

# 필기 한글 문자의 골격선 추출\*

박정선<sup>1</sup>, 홍기천, 오일석  
전북대학교 컴퓨터과학과

Jeong-Sun Park and Ki-Cheon Hong and Il-Seok Oh  
Department of Computer Science, Chonbuk National University, Korea

요약

필기 한글 인식에서 원래 패턴의 모양을 유지하는 골격선 추출은 중요하다. 세선화에 의존하는 기존 방법은 작은 잡음에 민감하다는 단점을 안고 있다. 본 논문은 필기 한글 패턴에 적합한 새로운 골격선 추출 방법을 제안한다. 먼저 한글 패턴은 T-접점과 B-접점이라는 두 가지 모양 특징을 중심으로 분할할 수 있다는 관찰에 근거하여 유사 불록으로 이루어진 부품 집합으로 분할한다. 또한 세 개 이상의 획이 복잡한 형태로 만나는 지점을 결합 부품으로 분할한다. 그런 다음, 각 부품에서 접점의 형태에 따라 결합 부품을 추가 탐지한다. 결합 부품과 인접한 부품들의 연관 관계에 따라 골격선을 구하고, 골격선의 연결성을 보장하기 위해서 선 분장을 수행한다. 본 논문에서 기존의 방법과의 비교를 위해 다섯 가지 비교 기준을 설정하고, 이를 기반으로 비교 분석하였다. 본 논문에서 제안한 방법이 여러 기준에서 세선화-기반 방법보다 우수함을 보였다.

## 1. 소개

한글은 부류수가 방대하고 글자간 유사성이 높기 때문에 문자 패턴을 구성하는 획을 추출하고 이를 분석하여 인식하는 방법이 주류를 이루고 있다 [이은주88]. 현재까지 개발된 필기 한글 인식 방법은 세선화(thinning)와 직선근사(line approximation)에 의존한다 [Kim98, 황영섭97, 강민오97, Kim96a, Kim96b]. 하지만 세선화에 기반한 획 추출 알고리즘을 필기 한글 패턴에 적용하였을 경우 발생하는 정보 손실 및 패턴 왜곡은 문자 인식을 매우 어렵게 만드는 요소이다[Oh98]. 이런 세선화의 단점을 보완하기 위한 방법론이 제시되어왔다 [한동협98, Suh97, Lee95]. 하지만 이러한 방법들 역시 입력 패턴에 대한 제약 조건들로 인한 한계점을 가지고 있다.

세선화에 의존하여 골격선을 추출하는 기존 방법은 작은 잡음에 원영상을 크게 왜곡한다는 단점을 안고 있다. 원 영상을 왜곡시키지 않고 충분한 심보율 표현할 수 있는 골격선 추출은 인식 시스템에서 가장 중요한 요소 중에 하나이다. 기존 화소 단위 처리 방법은 작은 잡음에 민감하기 때문에 새로운 방법이 필요하다.

한글 패턴의 분석 결과 두 개의 획이 T자 형태로 만나는 T-접점(T-junction), 하나의 획이 굽는 B-접점(B-junction), 그리고 세 개 이상의 획을 연결해주는 결합 부품(joint part)을 중심으로 분할할 수 있다는 결론을 얻었다. 본 논문에서는 이런 특징을 이용하여 유사 불록 부품을 분해한 후 각 부품을 대표할 수 있는 선분을 생성하는 방법을 제안한다. 먼저 분할은 두 단계에 걸쳐 수행되어지는데, 첫 번째 단계에서는 Teh-Chin[Teh89], 두 번째 단계에서는 Zimmer[Zimmer97]가 제안한 알고리즘을 이용하여 특징점을 탐지하고 이 점들에 대한 기하학적 조건을 검사하여 T-접점과 B-접점으로 분해하며 그리고 결합 부품을 탐지한다. 그런 다음, 분해된 각 부품에서 접점의 형태에 따라 결합 부품을 추가 탐지한다. 이 결합 부품과 인접한 부품들의 연관 관계에 따라 골격선을 추출하고, 골격선의 연결성을 보장하기 위해 선분 분장을 수행한다.

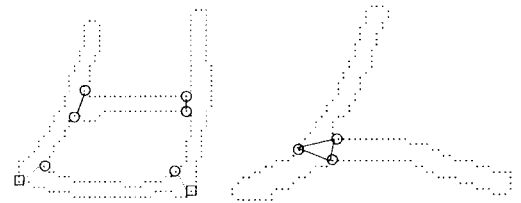
본 논문에서 제안한 방법을 잡음 가지(noise branch) 발생, 획의 직선성(straightness) 유지, 획의 길이 유지, 획 교차(stroke crossing), 복원성(reconstructibility) 등 다섯 가지 기준으로 기존 세선화 방법과 비교 분석한 결과 우수함을 보였다. 2장은 유사 불록 부품으로 분할 방법, 3장은 분할된 부품에서 골격선 추출 방법, 4장은 PE92 데이터베이스에 적용한 결과를 중심으로 기존의 방법과 비교 설명하고, 5장은 결론 및 향후 연구 과제에 대해서 논한다.

## 2. 유사 불록 부품 분할

### 2-1. 한글을 이루는 접점

한글 패턴은 두 가지 특징으로 구분된다. 하나의 획이 꾸부러진 지점(‘ㄱ’, ‘ㄴ’, ‘ㄷ’, ‘ㄹ’, ‘ㅁ’...)과 두 획이 T자 모양으로 만나는 지점(‘ㅏ’, ‘ㅑ’, ‘ㅓ’, ‘ㅕ’, ‘ㅗ’, ‘ㅛ’, ‘ㅜ’, ‘ㅠ’, ‘ㅡ’, ‘ㅣ’)가 있다. 전자의 지점을 B-접점(B junction)이라고 하고, 후자의 지점을 T-접점(T-junction)이라 정의한다. 이 두 지점으로 분할할 경우 불록으로 이루어진 부품인 유사 불록 부품 집합으로 분할되어진다.

그러나 무세약으로 필기된 한글 문자에서는 획이 수직으로 만나지 않고 그림1(b)와 같이 비스듬히 만나는 경우가 빈번하다. 이런 경우는 T-접점과 B-접점으로 분할하기가 자연스럽지 않고 모호성을 발생시킨다. 모호성이 발생하는 지점을 자연스럽게 처리하기 위해 하나의 부품으로 분할하고 이를 결합 부품(joint part)이라 정의한다. 결합 부품은 인접한 획들이 교차하는 한 점을 제공함으로써 골격선을 구할 때 유용하게 사용한다.



a) 접점의 위치(B-접점: 점선 T-접점: 실선)  
b) 결합 부품(joint part)  
그림 1. 한글의 특징 (O: 오목 특징점, □: 불록 특징점)

T-접점은 실선으로 표시된 오목 특징점과 오목 특징점의 쌍으로 이루어지며(그림1(a)), B-접점은 점선으로 표시된 오목 특징점과 불록 특징점의 쌍으로 이루어진다(그림1(a)). 또한 결합부품은 세 개 이상의 오목 특징점의 쌍으로 이루어진다(그림1(b)).

### 2-2. 분할 알고리즘

우세점 탐지 알고리즘 [Teh89]을 사용하여 패턴 외곽선에 존재하는 우세점(dominant point)을 탐지하고 이들을 특징점(feature point)으로 사용한다. T-접점은 탐지된 오목 특징점만을 대상으로 기하학적 조건을 검사하여 T-접점의 쌍을 추출하고, B-접점은 T-접점의 쌍으로 추출된 특징점을 제거하고 남은 특징점을 대상으로 기하학적 조건을 검사하여 B-접점의 쌍을 추출한다. 이를 첫 번째 단계라 한다.

그러나 난립된 쓴 패턴에 대해서는 심한 굴곡과 복잡한 획 연결 등으로 인하여 우세점 탐지 알고리즘만으로는 정확한 특징점 탐지가 불가능하기 때문에 유사 불록 부품으로 분할되지 못한 경우가 다수 발생한다. 이를 보완하기 위하여 불록 쉘(convex hull) 알고리즘[Zimmer97]을 적용하여 특징점을 탐지한다. 첫 번째 단계에서 분할된 부품 각각에 대해 불록 쉘 알고리즘을 적용하여 새롭게 생성된 영역을 오목 영역(concave region)이라 정의한다. 이 오목 영역에서 가장 내부에 있는 점을 탐지하여 오목 특징점으로 간주한다. 오목 특징점이 한 개일 때는 B-접점, 두 개일 때는 T-접점 그리고 세 개 이상일 때는 결합 부품으로 탐지한다. 이를 두 번째 단계라 한다.

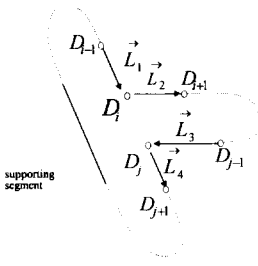
#### 2-2-1 첫 번째 단계

우세점(dominant point) 탐지 방법인 Teh-Chin 알고리즘을 적용하여 방향값이 급격히 변화된 점을 추출한다. 하지만 이 알고리즘은 불필요한 우세점을 발생시키는 단점이 있다. 이 우세점을 좀 미세한 돌출 부분에서 발생한 우세점을 제거하기 위하여 한 우세점에 대해 좌우 이

\*과학재단 핵심전문 과제(981-0919-098-2)의 지원에 의한 연구임.

옳한 우세점과의 각도를 계산한다. 임계값  $T_l$ 를 설정하여 각도가  $T_l$  이상이면 제거시킨다. 실험에서는  $T_l$ 를 145°로 하였다. 제거하고 남은 우세점을 특징점이라 한다. 각 점점에 입력으로 들어갈 점들로 구분하기 위하여 오목성에 따라 오목 특징점과 볼록 특징점으로 구분한다.

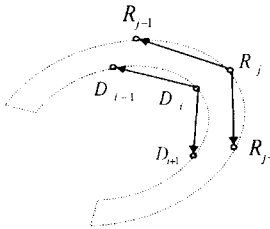
T-점점 검사를 위하여 이렇게 추출된 특징점들 중 오목 특징점만을 대상으로 T-점점에 위치하는지를 검사한다. 총  $n$ 개의 오목 특징점에 대한  $nC_2$ 개를 추출하여 다음 조건을 수행한다. 그림2에 있는 세 가지 조건을 모두 만족하면 T-점점으로 판정한다. 실험에서는  $T_1=10$ ,  $\epsilon_1=35^\circ$ , 그리고  $\epsilon_2=2$ 로 설정하였다.



- 조건1: 근접성(proximity)  
Distance( $D_i, D_j$ ) <  $T_1$
- 조건2: 공선성(collinearity)  
 $0^\circ - \epsilon_1 \leq \text{Angle}(\vec{L}_1, \vec{L}_4) \leq 0^\circ + \epsilon_1$  AND  
Distance( $\vec{L}_1, D_j$ ) <  $\epsilon_2$  AND  
Distance( $\vec{L}_4, D_i$ ) <  $\epsilon_2$
- 조건3: 반 평행성(antiparallelism)  
 $180^\circ - \epsilon_1 \leq \text{Angle}(\vec{L}_2, \vec{L}_3) \leq 180^\circ + \epsilon_1$

그림 2. T-점점 조건

B-점점 검사를 위하여 T-점점으로 판정된 특징점을 제거하고 남은 특징점을 대상으로 그림3의 조건을 적용하여 수행한다. 실험에서는  $T_2=145^\circ$ ,  $T_3=10$ ,  $\epsilon_3=10^\circ$ 로 설정하였다.



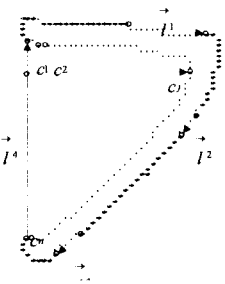
- 조건1: 안쪽 각도(inner angle)  
 $\angle D_{i-1}D_iD_{i+1} < T_2$
- 조건2: 바깥쪽 각도(outer angle)  
 $\angle D_{i-1}D_iD_{i+1} - \epsilon_3 < \angle R_{j-1}R_jR_{j+1} < \angle D_{i-1}D_iD_{i+1} + \epsilon_3$
- 조건3: 근접성(proximity)  
Distance( $D_i, R_j$ ) <  $T_3$

그림 3. B-점점 조건

2-2-2 두 번째 단계

정서체로 쓰여진 패턴은 첫 번째 단계에서 각 부품들이 유사 볼록으로 이루어진다. 하지만 난립로 쓰여진 데이터는 우세점 탐지 알고리즘만으로 정확한 특징점 탐지가 불가능하기 때문에 유사 볼록으로 부합되지 않는 부품이 다수 발생한다. 이런 이유로 두 번째 단계에 들어가 앞서 각 부품에 대한 유사 볼록으로 이루어졌는지를 판별할 필요가 있다. 또한 유사 볼록 부품으로 이루어진 부품에 대해서는 두 번째 단계를 수행하지 않으므로써 수행속도를 높일 수 있다.

하나의 부품에 대한 유사 볼록 검사는 다음과 같다. 각 부품에 대해 볼록 힐 알고리즘 [Zimmer97]을 적용하여 새롭게 생성된 영역을 구한다. 이 영역을 오목 영역이라 한다. 원 부품의 영역과 오목 영역으로 구분할 수 있는데, 유사 볼록 유부는 부품 영역과 오목 영역 면적의 비율을 계산하여 판별한다. 임계값  $T_b$ 를 설정하여 비율값이 임계값  $T_b$ 를 넘으면 유사 볼록하지 않다고 결정한다. 실험에서는  $T_b$  값으로 1/3을 사용하였다. 앞에서 유사 볼록하지 않다고 판정된 부품에 대해 오목 특징점 탐지를 시도한다.



- 오목 특징점 조건:  
length( $\vec{l}$ )  $\geq T_b$  AND  
maximum( $\vec{l}, C_j$ )  $\geq T_b$

그림 4. 오목 영역이기 위한 조건

그림4의 두 조건을 만족하면 오목영역이라 판정하고 가장 내부점을 추출한다. 이를 오목 특징점이라 간주한다. 실험에서는  $T_5=7$ ,  $T_6=5$ 로 사용하였다. 그림4는 네 개의 벡터를 중심으로 네 개의 오목영역이 존재하는데, 벡터  $\vec{l}$ 를 제외한 세 개의 벡터는 위의 두 조건을 만족하지 않으므로 무시한다. 그림4에서는 하나의 오목 특징점을 생성하지만, 경우에 따라서 여러 개의 오목 특징점이 탐지 될 수 있다. 이런 경우를 세 가지 경우로 압축할 수 있는데, 오목 특징점이 하나일 때 반대편의 볼록 특징점을 탐지하여 B-점점으로 분할하고, 두 개일 때 두 오목 특징점을 쌍으로 하는 T-점점으로 분할하고, 그리고 세 개 이상일 때 세 개 이상의 오목 특징점을 한 쌍으로 하는 결합 부품으로 분할한다.

3. 골격선 추출

3-1 각 부품의 형태

그림5는 유사 볼록 부품으로 분해된 최종 결과이다. 유사 볼록으로 분할된 부품을 분석해 보면 다음과 같이 나눌 수 있다. 점점을 하나만 가진 부품(그림5의 'a', 'b', 'c'로 표시), 부품 영역의 끝이 아닌 중간 변에 T-점점이 존재하는 부품 그리고 두 개 이상의 점점을 포함하는 부품(그림5의 'i', 'o', 'u', 'l', 's', 't', 'f', 'r'로 표시)들이 있다.

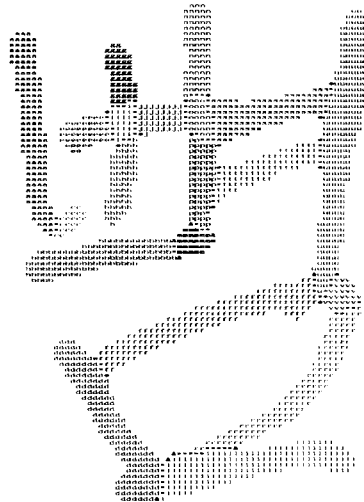


그림 5. 분할 결과 (&: T-점점 끝점, @: B-점점의 끝점)

첫 번째는 부품 양 끝점을 추출하여 선분 생성한다. 두 번째 경우는 부품 양 끝점을 추출하여 선분을 생성하고 T-점점 부분에서 선분의 연장이 이루어진다. 하지만 세 번째의 경우 살펴보자. 첫 번째, 그림5의 'i', 'o' 부품은 점점으로 둘러싸여 있다. 두 번째, 그림5의 'u'의 하단 부분, 'l'은 세 개 이상의 점이 한 점을 중심으로 교차하는 형태를 이룬다. 이 부분의 처리를 용이하게 하기 위하여 첫 번째 경우는 그 부품을 결합 부품으로 간주하고 두 번째 경우는 사실상 결합 부품은 아니지만 연결되지 않는 부분을 가상의 선으로 연결함으로써 새로운 결합 부품을 추출한다. 이 부품은 직접 선분을 구하지 못하고 인접한 확률이 공유하는 끝점으로 사용되고 인접 부품들과 관련하여 선분을 구한다. 앞에서 살펴본 다양한 경우를 고려하여, 우선 결합 부품을 탐지하고 각 부품에 대한 선분을 추출한다.

3-2 비 결합 부품을 결합 부품으로 탐지

우선 분해된 패턴의 점점들을 조사하여 끝점을 공유하는 경우를 모두 찾는다. 이 때 끝점을 공유하는 점점들이 폐곡선을 이루는지 여부에 따라 폐곡선으로 이루어졌다면 그 자체를(그림5의 'i', 'o'), 그렇지 않다면 두 개의 끝점에 새로운 분할 경로를 삽입하여 새로운 부품을 만들고 이 부품을 결합 부품으로 결정한다.(그림5의 'u'부품의 하단 부분)

끝점을 공유하지는 않지만 결합 부품으로 탐지해야하는 상황이 있다(그림5의 'l'부품). 다음 두 가지 조건을 검사하여 결합 부품을 탐지

한다. 첫 번째 두 개 점점간의 거리( $T_1$ )를 계산한다. 두 번째 두 개의 점점을 이어 새롭게 생성된 면적( $T_2$ )을 계산한다. 일제값  $T_1$ 과  $T_2$ 보다 작으면 결합 부품으로 일제한다. 실험에서 일제값  $T_2$ 은 평균 회 두께 값으로 사용하였고, 일제값  $T_1$ 는 원 부품의 면적에 대한 생성된 면적의 비율값 0.4로 하였다.

3-3 각 부품에서의 골격선 추출

골격선은 부품에 결합 부품의 연결 유무에 따라 다르게 추출한다. 결합 부품과 연결된 부품은 다음 세 가지 경우가 있다. 첫째 그림5의 's', 'p', 'f', 'v', 'r' 부품은 결합 부품의 중심과 다른 반대편의 점점의 중점을 두 끝점으로 하는 선분을 생성한다(그림6의 '1'로 표시). 둘째 그림5의 'j' 부품은 두 개의 결합 부품의 중심을 두 끝점으로 하는 선분을 생성한다(그림6의 '2'로 표시). 셋째 그림5의 'g', 'e', 'h', 'n', 'u', 'd', 'l' 부품은 결합 부품의 중심과 다른 반대편의 가장 먼 점을 끝점으로 하는 선분을 생성한다(그림6의 '3'로 표시).

결합부품과 연결되지 않는 부품은 다음 세 가지 경우가 있다. 첫째 그림5의 'm', 't' 부품은 각 점점의 중점을 양 끝점으로 하는 선분을 생성한다(그림6의 '4'로 표시). 둘째 그림2의 'a', 'b', 'c' 부품은 점점의 중점과 그 점에서의 가장 먼 거리에 있는 점을 끝점으로 하는 선분을 생성한다(그림6의 '5'로 표시). 셋째 아무런 점점을 포함하지 않는 부품은 거리가 가장 먼 점을 끝점으로 하는 선분을 생성한다.

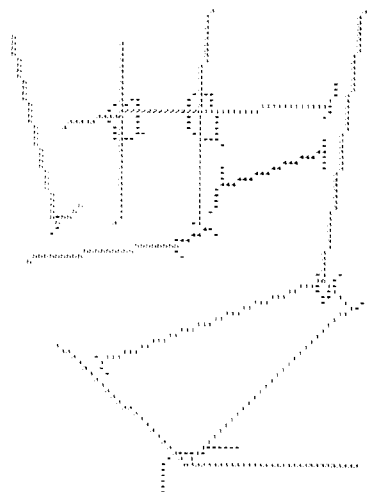


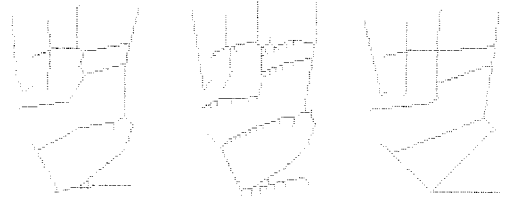
그림 6. 선분 생성 (\*: 점점, 숫자: 부품 형태에 따른 선분)

이제까지 모든 부품에 대해 그것을 대표하는 선분을 추출하였다. 이렇게 추출된 선분은 '1' 점점 부분에서 끊긴 현상이 발생한다(그림6의 'n', 'r'의 좌상 부분). 추출된 선분이 원래 패턴의 연결성을 그대로 유지하기 위해서는 선분 연결이 필요하다. 이를 위하여 점점의 방향을 이용하여 점점에 인접한 두 개의 부품 중, 점점의 방향과 유사한 부분을 도차 부품, 나머지 부분은 출발 부품으로 정한다. 출발 부품에 해당하는 선분의 끝점을 도차 부품에 해당하는 선분까지 연장하여 완전한 골격선을 추출한다.

4. 골격선 비교 실험

본 논문에서 결과로 나온 골격선을 기존의 골격선 추출 방법(STPA, Zhang-Suen)과 다섯 가지 기준으로 비교 실험하였다. 첫째 불필요한 잡영 가지를 발생시키지 않는다. 둘째 기존의 방법은 T-점점부분에서 Y패턴이 발생하여 세 개의 선분이 추출되지만, 제안한 방법은 원래 패턴의 직선성을 그대로 유지하고 있다. 셋째 기존 방법은 화소를 제거하는 방법으로 입력 패턴이 가지고 있던 것보다 짧게 추출되지만, 제안한 방법은 부품의 양 끝점을 찾아 선분을 계산하므로 원래 회 길이 그대로 유지한다. 넷째 교차 지점에서의 기존 방법은 세 개 혹은 그 이상의 회으로 분리되지만, 제안한 방법 교차하는 지점에서 결합 부품을 탐지하고 이 결합 부품이 인접한 부품들이 공유하는 한 개의 점을 세공하므로써 네 개의 선분이 연결되는 형태로 명확하게 출력된다.

다섯째 선분의 개수는 원 패턴의 모양을 그대로 유지하면서 기존의 방법보다 적은 개수로 추출되었다. 이는 인식기 측면에서 보면 큰 향상이라고 말할 수 있다. 그림7은 각 알고리즘의 결과이다.



a) STPA 세선화 b) Zhang-Suen 세선화 c) 본 논문 방법+ 그림 7. 세 가지 방법으로 추출된 골격선

5. 결론 및 향후 연구과제

본 논문에서는 먼저 유사 불록 부품으로 분할한 후 골격선 추출 방법을 제안하였다. 기존의 방법과 잡영 가지의 발생, 회의 직선성 유지, 회의 길이 유지, 회 교차, 복원성과 같이 다섯 가지의 기준을 가지고 비교해 본 결과 본 논문에서 제안한 골격선 추출 방법이 기존의 방법보다 우수함을 보였다. 따라서 본 논문에서 제안한 방법이 인식기의 입력으로 들어갈 경우 인식률 향상이 있을 것으로 기대된다.

향후 연구 과제로는, 추출한 골격선을 이용하여 인식 방법을 개발하는 것이며 기존 골격선을 이용한 인식기에서 골격선 추출모듈만을 본 논문의 모듈로 대체하여 인식률 향상을 측정해 보는 것이다.

참고문헌

[강민오97] 강민오, 조형세, "속성 그래프를 이용한 오프라인 필기체 한글 인식," 한국정보과학회 '97 봄 학술 발표 논문집(B), 한림대학교, pp.525-528, 1997.  
 [이은주88] 이은주, 권오석, 김태관, "필기체 한글에서 자도 분리와 인식," 한국정보과학회 논문지, Vol.15, No.6, pp.526-534, 1988.  
 [한동협98] 한동협, 조성배, "오프라인 필기체 한글 영상의 세선화를 위한 지식기반 시스템에 관한 연구," 제10회 영상처리 및 이해에 관한 워크샵 발표 논문집, 세우, pp.163-168, 1998.  
 [황영섭97] 황영섭, "제약조건 만족 그래프를 이용한 필기 한글인식," 박사학위 논문, 포항공과대학, 1997.  
 [Kim96a] W.S. Kim and R.H. Park, "Off-line recognition of handwritten Korean and alphanumeric characters using hidden markov models," Pattern Recognition, Vol.29, No.5, pp.845-858, 1996.  
 [Kim96b] H.J. Kim and P.K. Kim, "Recognition of off-line handwritten Korean characters," Pattern Recognition, Vol.29, No.2, pp.245-254, 1996.  
 [Kim98] H.Y. Kim and J.H. Kim, "Handwritten Korean character recognition based on hierarchical random graph modeling," Sixth International Workshop on Frontiers in Handwriting Recognition, Taejon, Korea, pp.577-586, August 1998.  
 [Lee95] S.W. Lee and Y.J. Kim, "Direct extraction of topographic features for gray scale character recognition," IEEE Tr. on Pattern Analysis and Machine Intelligence, Vol.17, No.7, pp.724-729, July 1995.  
 [Oh98] I.S. Oh, J.S. Lee, T.W. Yoo, J.S. Park, and K.C. Hong, "Region based shape decomposition algorithm for Hanguk character patterns," Sixth International Workshop on Frontiers in Handwriting Recognition, Taejon, Korea, pp.557-566, August 1998.  
 [Suh97] J.W. Suh and J.H. Kim, "Stroke extraction from gray-scale character image," Progress in Handwriting Recognition (Ed. by A.C.Downton and S.Impedovo), World Scientific, pp.593-598, 1997.  
 [Teh89] C.H. Teh and R.T. Chin, "On the detection of dominant points on digital curves," IEEE Tr. on Pattern Analysis and Machine Intelligence, Vol.11, No.8, pp.859-872, August 1989.  
 [Zimmer97] Y. Zimmer, R. Tepper, and S. Akselrod, "An improved method to compute the convex hull of a shape in a binary image," Pattern Recognition, Vol.30, pp.397-402, 1997.