

# 필기 한글 문자의 모양 분석과 표현

홍기천<sup>†</sup>·오일석<sup>††</sup>

## 요약

본 논문에서는 필기 한글의 모양을 분석하고 그 결과를 표현하는 방법을 제안한다. 하나의 한글 문자를 구성하는 연결 요소 각각은 다수 개의 부품들로 분해되고, 분해된 요소들을 표현하기 위해 골격선을 추출하였다. 추출된 골격선으로부터 속성 그래프를 이용하여 필기 한글 문자를 표현하였다. 속성 그래프의 노드는 하나의 부품에 해당하고 에지는 그들간의 연결 관계를 나타낸다. 노드와 에지에는 패턴의 모양을 나타내는 여러 가지 유용한 정보를 저장하였다.

## Shape Analysis and Representation of Handwritten Hangeul Characters

Ki-Cheon Hong<sup>†</sup>·Il-Seok Oh<sup>††</sup>

### ABSTRACT

This paper proposes a method of shape analysis and representation for the handwritten Hangeul character patterns. Each of the connected components composing a Hangeul character is decomposed into many parts, and skeletons are extracted from the decomposed parts. Using the results, we represent the shape of Hangeul characters using the attributed graph representation. A node of the attributed graph represents a part and an edge represents their relationships and they store valuable informations of the pattern shapes.

### 1. 서론

한글 문자를 인식하는 방법은 크게 두 가지로 나뉜다. 첫 번째는 문자 기반 인식 방법[1]이고 두 번째는 자소 기반 인식 방법[2, 3]이 있다. 각각의 방법은 나름대로의 장단점을 가지고 있다. 문자 기반 인식 방법은 문자를 자소들로 분리하는 작업이 필요 없다는 장점이 있다. 그러나 인식해야 할 문자의 수가 11,172개에 달하기 때문에 많은 문제를 내포하고 있다. 반면, 자소 기반 인식 방법은 문자를 초성, 중성, 종성으로 분리하는 부담을 가지고 있기 때문에, 분리 과정에서 초래되는 에러는 인식률에 나쁜 영향을 미칠 수 있다.

두 가지 방법 모두에서 한글 문자 패턴을 여러 개의 부품들로 분해하는 작업은 매우 중요하다. 또한 모양 분해는 한글 단어의 문자별 분리, 한글 문자의 자소 분리 등의 여러 작업에 있어서도 매우 중요하다. 워드 인식의 첫번째 단계는 워드를 개개의 낱자로 분리하는 작업이다. 이러한 작업을 위해서, 두개 문자 사이의 분할 경로를 결정하기 위한 휴리스틱한 방법이 제안되었다[4]. 그러나, 필기 한글 단어에서는 접촉유형이 많이 존재하기 때문에 휴리스틱한 방법은 한계가 있다.

기존의 한글 문자 인식 방법은 전처리로서 세선화 과정을 사용한다. 이은주[5]는 전처리 과정으로서 입력 패턴에서 골격선을 추출한다. 이러한 골격선으로부터 끝점, 교차점과 같은 특징점을 추출한다. 그 다음, 수직 모음 추정, 초성 인식, 수직 모음 복원, 수평 모음 인식, 수직 모음 인식, 종성 인식의 단계로 한글 문자

\* 이 연구는 한국과학재단 핵심과제 981-1212-036-2 지원을 받았다.

† 준회원 : 전북대학교 대학원 전산통계학과

†† 정회원 : 전북대학교 컴퓨터학과 교수

논문접수 : 1999년 7월 8일, 심사완료 : 2000년 4월 17일

를 인식한다. 이 방법은 수직 모음 추정에 있어서 세선화 과정을 사용하였기 때문에 PE92 데이터베이스와 같은 무계약 필기 문자의 경우 나쁜 성능을 보여줄 수 있다. 김향준[3]은 골격선으로부터 끝점, 분기점, 교차점과 같은 특징들을 추출하였다. 인식 과정의 첫 단계는 세선화된 한글 문자의 오른쪽위 부분의 끝점으로부터 수직 모음을 인식하는 것이다. 획간의 접촉 유형 300개와 11개의 분리 규칙을 사용하여, 입력 패턴을 28가지의 요소로 분리한다. 그 다음, 분리된 요소의 상대적 위치를 일련의 코드로 변환한다. 이러한 방법은 미리 정의된 접촉 유형과 분리 규칙이 무계약 필기체와 같이 거칠게 필기된 경우에 적용되지 않을 수 있다.

위에서 살펴본 여러 작업을 효과적으로 수행하기 위해서는 한글 문자의 모양을 분석하고 표현하는 것이 매우 중요하다. 본 논문에서는 필기 한글의 모양 표현 방법을 제안한다. (그림 1)은 본 논문에서 채택한 한글 문자 처리의 전체 과정을 보여준다. 모양 분해 알고리즘을 사용하여 한글 문자를 여러 개의 유사 볼록(near convex)한 부품으로 분해한다. 유사 볼록한 부품들을 얻기 위해서, T-접점과 B-접점이라는 두개의 모양 특징을 정의한다. 이 특징들은 특징점 탐지, T-접점 탐지, B-접점 탐지, 유사 볼록 검사, 오목 영역 중심으로 분할과 같은 다섯 단계를 거쳐서 추출된다. 검출된 접점 위치에서 문자의 획을 절단하면 볼록한 부품들을 얻을 수 있다[6].



(그림 1) 한글 문자 처리의 전체 과정

이렇게 얻어진 부품들의 집합을 가지고 골격선을 추출하기 위하여 부품들 중 결합 부품을 탐지한다. 탐지된 결합 부품을 기반으로하여 결합 부품과 인접한 부품의 골격선 추출, 결합 부품과 인접하지 않은 부품의 골격선 추출, 그리고 선분 연장과 같이 세 단계를 수행하여 한 픽셀 두개의 선분 집합으로 표현되는 골격선을 추출한다[7].

모양 분해 방법은 입력된 원래 패턴을 구성하는 획을 아무 손상 없이 명시적으로(explicitly) 분해하여 유사 볼록 부품으로 만들어 놓기 때문에, 각 부품에서 특징을 추출하는 것이 용이하다. 모양 분해 알고리즘을 이용하여 추출된 골격선은 세선화 기반 골격선 추출 방법보다 많은 장점을 가지고 있다. 그중 가장 두

드러진 장점을 들면 다음과 같다. 세선화 방법은 기본적으로 획의 외곽으로부터 화소를 제거하며 획의 중앙선을 찾아가는 연산을 수행함으로 획의 길이가 원래 입력 패턴이 가지고 있던 것보다 짧게 추출된다. 그러나 모양 분해를 이용하여 추출된 골격선은 분해된 부품의 외곽선상에 있는 점들 중에서 양 끝점을 찾아 선분을 계산하기 때문에 원래 획 길이를 유지한다. 한글에서는 '丨'와 '丨'의 수평획등이 필기시 짧게 쓰여지는 경우가 많다. 이런 경우 골격선 추출과정에서 획 길이가 짧아진다면 이들 획이 잡영가지로 처리되어 오인식을 발생시킬 가능성이 있다. 또한 골격선은 원래 패턴 영상이 가지고 있는 정보를 손실한다. 그러나 모양 분해를 이용한 골격선 추출 방법은 분해된 부품으로부터 골격선을 추출하기 때문에 필요하면 언제든지 원 영상 내에서 부품 영역을 참조할 수 있다. 때문에 필요한 경우에는 해당 부품 영역을 참조하여 다양한 특징을 추출할 수 있다. 특히 '口'와 'ㅇ'을 구별하는데 유용한 획의 굴곡성을 추출하는데 사용할 수 있다. 이와 같이 모양 분해를 이용하여 추출된 골격선은 세선화 기반 골격선 추출 방법보다 많은 장점을 가지고 있기 때문에, 높은 인식률을 갖는 한글 인식기 구축에 유용하게 사용할 수 있다. 하지만 모양 분해 및 골격선 추출 결과를 효율적이고 체계적으로 표현하는 작업이 남아있다. 본 논문은 이 작업을 연구대상으로 삼고 있다.

한글보다 더 복잡한 구조를 가진 한문을 표현하는 방법에 대해서는 많이 연구되어 왔다. Suen[8]은 계층적 속성 그래프(hierarchical attributed graph)를 이용하여 필기 한문 문자를 표현하였고, 높은 인식률을 보이는 인식기를 구축하였다. 이 방법은 문자를 연결 요소로 나눈 다음, 연결 요소를 속성 그래프로 표현하였다. 각각의 연결 요소에서 끝점, 교차점, 굴곡점과 같은 특징점을 이용하여 문자의 획을 추출하였다. 추출된 획에 대해 수평/수직-오른쪽-대각선/왼쪽-대각선과 같은 방향 속성과 선분/점과 같은 길이 속성을 부여하였다. 이렇게 속성화된 획을 이용하여 연결 요소에 대한 속성 그래프를 구한 다음, 이를 조합하여 입력 패턴에 대한 계층적 속성 그래프를 구성하였다. 이 방법은 간단하면서도 다양한 필기자의 필기 습관에 둔감하게 작용하는 장점이 있다.

본 논문에서는 필기 한글 패턴을 모양 분해하고 골격선을 추출한 결과를 효과적으로 표현하는 방법을 제안한다. 이를 위해 [8]에 기술된 계층적 속성 그래프와 인

집 행렬을 도입하고 이를 모양 분해된 한글에 맞게 수정하였다. 계층적 속성 그래프의 구성 과정은 크게 두 계층으로 나뉜다. 첫 번째 계층은 문자를 연결 요소로 나눈 다음, 각각의 연결 요소를 표현하는 속성 그래프를 구성한다. 두 번째 계층은 문자 전체에 대한 속성 그래프를 구성한다. 한 개의 연결 요소에 대한 표현을 속성 획 그래프(Attributed stroke graph ; ASG) 방법, 그리고 문자 전체에 대한 표현을 속성 계층 그래프(Attributed hierarchical graph ; AHG)이라 정의한다. 즉, 여러개의 ASG가 하나의 AHG를 구성한다. ASG와 AHG는 인접 행렬(adjacency matrix)형태로 기술한다.

제2장에서는 영역 기반 모양 분해 알고리즘과 골격선 추출 알고리즘에 대해서 기술한다. 제3장에서는 본 논문에서 제안한 모양 표현 방법을 기술하고, 제4장에서는 결론과 향후 연구 과제에 대해서 논한다.

## 2. 모양 분해 및 골격선 추출

### 2.1 영역 기반 모양 분해 알고리즘

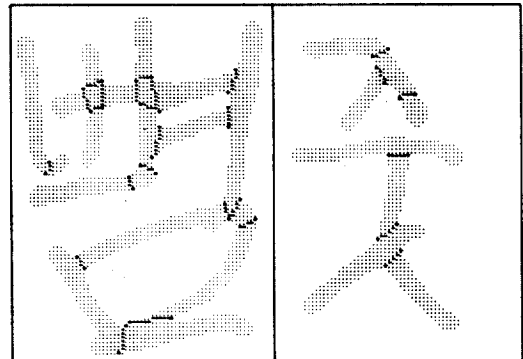
이 장에서는 모양 분해 알고리즘에 대해서 기술한다. 한글 입력 패턴을 불록한 부품들의 집합으로 분해하는 것은 매우 중요한 작업이다. 본 논문에서는 세션화 연산과정 중에 초래되는 부작용을 줄이기 위해 [6]에서 제안한 모양 분해 알고리즘을 도입하여 사용한다. 알고리즘은 다음의 다섯 단계로 구성되어 있다.

- (1) **특징점 탐지** : 입력 패턴으로부터 체인 코드를 구한 다음, 이 체인 코드에 우세점(dominant point) 탐지 알고리즘을 적용하여 방향이 급격히 변하는 점을 추출한다.
- (2) **T-접점 탐지** : T-접점은 ‘ㅌ’, ‘ㄴ’, ‘ㄷ’ 등과 같은 패턴내의 두 개의 획이 서로 T형태로 만나는 지점에서 형성된다. T-접점은 오목 특징점과 오목 특징점의 쌍에 대해 기하학적 조건을 검사하여 탐지한다.
- (3) **B-접점 탐지** : B-접점은 ‘ㄱ’, ‘ㄴ’, ‘ㄷ’, ‘ㄹ’ 등과 같은 패턴내에서 방향이 급격히 변하는 지점에서 생성된다. B-접점은 오목 특징점과 불록 특징점의 쌍에 대해 기하학적 조건을 검사하여 탐지한다. 이렇게 탐지된 T-접점과 B-접점을 중심으로 분할한다.
- (4) **유사 불록 검사** : 이렇게 얻어진 부품들이 유사 불록인지 판정하기 위하여 불록 형 알고리즘을 사용

하여 검사를 수행한다. 유사 불록하지 않다고 판정된 부품에 대해 다음 단계를 이용하여 분할을 한번 더 시도한다.

- (5) **오목 영역 중심으로 분할** : 오목한 영역의 가장 깊은 점을 중심으로 분할을 수행한다. 오목한 영역이 한 개일 때는 B-접점으로 분할하고, 두 개 이상일 때는 T-접점으로 분할한다.

(그림 2)는 위에 기술한 모양 분해 알고리즘으로 분할된 문자 패턴의 예를 보여준다. T-접점의 끝점은 ‘T’, B-접점의 끝점은 ‘B’, 분할경로는 ‘\*’로 표시하였다. ‘뱀’자는 두 개의 연결 요소로 구성되며 각각 2개와 16개의 부품으로 분해되었다. ‘웃’자는 2 개의 연결 요소로 구성되는데 각각 4개와 5개의 부품으로 분해되었다.



(그림 2) 모양 분해 결과

### 2.2 골격선 추출

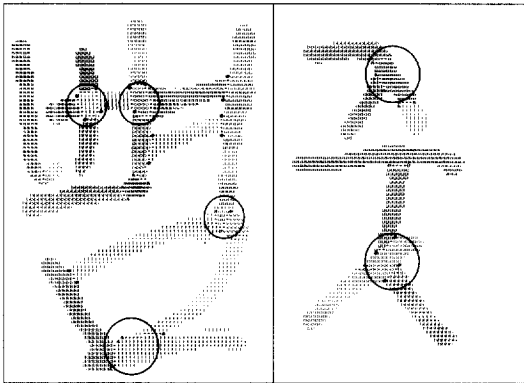
이 장에서는 골격선 추출 알고리즘에 대해서 기술한다. 알고리즘은 크게 결합 부품 탐지와 선분 생성의 두 단계로 나누어진다. 자세한 내용은 [7]을 참조하기 바란다.

- (1) **결합 부품 탐지** : 결합 부품을 효과적으로 탐지하기 위해 두 가지 경우로 나누었다. 첫 번째는 두 개 이상의 접점(T-접점 또는 B-접점)이 끝점을 공유하는 경우이다. 두 번째는 끝점 공유는 없지만 접점간의 거리가 가까운 경우이다.

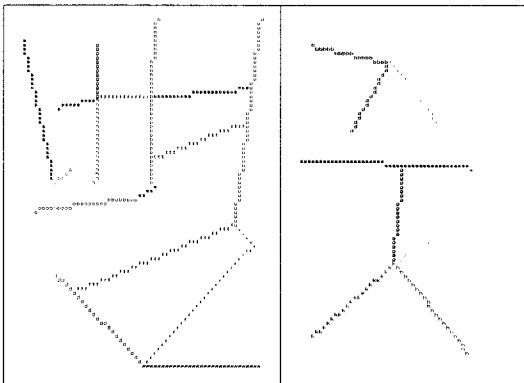
첫 번째 경우는 (그림 3(a))의 ‘뱀’자에서 부품 ‘ㄱ’과 ‘ㅇ’을 둘러싸고 있는 네 개의 접점이 폐곡선을 이루고 있는 경우와 부품 ‘u’의 아래 부분에 두 개의 접점이

한 개의 끝점을 공유하고 있는 경우이다. 원으로 표시한 부분이 끝점 공유를 통한 결합 부품 탐지의 예이다.

두 번째 경우는 (그림 3(a))의 '긷'자의 'ㄷ'과 'ㅅ'이 끝점을 공유하지는 않지만 결합 부품으로 탐지해야 하는 경우이다. 원으로 표시한 부분이 거리에 의한 결합 부품의 탐지 예이다. '@'은 T-점점의 끝점, '&'는 B-점점의 끝점, 그리고 '\*'은 분할 경로이다.



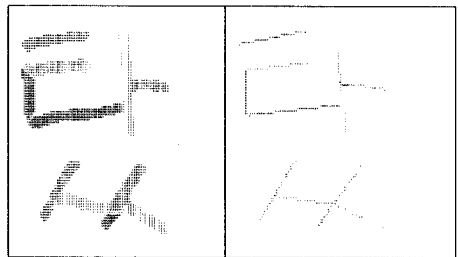
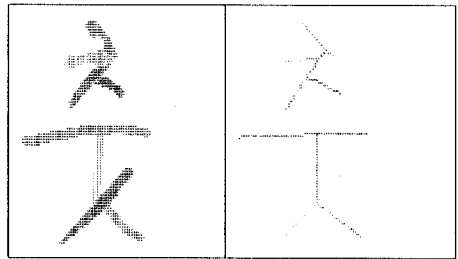
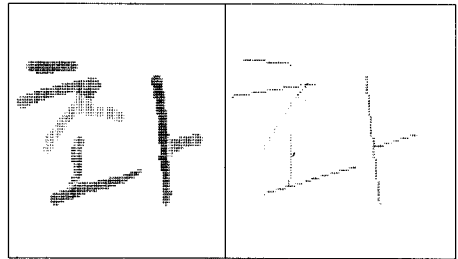
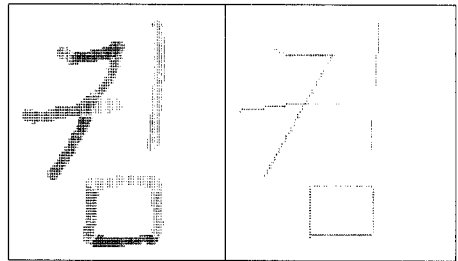
(a) 결합 부품



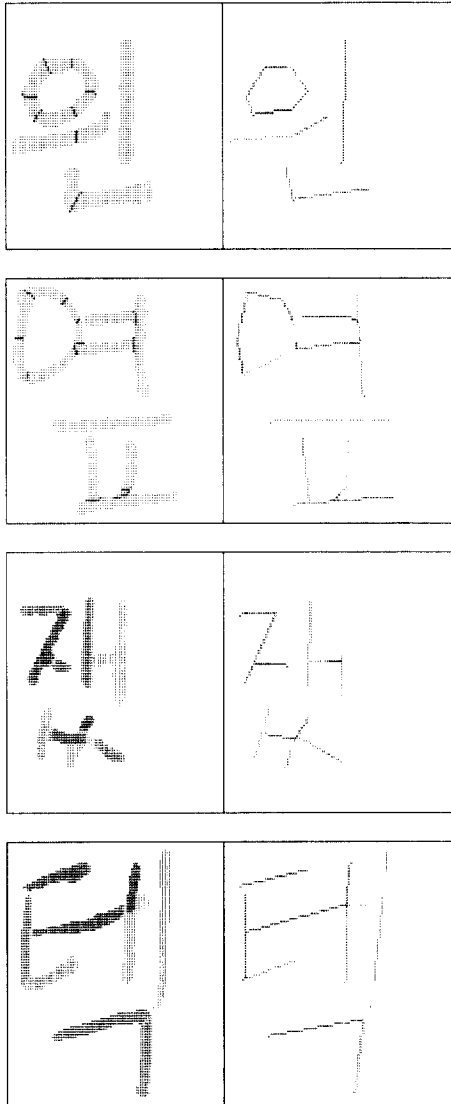
(b) 골격선

(그림 3) 골격선 추출

'뽕'자에서 'a', 'c', 'b', 'm', 't'와 '긷'자에서 부품 'i', 'a')으로 구분하여 처리한다. 이 과정은 크게 결합 부품과 인접한 부품에서 선분 추출, 결합 부품과 인접하지 않은 부품들에서 선분 추출, 그리고 선분들이 T형태로 접촉하는 부분에서 골격선의 연결성 확보를 위한 선분 연장과 같이 세 단계로 구성된다. (그림 3(b))는 이렇게 추출한 골격선을 보여준다. (그림 4)는 다양한 접촉 유형을 갖는 문자들에 대한 모양 분해된 문자와 이에 대한 골격선 예제를 보여준다.



(2) **선분 생성**: 이 과정은 위에서 추출한 결합 부품 (그림 3(a))을 이용하여, 각각의 부품에서 골격선 표현을 구성할 선분을 추출하는 단계이다. 이 과정을 효과적으로 수행하기 위해 부품을 결합 부품에 인접해 있는 지에 따라 결합 부품과 인접한 부품 (그림 3(a)의 '뽕'자에서 부품 'g', 'e', 'h', 'j', 'n', 'p', 's', 'u', 'f', 'v', 'd', 'r', 'l'과 '긷'자에서 부품 'b', 'd', 'e', 'g', 'c', 'h')과 인접하지 않은 부품(그림 3(a)의



(그림 4) 다양한 접속 유형을 갖는 한글 문자의 골격선 추출 예제

### 3. 필기 한글 문자의 모양 표현

본 장에서는 제2장에서 추출한 골격선을 기반으로 하여 필기 한글 문자를 표현하는 방법을 기술한다. 한 개의 한글 문자는 여러개의 연결 요소로 구성된다. 즉, 한글 문자를 표현하기 위해서 연결 요소를 먼저 표현하여야 한다. 연결 요소는 속성 획 그래프(Attributed stroke graph ; ASG)로 표현한다. 이 방법은 골격선의 한 직선

선분을 노드(node), 직선 선분간의 연결 관계를 에지(edge)로 하는 그래프이며 노드와 에지에 다양한 속성을 부여하였다. ASG를 기술하기 위하여 노드와 에지를 서로 다른 자료구조를 사용하여 저장한다. 노드는 9-tuple의 구조체에 저장하고 에지는 연결 요소내의 직선 선분간의 관계가 가지는 속성을 일련의 0과 1로 표현하여 인접 행렬형태로 저장한다. 문자 전체는 속성 계층 그래프(Attributed hierarchical graph ; AHG)로 표현한다. 이 AHG는 연결 요소를 노드, 연결 요소간의 상호 위치 관계(spatial relationship)를 에지로 하는 그래프이다. AHG도 ASG와 같이 에지를 모양 행렬을 이용하여 표현한다.

#### 3.1 부품 표현

(그림 3(b))의 ‘뽕’자와 ‘춧’자는 각각 18개, 9개의 직선 선분을 가지고 있다. 한 직선 선분  $P_i$ 에 해당하는 노드가 가질 수 있는 속성은 두 끝점의 좌표( $x_1, y_1, x_2, y_2$ ), 두 끝점의 유형, 방향( $\theta$ ), 획 두께, 그리고 중요도이다. 즉, 하나의 노드는 아래와 같이 표시할 수 있다.

- 노드 ::= <끝점 1의 좌표( $x_1, y_1$ ), 끝점 2의 좌표( $x_2, y_2$ ),  
 끝점 1의 유형, 끝점 2의 유형, 방향, 두께,  
 중요도>
- 끝점유형 = {O, B, T, J}
- 방향 = {H, V, RD, LD}
- $0.0 \leq \text{중요도}(S) \leq 1.0$

끝점 유형은 네 가지인데, 그 중 O(open)는 (그림 3(b)) ‘춧’의 직선 선분 ‘a’의 양 끝점과 같이 다른 선분과 연결되어 있지 않은 끝점을 나타낸다. B(B-접점)는 ‘ㄱ’, ‘ㄴ’, ‘ㄷ’등에 나타나는 선분의 방향이 급격히 변하는 지점이다. (그림 3(b)) ‘뽕’의 선분 ‘a’의 아래 끝점은 선분 ‘c’의 한 끝점과 B-접점으로 연결되어있다. 이와 같이 선분의 한 끝점이 B-접점상의 점일 때 끝점의 유형을 B라 한다. T(T-접점)는 ‘ㄱ’, ‘ㄴ’, ‘ㄷ’등과 같이 선분이 T형태로 만나는 지점이다. (그림 3(b)) ‘뽕’의 선분 ‘t’의 두 끝점이 한 끝점은 선분 ‘p’, 그리고 나머지 끝점은 선분 ‘u’와 모두 T-접점으로 만나고 있다. J(결합 부품)는 결합 부품의 중점을 의미한다. (그림 3(b)) ‘뽕’의 선분 ‘e’, ‘g’, ‘h’들이 공통적으로 공유하고 있는 점이 결합 부품의 중점이다. 따라서 직선 선

분 'e'의 한 끝점 유형은 O이고 나머지 끝점의 유형은 J이다.

직선 선분의 방향은 H(수평), RD(오른쪽 대각선), LD(왼쪽 대각선), V(수직)의 네 가지이다. 다음은 방향의 계산 방법이다.

$$\text{방향}(\theta) = \tan^{-1} \left( \frac{x_2 - x_1}{y_2 - y_1} \right)$$

$$H : 0^\circ \leq \theta \leq 22.5^\circ \text{ 또는 } 157.5^\circ \leq \theta \leq 180^\circ$$

$$RD : 22.5^\circ \leq \theta \leq 67.5^\circ$$

$$LD : 112^\circ \leq \theta \leq 157.5^\circ$$

$$V : 67.5^\circ \leq \theta \leq 112.5^\circ$$

두께는 부품의 평균 획 두께를 의미한다. 부품의 획 두께는 부품의 넓이를 길이로 나누어 계산한다. 넓이는 부품을 구성하는 화소의 개수, 그리고 길이는 그 부품에 해당하는 직선 선분의 길이로 한다.

중요도는 선분의 길이가 길수록 커진다. 중요도의 값은 0.0에서 1.0사이의 값이다. 골격선내의 모든 직선 선분을 조사하여 가장 길이가 긴 선분의 중요도를 1.0으로 하고 나머지 선분은 중요도가 1.0인 선분과의 상대적 길이로 표현된다. 예를 들어, (그림 3(b))의 '뺨'에서 가장 길이가 긴 선분의 번호는 'u'이다. 이 선분의 길이는 49이므로 중요도가 1.0이고 선분 'a'의 길이는 34이므로 선분 'a'의 중요도는  $34/49 = 0.69$ 이다.

예를 들어, (그림 3(b))의 문자 '뺨'에서 선분 'a'는 <6, 4, 39, 12, O, B, V, 4.60, 0.67>로 표현할 수 있다. <표 1>은 '뺨'과 '쫘'자의 모든 부품에 대한 표현을 보여준다.

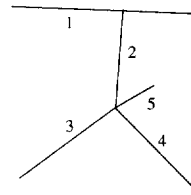
### 3.2 부품간 관계 표현

이 절에서는 부품간의 연결 관계인 에지의 표현 방법과 3.1절에서 표현된 노드를 기반으로하여 필기 한글 문자를 그래프로 구성하고 기술하는 방법에 대해서 논한다. 한글 문자는 1개 이상의 연결 요소로 구성된다. 즉, 하나의 한글 문자를 표현하기에 앞서 연결 요소를 먼저 표현해야 한다는 것이다. (그림 3)의 문자 '뺨'과 '쫘'은 각각 두 개의 연결 요소로 구성된다. 또한 세 개 이상의 연결 요소로 구성된 한글 문자도 많이 존재한다. 하나의 연결 요소는 속성 획 그래프(ASG)로 표현한다. (그림 5)는 연결 요소 'ㄸ'의 ASG와 연결 요소내의 노드와 에지가 가지고 있는 속성을 보여준다.

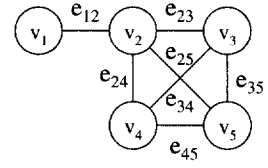
하나의 연결 요소내의 직선 선분이 노드이고 노드간

<표 1> 부품 표현

부품 번호	문자 '뺨'	문자 '쫘'
a	<6, 4, 39, 12, O, B, V, 4.60, 0.67>	<34, 4, 36, 48, O, O, H, 4.2, 1.00>
b	<46, 7, 42, 34, O, B, H, 4.66, 0.55>	<7, 8, 11, 27, O, J, H, 4.75, 0.44>
c	<39, 12, 36, 16, O, B, RD, 4.00, 0.10>	
d	<61, 12, 82, 35, O, J, LD, 7.78, 0.47>	<27, 17, 11, 27, O, J, RD, 4.35, 0.36>
e	<22, 13, 20, 23, O, J, H, 4.50, 0.20>	
f	<65, 17, 49, 59, T, J, RD, 4.49, 0.83>	
g	<7, 23, 20, 23, O, J, V, 3.62, 0.27>	<35, 30, 58, 28, T, J, V, 4.23, 0.48>
h	<39, 22, 20, 23, O, J, V, 3.15, 0.41>	<58, 28, 79, 47, O, J, LD, 5.42, 0.46>
i		<19, 34, 25, 39, O, B, LD, 6, 0.16>
j	<20, 23, 19, 37, J, J, H, 3.15, 0.27>	<58, 28, 53, 37, O, J, RD, 6.06, 0.20>
k		<75, 7, 58, 28, O, J, RD, 6.06, 0.48>
l		<11, 27, 19, 34, B, J, LD, 7.6, 0.18>
m	<42, 34, 40, 38, B, B, RD, 5.40, 0.10>	
n	<19, 37, 1, 39, O, J, V, 4.17, 0.37>	
o		
p	<19, 37, 40, 38, B, J, V, 3.19, 0.43>	
q		
r	<82, 35, 54, 64, B, J, V, 4.83, 0.60>	
s	<19, 37, 17, 63, T, J, H, 3.33, 0.51>	
t	<33, 37, 26, 62, T, T, H, 3.33, 0.49>	
u	<49, 59, 1, 66, O, J, V, 5.16, 1.00>	
v	<49, 59, 54, 64, B, J, V, 6.75, 0.08>	
w	<82, 35, 82, 65, O, J, H, 5.84, 0.63>	



(a) 연결 요소



(b) ASG

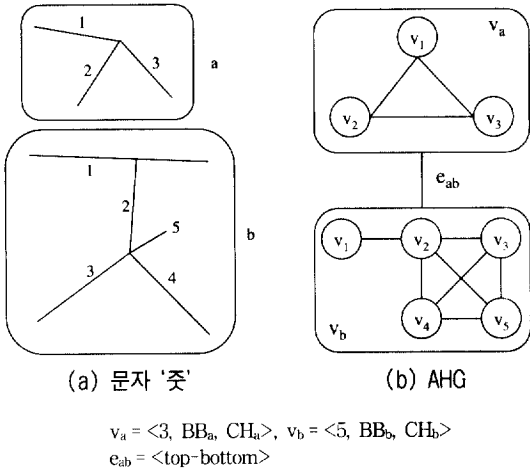
- $v_1 = \langle 34, 4, 36, 48, O, O, H, 4.20, 1.00 \rangle$
- $v_2 = \langle 35, 30, 58, 28, T, J, V, 4.23, 0.48 \rangle$
- $v_3 = \langle 75, 7, 58, 28, O, J, RD, 6.06, 0.48 \rangle$
- $v_4 = \langle 58, 28, 79, 47, O, J, LD, 5.42, 0.46 \rangle$
- $v_5 = \langle 58, 28, 53, 37, O, J, RD, 6.06, 0.20 \rangle$
- $e_{12} = \langle T \rangle, e_{23} = \langle J \rangle, e_{24} = \langle J \rangle, e_{25} = \langle J \rangle$
- $e_{34} = \langle J \rangle, e_{35} = \langle J \rangle, e_{45} = \langle J \rangle$

(c) 노드와 에지의 속성

(그림 5) 속성 획 그래프(ASG) 방법

의 연결 관계가 예지인 그래프가 ASG인 반면, AHG는 연결 요소가 노드이고 연결 요소간의 상호 위치 관계가 예지인 그래프이다. 연결 요소가 가질 수 있는 속성은 직선 선분의 개수, 최소 외접 사각형(Bounding box; BB), 볼록 껍(Convex hull; CH)이다. 즉, 3-tuple = <직선 선분의 개수, BB, CH>로 표현할 수 있다. 또한

연결 요소간의 상호 위치 관계가 가질 수 있는 속성은 상-하, 왼쪽-오른쪽의 두 가지이다. (그림 6)은 문자 'जू'의 각 연결 요소, AHG, 그리고 문자의 노드와 에지의 속성을 보여준다.



(그림 6) 한글 문자의 속성 계층 그래프 표현

AHG와 ASG의 에지를 기술하기 위해 인접 행렬(adjacency matrix) 형태의 모양 행렬(shape matrix)을 사용한다. 모양행렬은  $n \times n$  정방행렬이며, 대각선 요소를 중심으로 대칭(symmetric)이다. 모양 행렬의 비대각선 요소는 노드간의 관계인 에지의 속성을 나타낸다. 대각선 요소는 노드의 속성이므로 공백으로 비워둔다. 노드란 ASG에서는 직선 선분의 속성, AHG에서는 연결 요소의 속성을 의미한다. 에지란 ASG에서는 직선 선분간의 연결 관계 속성, AHG에서는 연결 요소간의 상호 위치 관계 속성을 의미한다. (그림 7)은 인접 행렬(I)의 각 요소를 나타낸다.

$$I = [a_{pq}]_{n \times n}$$

$$a_{pq} = \begin{cases} 0 & \text{if there is no edge between } v_p \text{ and } v_q \\ \text{blank} & \text{if } p \text{ is equal to } q \\ e_{pq} & \text{if } p \text{ is not equal to } q \end{cases}$$

$$I = \begin{pmatrix} - & e_{12} & e_{13} & e_{14} & \dots & e_{1n} \\ e_{21} & - & e_{23} & e_{24} & \dots & e_{2n} \\ e_{31} & e_{32} & - & e_{34} & \dots & e_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & e_{n3} & \dots & e_{n,n-1} & - \end{pmatrix}$$

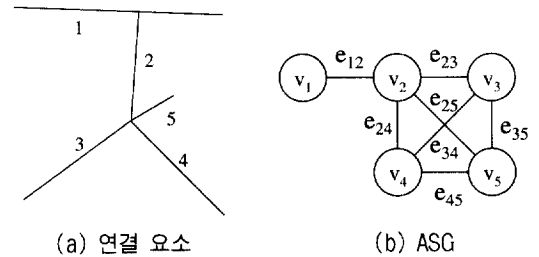
(그림 7) 모양 행렬

ASG를 위한 모양 행렬의 각 요소는 13개의 비트를 가지며 각 비트는 0, 1중의 하나이다. (그림 8)은 모양 행렬의 비대각선 요소의 각 비트의 의미를 보여준다. 왼쪽 3비트는 두 개의 직선 선분이 만나는 지점의 접촉 유형을 의미한다. 오른쪽 10비트는 두 개의 직선 선분의 접촉 유형을 의미한다. 예를 들어, (그림 5(a))에서 1번과 2번의 선분은 T-접점 영역에서 접촉하고 접촉 유형은 'T'이다. 그래서  $e_{12}(11) = e_{12}(5) = 1$ 이고 나머지 비트는 모두 0이므로  $e_{12}(=e_{21})$ 의 값은 0100000100000이다. 3번과 5번 선분의 경우, 결합 부품 영역에서 접촉하고 접촉 유형은 '-'이다. 그래서  $e_{35}(10) = e_{35}(1) = 1$ 이고 나머지 비트는 모두 0이므로 0010000000010이다. (그림 9)는 연결 요소 'ㅈ', ASG, 그리고 모양 행렬이다.

$e_{pq}(12)$	$e_{pq}(11)$	$e_{pq}(10)$	$e_{pq}(9)$	$e_{pq}(8)$	$e_{pq}(7)$	$e_{pq}(6)$	$e_{pq}(5)$	$e_{pq}(4)$	$e_{pq}(3)$	$e_{pq}(2)$	$e_{pq}(1)$	$e_{pq}(0)$
B	T	J	7	L	-	r	T	+	+	+	-	-

(b) 비대각선 요소

(그림 8) 모양 행렬의 비트와 속성 집합



$$I = \begin{pmatrix} - & 0100000100000 & 000000000000 & 000000000000 & 000000000000 \\ 0100000100000 & - & 0010010000000 & 0010100000000 & 0010100000000 \\ 0000000000000 & 0010010000000 & - & 0011000000000 & 0010000000001 \\ 0000000000000 & 0010100000000 & 0011000000000 & - & 0010001000000 \\ 0000000000000 & 0010100000000 & 0010000000001 & 0010001000000 & - \end{pmatrix}$$

(c) 모양 행렬

(그림 9) ASG의 모양 행렬

AHG의 에지도 ASG의 모양 행렬과 동일한 방법으로 구할 수 있다. AHG의 에지에 대한 모양 행렬 요소의 비트수는 2개이며 각 비트는 왼쪽부터 Left-Right, Top-Bottom의 속성을 0과 1로 표현할 수 있다.

#### 4. 결론

본 논문에서는 필기 한글 문자를 영역 기반 모양 분해 알고리즘을 사용하여 유사 불록한 부품들의 집합으

로 분해하고 이로부터 골격선을 추출하였고 이를 통해 필기 한글 문자의 모양을 표현하였다. 영역 기반 모양 분해 알고리즘을 사용한 이유는 무제약 필기 한글 문자 인식 과정에서 전처리로 사용되는 세선화 과정이 초래하는 문자의 왜곡현상을 줄여서 인식률의 향상을 기대하기 위함이다.

필기 한글 표현 방법으로는 속성 획 그래프(Attributed stroke graph ; ASG)와 속성 계층 그래프(Attributed hierarchical graph ; AHG)를 사용하였다. ASG는 연결 요소내의 직선 선분이 노드이고 직선 선분간의 연결 관계가 에지인 그래프이고 AHG는 연결 요소가 노드이고 연결 요소간의 위치 관계가 에지인 그래프이다. ASG를 기술하기 위한 자료구조는 노드와 에지를 구분하여 저장한다. 노드는 9-tuple의 구조체에 저장하였고 에지는 인접 행렬의 형태인 모양 행렬을 이용하여 저장하였다. 모양 행렬은  $n \times n$  정방행렬이며 대각선 요소를 중심으로 대칭이다. 모양 행렬의 요소는 ASG는 직선 선분간의 연결 관계 속성, AHG는 연결 요소간의 위치 관계 속성을 표현하였다.

### 참 고 문 헌

[1] H. S. Park, and S. W. Lee, "Off-line recognition of large-set handwritten characters with multiple hidden Markov models," *Pattern Recognition*, Vol.29, No.2, pp.231-244, 1996.

[2] 김호연, 서장원, 권재욱, 김진형, "필기 한글 인식에 관한 연구 : KAIST의 연구를 중심으로", *인지과학회 논문지*, Vol.6, No.4, pp.5-20, 1995.

[3] H. J. Kim, and P. K. Kim, "Recognition of off-line handwritten Korean characters," *Pattern Recognition*, Vol.29, No.2, pp.245-254, 1996.

[4] 황순자, 김문현, "접촉유형에 기반한 오프라인 필기체 한글문자분할", *인지과학회 논문지*, Vol.6, No.4, pp.51-71, 1995.

[5] 이은주, 권오석, 김태균, "필기체 한글에서 자모 분리 및 인식", *한국정보과학회 논문지*, Vol.15, No.6, pp.526-534, 1988.

[6] 박정선, 오일석, "필기 한글 패턴의 모양 분해", *한국정보과학회 논문지*, 1999년 12월(submitted).

[7] 홍기천, 오일석, "모양 분해를 이용한 필기 한글 문자의 골격선 추출", *한국정보과학회 논문지*, 2000년 1월(submitted).

[8] S. W. Lu, Y. Ren. and C. Y. Suen, "Hierarchical attributed graph representation and recognition of handwritten Chinese characters," *Pattern Recognition*, Vol.24, No.7, pp.617-632, 1991.



### 홍 기 천

e-mail : kchong@cs.chonbuk.ac.kr

1995년 전북대학교 전산통계학과 (학사)

1997년 전북대학교 전산통계학과 (석사)

1997년~현재 전북대학교 전산통계학과 박사과정

관심분야 : 패턴 인식, 컴퓨터 비전, 그래픽스



### 오 일 석

e-mail : isoh@moak.chonbuk.ac.kr

1984년 서울대학교 컴퓨터공학과 (학사)

1986년 한국과학기술원 전산학과 (석사)

1992년 한국과학기술원 전산학과 (박사)

1992년~현재 전북대학교 컴퓨터과학과 부교수

관심분야 : 컴퓨터 비전, 문서 및 문자 인식, 그래픽스, GIS