적용되어질 수가 있다

II. 배경

패턴 인식 방법은 패턴의 표현 방법과 판별 방법에 의하여 원형 비교 방법, 통계적 방법 및 구조적 방법 등으로 나누어진다 [Fu84]

2.1 원형 비교에 의한 패턴 인식 방법

이 방법에서는 패턴이 원래의 입력 형태인 2 차원 배열로 표현되어진다 입력 패턴을 각각의 원형 패턴들과 모두 비교하고 그중에서 차이를 최소로 하는, 즉 입력 패턴과 가장 가까운 원형을 찾아내서, 입력 패턴을 그 원형 패턴이 속하는 부류로 분류한다 이 방법은 초기의 패턴 인식 방법에서 주로 사용되었으며 인쇄체 문자와 같은 고정된 패턴의 인식에 사용되어질 수 있다.

2.2 통계적 패턴 인식 방법

패턴은 N 차원의 특징 벡터로 표현되는데, 그 특징 벡터들의 통계적인 분포를 알아냄으로써 그 벡터 공간을 각각의 부류로 분리시킨다 주어진 입력 패턴으로부터 특징 벡터를 추출한 다음에 그 특징 벡터가 통계적인 분포상에서 어느 곳에 속하는 가를 알아냄으로써 그 입력 패턴을 분류한다 이 방법은 수학적으로 모델링이 가능하다는 장점이 있으나, 문자가 계층적인 구조를 가지고 있을 때 효율적으로 인식할 수 없다는 단점이 있다

2.3 구조적 패턴 인식 방법

패턴은 그 패턴을 구성하고 있는 기본 원소들의 스트링 형태나, 원소들과 그것들 사이의 관계를 나타내는 트리나 그래프 형태로 표현된다. 입력 패턴이 주어졌을 때 먼저 그 패턴을 구성하고 있는 기본 원소들을 인식한 다음, 그것으로부터 패턴 그래프와 같은 자료 구조를 생성한다 예를 들어 패턴 그래프는 특징점이 노드가 되고, 자획이 호가 되며, 자획의 형태가 호의 라벨이 되는 식으로 만들수 있다 이 패턴 그래프를 문법의 도출 과정을 이용하여 처리함으로써 패턴을 인식한다 이 방법은 한글과 같이 계층적인 구조를 갖는 문자의 인식에 적합하고 글자 모양의 변형이 심한 필기체 인식에도 적합하므로, 1970 년대 중반 이후로 많은 연구가 되고 있는 분야이다.

III. 전처리 단계

전처리 단계는 다음 단계인 특성 추출 단계에서의 특성 추출을 용이하게 하기위하여 입력 영상의 잡음 및 중복성을 제거하는 단계로서, 본 논문에서는 문자 영상의 위치 조정, 평골화, 형상 추적 및 세선화 과정 등으로 구성되어 있다

3.1 문자 영상의 위치 조정

입력된 문자 영상에서 글자 부분이 중앙에 위치하지 않고 어느 한쪽으로 치우칠 수가 있다 그러한 입력 영상은 글자 부분이 가운데 위치하도록 글자 부분의 영상을 중앙으로 이동시켜줄 필요가 있다 통계적인 방법에 의한 문자 인식 시스템에서는 이 위치 조정 루틴이 반드시 필요하다 구조적인 문자 인식 시스템에는 이것이 별로 영향을 미치지 않으나 문자 영상의 맨끝 가장자리에 값이 '1'인 화소가 존재할 경우에 3 x 3 마스크 윈도우를 가지고서 문자 영상 배열을 처리하는 과정에서 연산을 수행할 때마다 경계 조건을 항상 테스트해야만 하는 번거로움이 발생한다. 이러한 번거러움을 제거하기 위해서는 이 문자 영상의 위치 조정 루틴이 반드시 필요하다.

3.2 평골화

입력된 영상에는 반드시 잡음이 존재하게 된다. 잡음이란 영상 화소의 값이 원래 '1'이어야 할 것이 '0'이 되거나 혹은 그 반대로 되는 경우를 말한다 스캐너를 통하여 얻어진 영상에는 이러한 잡음이 존재하는데, 이것으로 인하여 영상의 글자 부분 한가운데 구멍이 존재하거나 글자 부분의 가장자리가 거칠어진 모양을 하게 된다. 이러한 잡음은 세선화 처리에 나쁜 영향을 주므로 평골화 처리를 함으로써 제거하여야 한다 평골화에 대해서는 많은 방법들이 있으나 이 논문에서는 Stentiford의 알고리즘을 사용하였다 [Ste83]

3.3 형상 추적

세선화 과정에서 글자 부분의 형상 추적을 필요로 한다. 여기서 형상이란 값이 '1'인 화소들이 연결된 덩어리를 '블록'으로 정의할 때, 각 블록의 가장자리에 속하는 화소들의 집합을 그 블록의 형상이라고 한다 형상에는 외부 형상과 내부 형상이 있는데, 대부분의 블록은 외부 형상만 존재하나 루프 형태로 구성된 블록은 외부 형상과 내부 형상을 모두 갖는다. 한글의 경우 루프 형태를 갖는 자모로는 ㅁ, ㅇ, ㅂ, ㅍ, ㅎ, ㅐ, ㅒ, ㅖ 등이 있다. 이 과정을 거치고 나면 각 블록의 가장자리에 있는 화소들의 값은 형상 추적을 하면서 지나간 횟수만큼 그 값이 증가된다

본 논문에서는 Pavlidis의 알고리즘을 문자 영상에 알맞게 고쳐서 사용하였는데 간략하게 기술하면 다음과 같다 [Pav82] 먼저 형상의 시작점을 찾아낸 다음 이 점이 외부 형상의 시작점인지 내부 형상의 시작점인지를 판별한 다음에 각각에 대하여 외부 형상 추적 및 내부 형상 추적 알고리즘을 수행한다 형상의 시작점은 Raster Scanning 순서로 영상을 스캐닝해 나가면서 화소 값이 '1'인 화소들 중에서, 바로 왼쪽 이웃 화소의 값이 '0'인 경우에는 외부 형상의 시작점으로 간주하고, 바로 오른쪽 이웃 화소의 값이 '0'인 경우에는 내부 형상의 시작점으로 간주하였다

3.4 세선화 과정

스캐너로부터 얻어진 입력 영상은 불필요한 화소들을 많이 포함하고 있기 때문에 세선화를 시킴으로써 필요한 최소한의 정보만을 갖도록 할 수 있다. 지난 20 여년 동안 세선화에 대한 많은 연구가 있어 왔는데 요즘은 3 차원 영상에 대한 세선화 연구도 많이 되고 있는 중이다 이 논문에서는 Hideo Ogawa의 문자 영상을 위한 세선화 알고리즘을 사용하였는데, 이 알고리즘은 각 화소 라인들의 접점 부분에서는 완벽하게 세선화가 안되므로 후처리 단계를 필요로 한다 [Hid82] 이 알고리즘의 후처리 단계로는 SPTA(Safe Point Thinning Algorithm) 알고리즘을 사용하였다 [Nac84] 따라서 본 논문에서 사용한 세선화 알고리즘은 다음의 세 단계로 구성되었다

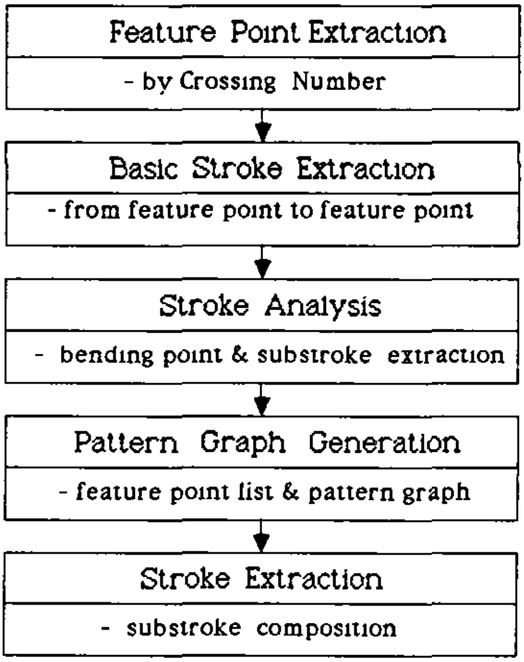
1. 입력 문자 영상의 두께를 2 내지 3 정도로 줄인다

2. 2 x 2 정방 윈도우를 가지고서 문자 영상의 중심선을 찾아 추적한다.
3. 연결점 부분의 비세선화 부분을 SPTA 알고리즘을 적용하여 완전히 세선화시킨다.

이 알고리즘은 자획으로만 구성된 문자 영상을 전제로 하여 만들어진 알고리즘이므로 다른 세선화 알고리즘들과 비교해볼 때, 세선화된 자획의 길이가 가장 적게 줄어들고 또한 자획의 직선 성질을 잘 유지시켜주는 장점이 있는 반면에, 다른 알고리즘들에 비하여 복잡도가 크고 수행 속도가 느린 것이 단점이다.

IV. 자획 추출 알고리즘

문자 인식을 위하여 자획을 분리해내는 방법에는 자획 추출 방법과 다각형 근사에 의한 방법과 양방향 자획 전파 방법 등 3 가지가 있는데 본 논문에서는 세선화를 이용한 자획 추출 방법을 사용해서 자획을 분리하였다 자획 추출 알고리즘에 대한 흐름도가 그림 4-1에 나와 있다.



< 그림 4-1 · 자획 추출 알고리즘의 흐름도 >

4 1 한글 문자의 자획 모델

한글 문자 영상에서 자획을 추출해내기 위해서는 먼저 한글이 어떻게 구성되어 있는가를 정확히 분석한 다음에 그것에 알맞은 자획에 관한 모델을 세우는 과정이 필요하다. 한글은 기본적으로는 24 개의 자소로 구성되어 있지만, 그것들의 조합으로써 30 개의 자음과 21 개의 모음이 존재한다

한글의 각 자소들은 하나 이상의 자획들로 구성되어 있다 여기에서 자획이란 사람들이 일반적으로 한글 글자를 쓸 때 펜이 지면에 닿는 점으로부터 떨어지는 점까지의 펜이 지나간 경로를 말한다 한글 자소들을 구성하는 자획으로는 다섯 가지의 기본 자획과 두 가지 형태의 복합 자획이 있다. 복합 자획은 두 개의 기본 자획이 모여 구성되는데 연결되는 각도에 따라서 약간의 변형이 존재한다 각 자획들은 다음과 같다

기본 자획
수평 자획 (Horizontal type) : |
수직 자획 (Vertical type) : _
좌사 자획 (Left-down type) : /
우사 자획 (Right-down type) : \
원형 자획 (Circle type) . o

복합 자획
상곡 자획 (Upper-bending type) : ㄱ
하곡 자획 (loWer-bending type) : ㄴ

4 2 Crossing Number에 의한 특징점 추출

세선화된 문자 영상에서의 각 화소들은 고립점(Isolated Point), 단점(End Point), 연결점(Connection Point), 굴곡점(Bending Point), 분기점(Branch Point) 및 교차점(Cross Point) 등의 다섯 가지 종류의 점들중 하나로 분류되어진다. 화소들의 대부분은 연결점에 속하게 되는데, 고립점은 잡음으로 간주하여 제거될 수 있으며, 이 두가지 점들을 제외한 나머지 점들은 문자 영상의 형태에 대한 정보를 알려주는 특징점이 된다

한 화소에 대한 Crossing Number는 그 화소 주위에서의 분기도의 정도를 나타내는 값으로, 이에 대한 정의는 다음과 같다 [Hid82]

Crossing Number : Xc
8 neighborhood configuration of pixel (i,j)

3	2	1
4	(i,j)	0
5	6	7

Xc = 1/2 * (Summation (k=0,7) of |N(k+1) - N(k)|)
N(k) = k-th neighbor of pixel (i,j)
N(8) = N(0)

< 그림 4-2 · Crossing Number의 정의 >

세선화된 영상에서 각 화소의 Crossing Number는 0에서 4까지의 정수들 중에서 어느 한 값만을 갖게 되는데, 그 값에 의하여 굴곡점을 제외한 나머지 특징점들을 추출해낼 수 있다. 굴곡점과 연결점은 Crossing Number의 값이 2로서 서로 같기 때문에 이 단계에서는 굴곡점을 추출해낼 수가 없다 각 점들에 대한 Crossing Number의 값은 고립점은 0, 단점은 1, 연결점 또는 굴곡점은 2, 분기점은 3 그리고 교차점은 4가 된다. 이해를 돕기 위하여 여기에 대한 예들을 그림 4-3에 나타내었다

0 0 0	0 0 0	0 0 0	0 0 0	0 1 0	0 1 0
0 P 0	0 P 1	1 P 1	1 P 0	1 P 0	1 P 1
0 0 0	0 0 0	0 0 0	0 1 0	0 1 0	0 1 0
Xc=0	Xc=1	Xc=2	Xc=2	Xc=3	Xc=4
고립점	단점	연결점	굴곡점	분기점	교차점

< 그림 4-3 . 특징점들에 대한 Crossing Number 값 >

4 3 기본자획의 추출 및 분석

이 단계에서는 두 개의 특징점 사이를 서로 연결하고 있는 모든 자획들을 추출하게 되는데, 이것을 기본 자획으로 정의한다 문자 영상을 스캐닝하면서 먼저 특징점을 발견한 다음,

그 화소로부터 연결된 이웃 화소들을 따라서 진행해 나가다가 다른 특징점이 발견되어지면 두 특징점 사이의 자획을 하나의 기본 자획으로 한다. 기본 자획은 크게 2 가지 형태로 분류되는데 하나는 루프 형태이고 다른 하나는 직선 형태이다.

기본 자획들이 모두 추출되어지면 그 다음에는 기본 자획들을 하나씩 분석하여 굴곡점의 존재 여부를 알아낸다. 굴곡점이 존재하면 기본 자획은 여러 개의 부분 자획으로 분리되고, 그렇지 않을 경우에는 하나의 부분 자획으로 간주한다. 추출된 부분 자획들은 시작점의 위치, 끝점의 위치, 부분 자획의 길이 및 자획의 형태에 대한 정보를 갖게 된다. 부분 자획의 형태로는 앞의 4.1 절에서 기술되어진 바와 같이 H, V, L, R 및 C 형태가 있다. 이 과정에서 길이가 짧으면서 자획의 형태와 양끝점에서의 Crossing Number 값에 의하여 잡음으로 판별되어진 부분 자획들은 제거된다.

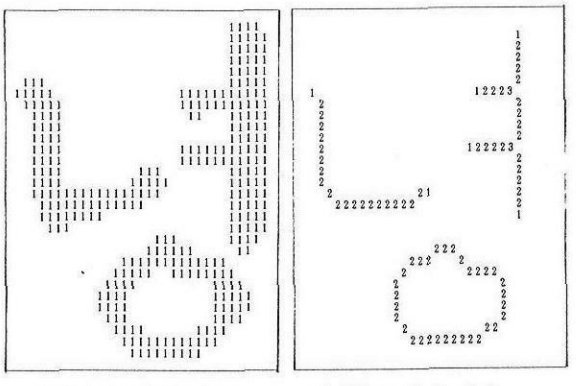
4.4 패턴 그래프의 생성

앞 단계에서 추출된 부분 자획들과 굴곡점을 포함한 특징점들을 가지고서 패턴 그래프라는 자료 구조를 생성하게 된다. 이 패턴 그래프는 행렬 형태로 구현되어지는데, i 번째 특징점과 j 번째 특징점 사이에 부분 자획이 존재하면 패턴 그래프의 (i,j) 항에 그 부분 자획의 형태와 길이를 저장하게 된다. 이 그래프는 다음 단계인 조합 자획의 추출을 위한 입력으로 사용된다.

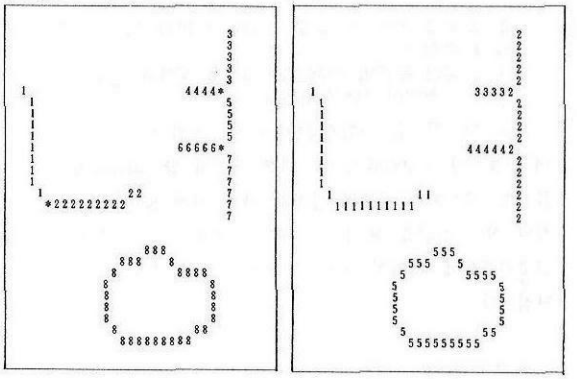
4.5 게슈탈트 법칙에 의한 조합 자획의 추출

부분 자획들은 매끄러운 연속의 법칙에 의하여 서로 연결되어질 수가 있다. 그 연결되어진 부분 자획들은 하나의 자획을 구성하게 된다. 예를 들어 '논' 자에는 8 개의 부분 자획이 존재하게 되나 이것을 매끄러운 연속의 법칙에 의하여 부분 자획들을 서로 연결하게 되면 4 개의 자획으로 줄어들게 된다. 사람들이 글자를 쓰는 순서는 이 매끄러운 연속의 법칙에 따르므로, 이 법칙을 If-Then-Else 규칙에 의하여 구현함으로써 문자 영상 중에서 서로 연결 가능한 부분 자획들을 서로 연결하여 하나의 자획으로 만들어 낼 수 있다.

이 과정에서는 부분 자획들을 서로 연결시켜주는 것이 대부분이지만, 입력 영상에 잡음이 생겨서 부분 자획들이 서로 연결되었을 경우에는 필요에 따라서 그 부분 자획을 분리시키는 작업도 필요하다. 자획 추출 과정의 각 결과들을 그림 4-4에 나타내었다.



< Original Font > < Thinned Font >



< Substrokes > < Strokes >

< 그림 4-4 : 자획 추출 과정의 결과 >

이 단계에서 추출되어진 자획들을 가지고서 다시 새로운 패턴 그래프를 생성하게 되는데, 이 그래프는 다음의 인식 단계에서 자소의 인식을 위한 정보로 사용되어진다.

V. 결론

본 논문에서는 한글의 구조적 특성을 이용하여 입력된 문자 영상으로부터 사람이 쓰는 것과 같은 형태의 정확한 자획을 추출하는 방법을 제안하였다. 그리고 그 결과를 이용하여 구조적 한글 인식 시스템을 구현하였다.

한글은 각 자소들의 계층 구조적인 조합에 의하여 각 글자가 구성되므로 한 글자를 구성하고 있는 자소들의 상대적인 위치는 일정하다. 이러한 한글의 구조적인 특성과 자획 추출 단계에서 얻어진 정보들을 이용하여 각 자소들에 대한 인식 규칙을 정하였다. 입력 문자 영상으로부터 추출되어진 자획들에 대하여 먼저 존재할 가능성이 있는 자소들을 찾아낸 다음, 그 자소들에 대한 인식 규칙을 차례로 적용해 나감으로써 문자 내에 포함되어 있는 각 자소들을 추출해낼 수 있었다.

본 논문에서 제안된 자획 추출 방법은 자획들을 추출해내는 과정에서 잡음으로 판명된 자획들은 제거하고, 또 잘못 연결된 자획들은 서로 분리시킬 수 있으므로 입력 영상에서의 잡음을 제거할 수가 있다. 그리고 한문의 구조적 인식에 있어서 특성 추출 단계로 사용되어질 수가 있고, 필기체 인식에도 적용 가능하다는 장점을 갖고 있지만 처리 시간이 다른 자획 분리 방법들보다 오래 걸린다는 단점이 있다.

참고 문헌

[1] S.F. Fu and Azriel Rosenfeld, "Pattern Recognition and Computer Vision," IEEE Computer, pp. 274-282, October 1984.

[2] F.W.M. Stentiford and R.G. Mortimer, "Some New Heuristics for Thinning Binary Handprinted Characters for OCR," IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-13, No. 1, January/February 1983.

[3] T. Pavlidis, *Algorithms for Graphics and Image Processing*, Washington, D.C., Computer Science Press, 1982.

[4] N.J. Naccache and R. Shinghal, "SPTA: A Proposed Algorithm for Thinning Binary Patterns," IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-14, No. 3, pp. 409-418, May/June 1984.

[5] H. Ogawa and K. Taniguchi, "Thinning and Stroke Segmentation for Handwritten Chinese Character recognition," Pattern Recognition, Vol. 15, No. 4, pp. 299-308, 1982.