GAP 군집화에 기반한 필기 한글 단어 분리

(Word Segmentation in Handwritten Korean Text Lines based on GAP Clustering)

정 선 화 † 김 수 형 ††

(Seon Hwa Jeong) (Soo Hyung Kim)

요 약 본 논문에서는 필기 한글 문자열 영상에 대한 단어 분리 방법을 제안한다. 제안된 방법은 gap의 크기 정보를 사용하여 단어를 분리하는데, 이때 gap은 문자열 영상을 수직방향으로 투영한 후 흰-런(white-run)을 찾음으로써 구할 수 있다. 문자열 영상으로부터 얻어지는 gap들의 크기를 측정한 후, 각각의 gap을 단어와 단어사이에 존재하는 gap과 문자와 문자사이에 존재하는 gap 중 하나로 분류한다. 본 논문에서는 필기 영문 문자열의 단어 분리를 위해 제안된 기존의 세 가지 거리 척도를 채택하고 군집화에 기반한 세 가지 분류방법을 적용하여 한글 문자열의 단어 분리를 위한 최적의 조합을 선정하였다. 우편봉투 상에 작성된 주소열로부터 수작업으로 추출한 305 개의 문자열 영상을 사용하여 실험한 결과 BB(bounding box) 거리를 사용하여 순차적 군집 방법을 적용하는 경우 3 순위까지의 누적 단어 분리 성공률이 88.52% 로서 가장 우수한 성능을 보여 주었다. 또한 하나의 문자열 영상에 대한 단어 분리 속도는 약 0.05초이다.

Abstract In this paper, a word segmentation method for handwritten Korean text line images is proposed. The method uses gap information to segment words in line images, where the gap is defined as a white run obtained after vertical projection of line images. Each gap is assigned to one of inter-word gap and inter-character gap based on gap distance. We take up three distance measures which have been proposed for the word segmentation of handwritten English text line images. Then we test three clustering techniques to detect the best combination of gap metrics and classification techniques for Korean text line images. The experiment has been done with 305 text line images extracted manually from live mail pieces. The experimental result demonstrates the superiority of BB(Bounding Box) distance measure and sequential clustering approach, in which the cumulative word segmentation accuracy up to the third hypothesis is 88.52%. Given a line image, the processing time is about 0.05 second.

1. 서 론

정보화 사회에 들어서면서 사회 도처에 존재하는 자료를 데이타베이스화하여 정보로 활용하는 것이 매우 중요하게 되었다. 기존에 존재하는 자료를 사람이 직접 컴퓨터에 입력하기에는 시간적 금전적 비용이 매우 크다. 따라서 사람이 하는 작업을 줄이기 위하여 기계가 일부 대신하여 수행할 수 있는데 이때 문서인식기술이 요구된다. 이러한 문서인식기술은 우편봉투 주소열 인식[1], 전표 인식[2], 팩스 배달[3] 등의 여러 응용분야를 갖는다. 이들 문서는 하나 이상의 문자열로 구성되어 있고 각각의 문자열은 단어들로 구성되어 있다. 따라서 문서 인식을 수행하기 위해서 처리 대상이 되는 문서는 문자열 영상으로 분리되어야 하고 각각의 문자열 영상은 단어 단위로의 분리가 필요하다. 본 논문에서는 필기 한글 문자열 영상에 대한 단어 분리 방법을 제안한다.

단어 분리 방법에 관한 연구는 주로 영문 문자열에 국한되어 진행되어 왔으며[4, 5, 6, 7, 8], 필기 한글에

· 본 논문은 1999년도 한국전자통신연구소 국내 위탁 연구과제의 지원에 의하여 연구되었음

† 비 회 원 : 전남대학교 전산학과
swjong@chonnam.chonnam.ac.kr
†† 종신회원 : 전남대학교 정보통신연구소 교수
shkim@chonnam.chonnam.ac.kr

논문접수 : 1999년 10월 16일
심사완료 : 2000년 4월 14일

대한 단어 분리 연구는 아직까지 이루어지지 않고 있는 실정이다. 단어 분리에 관한 기존의 연구는 사용되는 특징에 따라 크게 두 가지로 구분할 수 있다. 인접한 화소들의 집합을 연결요소(connected component)로 정의할 때, 첫 번째 방법에서는 연결 요소 사이에 존재하는 gap의 크기 정보만을 사용하여 단어 분리를 수행한다[5, 9, 10, 11, 12]. 두 번째 방법은 사람의 단어 분리 방법을 모방한다[4, 6, 13, 14]. 사람은 단어 분리를 하기 위해 gap의 크기 정보뿐 아니라, 쉼표나 마침표의 존재, 영문의 경우 단어의 첫 문자가 대문자인지 소문자인지의 여부, 숫자의 존재에 관한 정보 등을 사용한다. 즉 두 번째 방법에서는 단어 분리를 위해 gap의 크기 정보 외에 위에서 언급한 다른 정보들을 추가로 사용한다. 그러나 인식과정을 거치지 않고 두 번째 방법에서 사용되는 정보들을 얻기는 매우 어렵다.

본 논문에서는 gap 크기만을 사용하여 단어 분리를 수행한다. 이때 단어 분리 문제는 gap 크기 측정 단계와 분류 단계로 나누어 생각할 수 있다. Gap으로 분리되는 검은 화소의 집합을 단어후보라고 정의할 때, gap의 크기 측정 단계에서는 인접한 두 단어후보 사이의 거리를 측정한다. 영문 문자열 영상에서의 gap 거리 척도에 관한 연구가 많이 이루어져 있다. [4]에서는 다수의 거리 척도를 비교/평가하고 있으며, [5]에서는 [4]에서 우수하다고 평가된 일부 방법들과 제안된 방법의 성능을 비교하고 있다. 본 논문에서는 기존에 제안된 방법들 중 한글에 적합하다고 판단되는 BB(Bounding Box), RLE(Run-Length/Euclidean), CH(Convex Hull) 거리 척도를 선택하여 사용하였다. 그러나 영문의 경우 인접한 두 연결요소 사이의 거리를 측정하는데 이들 거리 척도들을 사용하는 반면, 여기서는 인접한 두 단어후보 사이의 거리를 측정하는데 사용한다. BB 거리는 단어후보의 bounding box 사이의 수평 거리를 나타내며, RLE 거리는 휴리스틱 규칙을 기준으로 단어후보 사이의 최소 런-길이나 유클리디안 거리를 사용한다. CH 거리는 단어후보들의 convex hull을 구한 후 인접한 두 convex hull의 중심을 잇는 선분과 만나는 convex hull 상의 두 점 사이의 유클리디안 거리로 정의된다.

분류 단계에서는 임의의 gap을 단어와 단어사이에 존재하는 gap과 문자와 문자사이에 존재하는 gap중 하나로 분류한다. 본 논문에서는 군집화에 기반한 세 가지 방법을 고려하였다. 첫 번째 방법은 기존의 계층적 군집화 방법 중 하나인 최소 평균 거리법이다. 두 번째 방법은 군집을 두 개로 고정시킨 후 gap을 하나씩 순차적으로 분류해 가는 방법이고, 세 번째 방법은 크기 순으로 정렬된 gap에서 이웃하는 gap들간의 거리와 크기 비율을 사용하여 분류하는 방법이다.

문자열 영상에서 분리는 인식을 위해서 수행되는 단계이다. 단어 분리 과정에서는 높은 분리 신뢰도를 갖는 단 하나의 단어 분리 결과를 생성할 수도 있고, 여러 분리 후보를 인식 과정으로 넘겨줌으로써 인식과 함께 단어 분리 결과를 정하도록 할 수도 있다. 전자의 방법은 처리 속도를 향상시키지만 잘못된 단어 분리 결과를 수정할 수 없다는 단점이 있다. 후자의 방법[7, 15, 16, 17, 18]은 여러 단어 분리 결과를 수용하므로 단어 인식률은 좋지만 넘겨온 분리 결과만큼 여러 번의 인식을 수행함으로 처리 속도의 부담이 크다. 또한 이때 단어 인식 단계에서 사용되는 인식 기법은 단어후보들에 대하여 타당한 인식점수(예 :사후확률)를 출력할 수 있어야 한다. 여러 개의 단어 분리 후보를 내주는 방법에 관한 기존 연구로는 최소 평균 거리법을 사용하여 군집화를 수행한 후 분리 후보를 결정하는 방법[15]과 단순히 주어진 gap들을 조합하여 분리 후보를 내주는 방법[16, 17, 18] 등이 있다. 본 논문에서도 여러 개의 단어후보를 내주는 방법을 고려한다.

본 논문의 목표는 필기 영문 문자열의 단어 분리를 위해 제안된 세 가지 거리 척도에 대하여 군집화에 기반한 세 가지 분류방법을 적용하여 한글 문자열의 단어 분리를 위한 최적의 조합을 찾는 것이다. 본 논문의 구성은 다음과 같다. 2절에서 세 가지 거리 척도를 기술하며, 분류방법은 3절에서 기술한다. 4절에서 실험 결과를 제시하며, 5절에서 결론을 맺는다.

2. 거리 척도

필기 한글 문자열 영상에 대한 단어 분리를 위하여 gap 크기 정보를 사용한다. Gap 크기만을 사용하여 단어 분리를 수행하기 위해 주어진 문자열 영상에 두 가지 가정을 한다. 첫 번째, 단어와 단어사이에는 띄어쓰기가 되어있다고 가정한다. 두 번째, 문자와 문자 사이에 존재하는 gap(ICG: Inter-Character Gap)의 크기보다 단어와 단어 사이에 존재하는 gap(IWG: Inter-Word Gap)의 크기가 더 크다고 가정한다. 본 논문에서는 첫 번째 가정 하에 문자열 영상을 수직방향으로 투영한 후 0 의 값을 갖는 부분, 즉 흰-런(white-run)을 찾아 gap이라 한다. 또한 수직방향 투영방법으로 gap을 찾기 위해서는 문자열 영상의 기울어짐이 적어야 한다. 그림 1에는 임의의 문자열 영상을 수직방향으로 투영하여 구한 일곱 개의 gap을 보여주고 있다.

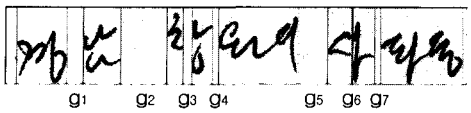


그림 1 문자열 영상의 수직방향 투영 결과

2.1 BB 거리

BB(bounding box) 거리는 인접한 두 단어후보의 bounding box의 수평거리로 정의된다. 삐침획(ligature)을 많이 포함하는 영문의 경우 BB 거리는 그다지 적합한 거리 척도가 아니다[4]. 그러나 문자의 전체적인 크기가 비교적 비슷하면서 직사각형 모양으로 작성되는 한글의 경우에 적당한 거리 척도로 사용될 수 있다. 이는 4절에 기술된 실험결과에서 보여준다. 그림 2는 그림 1에 주어진 문자열 영상에 대하여 BB 거리를 측정한 예를 보여주고 있다.

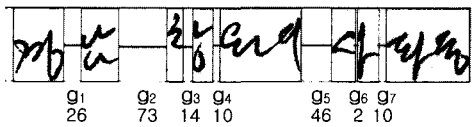


그림 2 BB 거리 측정의 예 (단위: 화소)

2.2 RLE 거리

RLE(run-length/Euclidean) 거리는 인접한 두 단어후보사이의 수평방향으로의 겹침 정보를 고려한다. 겹침의 정도는 아래와 같이 계산된다. 여기서 O는 두 단어후보를 수평방향으로 투영하였을 때 겹치는 크기를 나타내고, H_1과 H_2 각각은 두 단어후보의 세로방향 크기를 나타낸다.

$$\frac{O}{H_1+H_2}$$

두 단어후보사이의 거리를 구하기 위해서 RLE 거리 척도에서는 겹침의 정도가 임계값보다 크면 최소 런-길이를 사용하고 그렇지 않으면 유클리디안 거리를 사용한다. 최소 런-길이는 두 단어후보가 겹치는 구간에서의 런-길이 중 최소의 거리이다. 본 논문에서는 임계값으로 0.25를 사용하였다. 그림 3은 주어진 문자열 영상에서 최소 런-길이와 유클리디안 거리를 계산한 예를 보여주고 있다.

한글은 10개의 모음과 14개의 자음으로 구성되며, 그 구성 형태는 여섯 가지로 분류될 수 있다[19]. 이중 다섯 가지 구성형태가 받침을 가지며, 한글이 빠르게 필기되는 경우 받침은 종종 나머지 글자를 구성하는 원소들

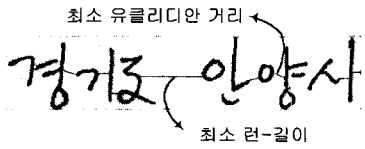


그림 3 최소 런-길이와 최소 유클리디안 거리의 예

과 떨어져서 작성된다. RLE 거리는 이 경우를 고려한다. 즉 받침이 하나의 단어후보이고 인접한 단어후보와 거리를 구할 때 겹침의 정도가 작은 경우 최소 유클리디안 거리를 사용함으로써 최소 런-길이보다 더 작은 거리를 갖도록 한다. 그림 4는 그림 1에 주어진 문자열 영상에 대하여 RLE 거리 척도를 적용한 예를 보여주고 있다. 이 문자열의 경우 단어후보들 사이의 겹침이 임계치보다 커서 모든 거리가 최소 런-길이를 사용하여 계산되어졌다. 또한 아래 그림에서 볼 수 있는 것처럼 RLE 거리는 BB 거리 보다 항상 큰 값을 갖는다.

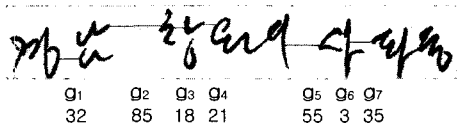


그림 4 RLE 거리 측정의 예 (단위: 화소)

2.3 CH 거리

CH 거리는 단어후보 사이의 거리를 계산할 때 그들의 모양을 반영한다. 먼저 단어후보의 convex hull를 구하고[20] 이를 구성하는 좌표들 $(x_1,y_1),\cdots,(x_n,y_n)$로부터 convex hull의 중심을 아래와 같이 좌표의 평균값으로 계산한다.

$$\left(\frac{1}{n}\sum_{i=1}^{n}x_i,\ \frac{1}{n}\sum_{i=1}^{n}y_i\right)$$

두 인접한 단어후보 사이의 CH 거리는 각 convex hull의 중심을 잇는 선분과 convex hull 이 만나는 점 P_1과 P_2의 유클리디안 거리로 계산된다. 그림 5에는 그림1에 주어진 문자열 영상에 대하여 CH 거리를 측정한 예를 보여주고 있다.

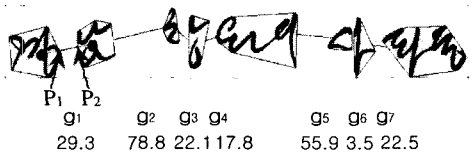


그림 5 CH 거리 측정의 예

3. 분류방법

문자열 영상을 수직방향으로 투영한 후 얻어진 gap들의 집합을 $\{g_1,\cdots,g_n\}$ 이라 하자. 그리고 이들이 분류될 클래스의 집합을 {ICG,IWG} 이라 정의하자. 여기서 ICG는 문자와 문자사이에 존재하는 gap을 의미하며, IWG는 단어와 단어사이에 존재하는 gap을 나타낸다. 임의의 문자열 영상을 단어 단위로 분리하는 문제는 정의역 $\{g_1,\cdots,g_n\}$에서 공역 {ICG,IWG}로 대응되는 함수 f를 구하는 것과 동일하다. 이때 함수 f를 세 가지 형태로 나눌 수 있다. 첫 번째는 모든 정의역의 원소가 공역의 ICG로 대응되는 경우로 이는 주어진 문자열 영상이 단 하나의 단어로 구성되어 있음을 의미한다. 두 번째는 모든 정의역의 원소가 IWG로 대응되는 경우이며 이는 주어진 문자열 영상이 n+1개의 단어로 구성됨을 의미한다. 세 번째는 ICG와 IWG로 골고루 대응되는 경우이며 이때 문자열 영상을 구성하는 단어의 개수는 IWG로 대응되는 정의역의 원소 개수 + 1이 된다.

함수 f의 첫 번째와 두 번째의 경우는 얻어진 gap들이 하나의 부류로 대응되는 경우로, 본 논문에서는 이들을 위해서 별도의 처리를 한다. 먼저 집합 $\{g_1,\cdots,g_n\}$의 분산을 구한 후 주어진 임계치보다 작으면 모든 gap들을 ICG와 IWG 중 하나의 부류로 판단한다. 분산이 작다는 의미는 g_i들이 하나의 대표값(평균)을 중심으로 서로 근접하게 분포하고 있음을 나타낸다. 그 다음 대표값을 기준으로 이 값이 작으면 모든 gap들을 ICG로, 그렇지 않으면 IWG로 분류한다. 분산이 주어진 임계치보다 크면 세 번째의 경우로 판단하고 아래의 절들에서 설명하고 있는 세 가지 분류방법 중 하나를 사용하여 주어진 gap을 두 부류 중 하나의 부류로 할당한다.

3.1 최소 평균 거리법

최소 평균 거리법(ALM : Average Linkage Method)은 계층적 군집화 방법 중 하나로 단어분리를 위하여 [15]에서 처음 도입되었다. 우선 각각의 gap을 원소로 갖는 n개의 군집 $C_1,\cdots,C_n$을 고려한다. 다음 n개의 군집 중 가장 가까운 군집 두 개를 선택하여 하나의 군집으로 묶는다. 이때 가까움의 정도는 다음의 식을 이용하여 측정한다.

$$D(C_i,C_j) = \frac{1}{n_i, n_j} \sum_{a\in C_i, b\in C_j} d(a,b)$$

여기서 n_i는 군집 C_i에 속해있는 gap의 개수이며, d는 거리함수로 본 논문에서는 맨하탄(Manhattan)거리를 사용한다. 이와 같이 가까움 정도를 측정하여 묶는 과정을 두 개의 군집이 남을 때까지 반복한다. 그림 6의 수형도(dendrogram)에서는 n개의 gap이 두 개의 군집으로 묶여 가는 순서를 알파벳을 사용하여 보여주고 있다. 이때 사용된 gap 크기는 그림 2에 주어진 BB 거리이다.

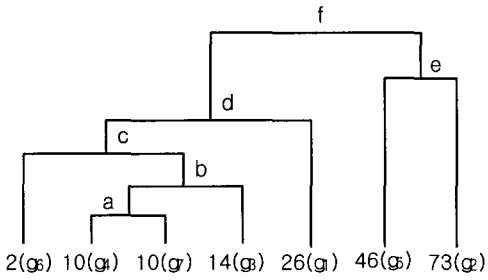


그림 6 수형도 - 그림 2의 BB 거리들을 최소 평균 거리법으로 군집화한 결과

그림 6의 수형도는 오른쪽 군집이 왼쪽 군집보다 더 큰 평균을 가지도록 정렬하고 군집 내에서는 올림차순 정렬함으로써 얻어진다. 여기서 1 순위 단어 후보는 가장 마지막의 군집화 결과(f)로서 g_5와 g_2 두 개가 IWG에 해당됨을 알 수 있으며, 그 결과 입력 문자열이 세 개의 단어로 분리된다. 이때 i 순위 단어 분리 후보집합은 n-1-i 번째 군집 결과 중 오른쪽에 위치하는 군집들의 집합이다. 그림 7은 단어 분리 후보집합 및 해당되는 단어 분리 결과를 보여준다.

순위	IWG 후보	
1(f)	g_5, g_2	깨끗한물에 수박을
2(e)	g_2	깨끗한물에 수박을
3(d)	g_1, g_5, g_2	깨끗한물에 수박을
⋮	⋮	

그림 7 최소 평균 거리법에 의한 단어 분리 후보 - 그림 2의 문자열영상, BB거리

3.2 순차적 군집 방법

순차적 군집 방법에서는 먼저 $\{g_1,\cdots,g_n\}$을 올림차순으로 정렬한다. 그 결과를 g(1),…,g(n) 이라 하자. 그 다음 두 개의 군집 C_{ICG}와 C_{IWG}을 다음과 같이 초기화한다.

$$C_{ICG} = \{g(0)\},\quad C_{IWG} = \{g(n)\}$$

여기서 g(0)은 널(null) gap으로 ICG 부류 중 가장 작

은 거리 값을 갖는 gap을 표시하기 위하여 사용된다. 즉 문자와 문자사이에 존재하는 gap의 가장 작은 값으로 '0'을 고려하였다. 초기화 후 분류되지 않은 나머지 n−1개의 gap들을 C_{IWG}와 C_{ICG} 중 하나의 부류로 할당한다. g(i)에 대한 분류 방법은 다음과 같다. g(i)와 C_{ICG}와 C_{IWG} 각각의 평균과의 거리를 계산한 후 가까운 부류로 g(i)를 할당한다. 이때 참조하는 gap의 순서를 작은 값, 큰 값 순서로 정함으로써 주어진 g(i)가 계속해서 C_{ICG}으로 분류된다거나 또는 계속해서 C_{IWG}으로 분류되는 것을 방지하였다. 이는 한 부류에 대한 정보만을 계속 추가함으로써 생기는 정보의 편의(bias)를 줄이기 위함이다. 순차적 분류 방법의 종료 시점은 $i \le (n-1)/2$ 일 때 g(i)가 C_{IWG}으로 분류될 때와 $i > (n-1)/2$ 일 때 g(i)가 C_{ICG}으로 분류되는 때이다.

순차적 군집 방법을 사용하여 IWG를 찾을 때, 1 순위 단어 분리 후보는 두 개의 군집 중 C_{IWG}에 속한 gap들로 구성된다. 크기 순으로 gap들을 분류하기 때문에 C_{IWG}에 마지막으로 분류된 gap이 C_{IWG}에 속하는 gap 중 가장 작은 값을 갖는다. 1 순위 단어 분리 후보에 속하는 gap 중 가장 작은 gap을 g(j)라고 하자. 이때 i 순위 단어 분리 후보 집합은 다음과 같다.

$$\{g(j+i-1),\cdots,g(n)\},\ \text{if } j+i-1 \le n$$
$$\{g(j-K),\cdots,g(n)\},\ \text{if } j+i-1 = n+K,\ K>0$$

그림 8은 그림 2에 순차적 군집 방법을 적용한 예이다. 이때 1 순위 단어후보집합은 { g(6)($=g_5$), g(7)($=g_2$) }이다. 여기서 j는 6이 되고 n은 7이 된다. 2 순위 단어후보집합은 $j+i-1 \le n$($6+2-1 \le 7$)이므로 위의 첫 번째 식으로부터 {g(7)}이 된다. 마지막으로 3 순위 단어후보집합은 $j+i-1 > n$($6+3-1 = 7+1$)이므로 위의 두 번째 식으로부터 { g(5), g(6), g(7) }이 된다. 이때 K=1이다.

C_{ICG} C_{IWG}
2(g_6) 10(g_4) 10(g_7) 14(g_3) 26(g_1) | 46(g_5) 73(g_2)

(a) 1 순위 단어 분리 후보

C_{ICG} C_{IWG}
2(g_6) 10(g_4) 10(g_7) 14(g_3) 26(g_1) 46(g_5) | 73(g_2)

(b) 2 순위 단어 분리 후보

C_{ICG} C_{IWG}
2(g_6) 10(g_4) 10(g_7) 14(g_3) | 26(g_1) 46(g_5) 73(g_2)

(c) 3 순위 단어 분리 후보

그림 8 순차적 군집 방법 적용 예 - 그림 2의 문자열영상, BB거리

3.3 거리/크기비율에 기반을 둔 분류 방법

거리/크기비율에 기반을 둔 분류 방법에서는 gap을 올림차순으로 정렬한 후 인접한 두 gap 사이의 거리와 크기비율 정보를 사용하여 주어진 gap을 IWG과 ICG 중 하나로 할당한다. 인접한 g(i)와 g(i+1) 사이의 거리 d_i ($i = 1,\cdots,n-1$)는 유클리디안 거리 함수를 사용하여 계산되며, 비율 r_i는

$$\frac{|g(i)|}{|g(i+1)|}$$

로 계산된다. 여기서 g(i)가 벡터 값을 가질 때 |·|는 벡터의 크기(norm)를 의미한다. 주어진 gap들을 분류하기 위해서 거리와 크기비율을 곱한 값이 사용된다. 분류 기준은 다음과 같다.

$$j = \underset{1 \le i \le n-1}{\text{argmax}} (d_i * r_i)$$

일 때, g(j+1),…,g(n)을 C_{IWG}으로 분류한다. 이 분류 방법에서 i 순위 단어후보집합은 i 번째로 큰 $d_i * r_i$를 찾음으로써 얻어진다. 그림 9에서는 그림 2에 거리/크기비율에 기반을 둔 분류 방법을 적용하여 단어 분리 후보집합을 구하는 과정을 보여주고 있다.

2(g_6) 10(g_4) 10(g_7) 14(g_3) 26(g_1) 46(g_5) 73(g_2)

d	8	0	4	12	20	27
r	5	1	1.4	1.86	1.77	1.56
d*r	40	0	5.6	22.32	35.4	42.12

순위	IWG gap
1	g_2
2	g_5, g_2
3	g_1, g_5, g_2
⋮	⋮

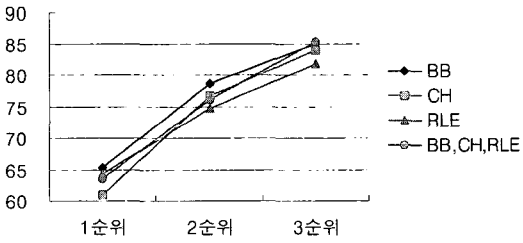
그림 9 거리/크기비율에 기반을 둔 분류 방법의 적용 예 - 그림 2의 문자열영상, BB거리

4. 실험 및 결과

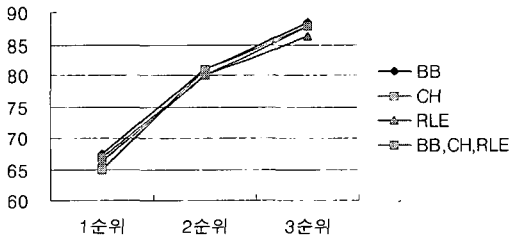
본 논문에서 사용한 거리 척도와 분류방법을 평가하기 위한 실험에서 사용된 총 305개의 문자열 영상 데이타는 우편봉투 상에 작성된 주소열로부터 수작업으로 추출되었다. TWAIN 호환 스캐너에 의해 해상도 300dpi 트루 칼라로 스캐닝된 우편봉투 영상들은 한국과학기술원 전산학과 인공지능 연구실로부터 제공받은 데이타이며 본 실험에서는 문자열 영상을 이진화하여 사용하였다. 실험은 Pentium II 366MHz PC 상에서 수행되었다.

4.1 거리척도/분류방법의 성능 비교

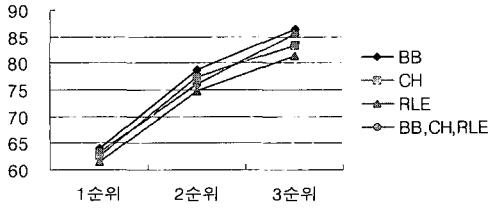
본 연구에서는 BB, RLE, CH 거리 각각을 세 가지 분류방법에 적용하여 보고 각각의 성능을 평가/비교하여 보았으며, 또한 세 가지 거리를 모두 고려하였을 때의 성능 변화도 관측하였다. 그림 10에는 각 방법의 성능이 제시되어 있다. 이때 정확도는 문자열 단위로 계산되었다. 그림 10으로부터 거리 척도 중 BB 거리가 세 분류방법 모두에서 가장 우수한 성능을 보여 주며 분류방법 중에서는 순차적 군집 방법이 가장 우수함을 관측할 수 있다. 즉 BB 거리를 사용하는 순차적 군집 방법이 가장 우수한 단어 분리 성공률을 주었다. 이때 1 순위 후보만을 고려할 경우 성공률은 67.54%이며, 2 순위 후보까지 고려할 경우 80.98% 그리고 3 순위 후보까지 고려할 경우 88.52%이었다. 또한 세 가지 거리를 모두 고려한 경우에 그 성능은 BB, RLE, CH 각각의 거리를 적용하여 얻어진 성능의 평균에 가깝다는 사실 또한 관측할 수 있었다. 이는 각 거리 척도들이 서로의 장단점을 잘 보완하지 못함을 시사한다.



(a) 최소 평균 거리법



(b) 순차적 군집 방법



(c) 거리/크기비율에 기반을 둔 분류 방법

그림 10 단어 분리에 대한 누적 성공률

[15]에서 Mahadevan 등은 위의 세 가지 거리 척도를 사용하여 영문 문자열에 대한 단어 분리 실험을 하였다. 이때 사용된 분류 방법은 Max 방법이다. Max 방법은 거리 크기/비율에 기반을 둔 방법에서 크기만을 고려한 방법이다. 영문 문자열에 대한 실험 결과 CH거리가 가장 좋은 결과를 보여주며, 그 다음 RLE거리, BB거리 순으로 주어졌다. 한글 문자열 영상에 대하여 실험한 본 논문에서는 BB거리, RLE거리, CH거리 순으로 성능이 주어졌다. 영문의 경우 알파벳들의 크기들이 서로 다르고 또한 삐침획들이 많이 포함되어 있어 글자의 모양을 반영하는 CH거리가 가장 우수한 성능을 주는 반면, 한글의 경우 문자들이 거의 동일한 크기의 직사각형 모양을 가지는 특성으로 인해 단어후보사이의 bounding box의 수평거리를 사용하는 BB거리가 가장 우수한 성능을 보여준다.

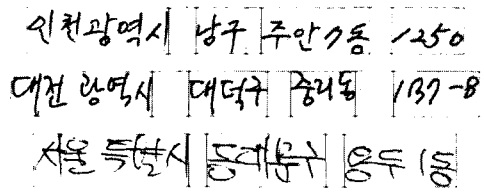
군집화 방법은 최종 군집 개수에 대한 정보를 사용하는 경우와 사용하지 않는 경우로 나누어질 수 있다. 본 논문에서 주어진 문제는 gap들을 두 클래스 - IWG, ICG -로 분류하는 문제로 최종 군집 개수에 대한 정보가 주어져 있다. 고려한 분류방법 중 최소 평균 거리법은 계층적 군집화 기법 중 하나로 가장 가까운 군집들을 하나의 군집으로 만들어질 때까지 묶어가는 방법이다. 즉 최종 군집 개수에 대한 정보를 사용하지 않고 있다. 그러나 순차적 군집 방법은 이 정보를 사용하여 군집화를 수행하고 있다. 이렇게 부가적인 정보를 사용함으로써 순차적 군집 방법이 더 좋은 성능을 보여주는 것으로 분석된다. 마지막으로 하나의 문자열 영상이 주어질 때 단어분리를 수행하는데 약 0.05초가 소요되었다.

4.2 단어 분리 실패 분석

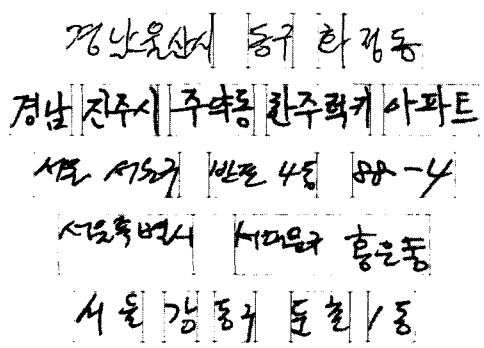
본 논문에서는 문자열 영상을 단어 단위로 분리하기 위하여 다음의 두 가정을 하였다. 즉 문자열 영상은 단

어 단위로 띄어쓰기가 되어있고, 문자와 문자사이의 gap 크기보다 단어와 단어사이에 존재하는 gap 크기가 더 크다라고 가정하였다. 그러나 본 논문에서 사용된 총 305개의 실험 데이타를 분석한 결과 전자의 가정에 위배된 영상이 2개 존재하였고, 후자의 가정을 위배한 영상이 22개 존재하였다. 이 수치로부터 gap 정보만을 사용하여 단어 분리를 할 경우 얻을 수 있는 최대 성능은 92.13%(281/305)임을 알 수 있다. 본 실험에서 얻어진 가장 높은 단어 분리 성공률은 BB 거리를 사용하는 순차적 군집 방법에 의해서 얻어진 3 순위 누적 성공률로서 88.52%(270/305)이었다. 이때 가정에 위배되는 데이타를 기각시키고 단어 분리 성공률을 계산하면 96.09%(270/281)가 된다.

그림 11은 단어 분리를 수행한 몇 개의 문자열 영상을 보여 주고 있다. (a)는 성공 사례를, (b)는 실패 사례를 보여주고 있다. 처음 세 개의 문자열 영상은 가정에 위배되어 단어 분리에 실패한 경우이다. (b)의 첫 번째 문자열 영상은 단어와 단어 사이가 띄어쓰기가 되어 있지 않은 경우이며, 그 다음 두 개의 문자열 영상은 IWG의 크기보다 ICG의 크기가 더 큰 경우를 나타내고 있다. 네 번째 문자열 영상은 가장 대표적인 실패 사례로 매우 큰 IWG로 인해 나머지 IWG들이 ICG로 잘못 분류되고



(a) 단어 분리에 성공한 사례



(b) 단어 분리의 실패 사례

그림 11 BB 거리를 사용하는 순차적 군집 방법을 적용한 단어 분리 결과

있다. 다섯 번째 문자열 영상에서는 IWG들과 ICG들이 비슷한 크기를 가짐으로써 단어 분리에 실패하였다.

5. 결론 및 향후연구

본 논문에서는 필기 영문 문자열의 단어 분리를 위해 제안된 기존의 세 가지 거리 척도에 대하여 군집화에 기반한 세 가지 분류방법을 적용하여 한글 문자열의 단어 분리를 위한 최적의 조합을 선정하였다. 세 가지 거리 척도로 BB, RLE, CH 거리를 고려하였고, 분류방법 최소 평균 거리법, 순차적 군집 방법, 거리/크기비율에 기반을 둔 분류방법을 사용하였다. 이들의 총 9가지 조합과, 세 가지 거리를 모두 고려한 경우까지 포함하여 총 12 가지 조합에 대하여 단어 분리 성공률을 조사하였다. 이때 사용된 문자열 영상 데이타는 실제 우편봉투 상에서 수작업으로 추출된 305개의 주소 데이타이다. 그 결과 BB 거리를 사용하는 순차적 군집 방법이 가장 우수한 성능을 내주었으며, 1순위부터 3순위까지의 누적 단어 분리 성공률이 각각 67.54%, 80.98%, 88.52%이었다. 또한 하나의 문자열 영상에 대한 단어 단위 분리 수행하는데 약 0.05초가 소요되었다.

본 논문에서는 주어진 gap을 두 부류로 분류하기 위한 특징으로 gap 크기 정보만을 사용하고, 이 때문에 문자열 영상에 대하여 두 가지 가정을 하였다. 그러나 실험 데이타에서도 관측할 수 있듯이 실제 문자열 영상 데이타들에서 그러한 가정들을 완전히 만족하지 못하는 경우가 종종 발생한다. 향후 연구 과제로 문자열 영상에 대하여 어떠한 가정도 하지 않고 단어 분리를 수행하기 위해서 gap 크기이외에 한글의 특성을 잘 반영하는 특징들을 구상하고자 한다.

참 고 문 헌

[1] S.N. Srihari and E.J. Keubert, "Integration of hand-written address interpretation technology into the United States Postal Service remote computer reader system," *Proc. 4th International Conference on Document Analysis and Recognition*, pp. 892-896, Ulm, Germany, Aug. 1997.

[2] S.N. Srihari, Y.C. Shin, V. Ramanaprasad and D.S. Lee, "A system to read names and addresses on tax forms," *Technical Report CEDAR-TR-94-2*, CEDAR, SUNY Buffalo, Oct. 1994.

[3] A.J. Elms, S. Procter and J. Illingworth, "The advantage of using HMM-based approach for faxed word recognition," *International Journal of Document Analysis and Recognition*, Vol. 1, No. 1, pp. 18-36, 1998.

[4] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines," *Pattern Recognition*, Vol. 27, No. 1, pp. 41-52, 1994.
[5] U. Mahadevan and R.C. Nagabushnam, "Gap metrics for word separation in handwritten lines," *Proc. Third International Conference on Document Analysis and Recognition*, pp. 124-127, Montreal, Canada, 1995.
[6] G. Kim, "Architecture for handwritten text recognition systems," *Proc. Sixth International Workshop on Frontiers in Handwritten Recognition*, pp. 113-122, Taejon, Korea, August 1998.
[7] G. Dzuba, A. Filatov and A. Volgunin, "Handwritten ZIP code recognition," *Proc. Fourth International Conference on Document Analysis and Recognition*, pp. 766-770, Ulm-Germany, August 1997.
[8] A.C. Downton, R.W.S. Tregidgo, et al., "Recognition of handwritten British postal addresses," *From Pixels to Features III: Frontiers in Handwriting Recognition*, S. Impedovo and J.C. Simon, eds., pp. 129-143, 1992.
[9] D. Guillevic and C.Y. Suen, "Cursive script recognition: A sentence level recognition scheme," *Proc. Fourth International Workshop on Frontiers in Handwritten Recognition*, pp. 216-223, Taipei, Taiwan, 1994.
[10] J.T. Favata, S.N. Srihari and V. Govindaraju, "Off-line handwritten sentence recognition," *Proc. Fifth International Workshop on Frontiers in Handwritten Recognition*, pp. 171-176, Essex, England, 1996.
[11] S.N. Srihari, R.K. Srihari and V. Govindaraju, "Handwritten text recognition," *Proc. Fourth International Workshop on Frontiers in Handwritten Recognition*, pp. 265-274, Taipei, Taiwan, 1994.
[12] B. Yanikoglu and P. Sandon, "Segmentation of off-line cursive handwriting using linear programming," *Pattern Recognition*, Vol. 31, No. 12, pp. 1825-1833, 1998.
[13] G. Kim and V. Govindaraju, "Handwritten phrase recognition as applied to street name images," *Pattern Recognition*, Vol. 31, No. 1, pp. 41-51, 1998.
[14] 윤정석, 김경환, "시간지연 신경망을 이용한 영문 필기체 단어 분리", *정보과학회 '99 춘계 학술발표 논문집*, Vol. 26, No. 1, pp. 490-492, 1999.
[15] U. Mahadevan and S.N.Srihari, "Hypotheses generation for word-separation in handwritten lines," *Proc. Fifth International Workshop on Frontiers in Handwritten Recognition*, pp. 453-456, Essex, England, 1996.
[16] E. Cohen, J.J. Hull and S.N. Srihari, "Control structure for interpreting handwritten addresses," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 10, pp. 1049-1055, 1994.
[17] S.N. Srihari, V. Govindaraju and A. Shekhawat, "Interpretation of handwritten addresses in US mail stream," *Proc. Second International Conference on Document Analysis and Recognition*, pp. 291-294, Tsukuba, Japan, 1993.
[18] V. Govindaraju, et al., "Interpretation of handwritten addresses in US mail stream," *Proc. Third Sixth International Workshop on Frontiers in Handwritten Recognition*, pp. 197-206, Buffalo, USA, 1993.
[19] P.K. Kim and H.J. Kim, "Off-line handwritten Korean character recognition based on stroke extraction and representation," *Pattern Recognition Letters*, Vol. 15, No. 12, pp. 1245-1253, 1994.
[20] U. Manber, *Introduction to Algorithms: A Creative Approach*, Addison Wesley, 1989.

정 선 화

1996년 전남대학교 통계학과 졸업(학사). 1998년 전남대학교 전산통계학과 대학원 졸업(이학석사). 1998년 ~ 현재 전남대학교 전산학과 박사과정. 관심분야는 패턴인식, 신경망 학습.

김 수 형

1986년 서울대학교 컴퓨터공학과 졸업(학사). 1988년 한국과학기술원 전산학과 졸업(공학석사). 1993년 한국과학기술원 전산학과 졸업(공학박사). 1990년 ~ 1996년 삼성전자 멀티미디어연구소 선임연구원. 1997년 ~ 현재 전남대학교 전산학과 조교수. 관심분야는 패턴인식, 컴퓨터 비전, 신경망 학습.