온라인 무제약 영어 필기 단어 인식을 위한 전역적 특성 분석

김재륜, 하진영
강원대학교 컴퓨터공학과

Global Feature Analysis in On-Line Unconstrained Handwritten English Word Recognition

Jaeryoon Kim and Jin-Young Ha
Dept. of Computer Eng., Kangwon National University

요 약

온라인 문자인식에 대한 연구는 지난 30여년에 걸쳐 수행되었지만, 무제약 필기 단어 인식에 대한 연구는 활성화 된지가 오래되지 않은 실정이다 필기자에 따른 다양한 시체의 변이와 방대한 탐색공간, 그리고 PDA(personal digital assistant)등의 제약된 계산 능력으로 인해 학문적으로는 좋은 연구결과가 나오고 있지만 실용화에는 아직도 해결해야 할 문제가 많다. 대부분의 온라인 문자 인식 시스템에서는 인식 시스템 자체의 성능만으로는 인식 성능의 한계가 있기 때문에 여러 가지 외부 지식을 사용한다 그 중 대표적인 것이 단어 사전을 이용하는 것인데, 단어 사전의 크기를 미리 줄일 수 있다면 인식기의 성능이 좋아질 수 있다. 본 연구에서는 온라인 무제약 영어 필기 단어 인식을 위한 필기 데이터의 전역적 특성을 분석하고, 각각의 특성에 따른 사전 감축 비율과 오류에 대해 연구하고자 한다.

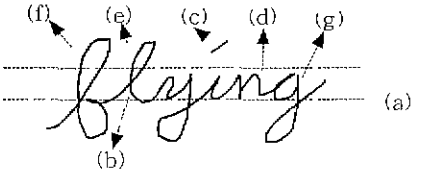
1. 서 론[1)]

온라인 문자 인식시스템 분야에 있어서, 인식 시스템 자체의 성능에는 한계가 있기 때문에, 많은 인식시스템에서 여러 외부지식을 사용한다[1-4] 그 중 본 논문에서 사용하는 외부지식은 단어 사전이다 이 단어 사전은 인식 시에 생성되는 후보를 한정시켜주는 것으로 단어 사전의 크기가 작을수록 인식 시스템에 좋은 영향을 미치게 된다. 단어 사전을 이용하는 데 있어서 단어 사전의 크기가 그 인식 시스템의 성능에 많은 영향을 미치므로, 단어 사전의 크기를 줄이는 연구가 필요하다[5] 단어 사전을 줄이기 위해서는 필기된 데이터의 전역적 특성(global feature)을 이용한다. 이 전역적 특성이라 할 수 있는 것으로는 단어 자체의 모양(word contour), 점(dot), 수평바(horizontal bar), 하향획의 수 (number of down strokes), 하향획의 유형(down stroke type), 획의 밀도(stroke density)등을 들 수 있으며 필기 시에 나타나는 루프(loop), 수직바(vertical bar), 뾰족점(cusp), 혹(hump)을 들 수 있다[5,6]

본 연구를 통하여 무제약 영어 필기 단어 인식을 위한 전역적 특성을 분석하여 수만 단어 이상의 탐색공간의 크기를 수백 단어 이내로 줄임으로써 속도 향상을 꾀하고, 터무니없는 오인식 결과도 미리 전역적 특성을 통해 후보에서 제외시킴으로써 회피하고자 한다.

2. 전역적 특성

영어 필기에서 고려할 수 있는 전역적 특성으로는 단어 자체의 모양(word contour), 점의 개수, 수평바(horizontal bar)의 개수, 하향획(down stroke)의 개수, 하향 획(down stroke)의 유형, 루프(loop), 수직바(vertical bar), 뾰족점(cusp), 혹(hump), 획의 밀도(stroke density)등을 들 수 있다 본 연구에서 사용한 전역적 특성을 자세히 살펴보면 다음과 같다



(a) base line (b) down stroke (c) dot (d) a-type down stroke (e) l-type down stroke (f) f-type down stroke (g) g-type down stroke

<그림 1> 필기에서 얻어지는 전역적 특성

2.1. 기준선 (base line)

영어 필기를 할 때에는 가상의 가로 선을 생각하고 필기를 하게 된다. 이렇게 영어 필기의 기준이 되는 가로선을 기준선이라고 한다 기준선은 다른 전역적 특성들을 찾아내는데 주요한 지표가 되는 특성이다[7].

2.2.하향획의 개수

1) 본 연구는 한국과학재단의 우수공학연구센터 지원에 의한 인공지능연구센터의 기초연구비지원을 받았음

영이 필기를 할 때에 나타나는 여러 획들 중 위에서 아래로 내려 긋는 획을 하향획이라고 한다. 하향획은 같은 단이일 경우 그 수기 대부분 일정하고, 그렇기 때문에 단어의 특성을 잘 나타내주는 전역적 특성이다

2.3. 하향획의 유형

하향획은 그 길이와 위치에 따라서 4가지로 분류할 수 있다 영어 단어 필기 시에 기준이 되는 선은 모두 4개이다. 이 기준이 되는 선들에 의해서 필기 영역을 윗 부분, 중간 부분, 아래 부분의 3개로 나눌 수 있다. 하향획이 각 영역의 어디에 위치하는 가에 따라서 a-유형, l-유형, g-유형, 그리고 f-유형의 4개의 유형으로 분류할 수 있다

a-유형	: 중간 부분에만 그어지는 하향획이다. 'a', 'e'등을 쓸 때 나타나며 <그림 1>에서는 (d)가 그 예이다
l-유형	: 윗부분과 중간 부분에 걸쳐 그어지는 하향획이다. 'l','k'등을 쓸 때 나타나며 <그림 1>에서는 (e)가 그 예이다.
g-유형	: 중간부분과 아래부분에 걸쳐 그이지는 하향획이다. 'p', 'g'등을 쓸 때 나타나며 <그림 1>에서는 (g)가 그 예이다
f-유형	위, 중간, 아래 세 영역에 걸치서 그어지는 하향획이다. 필기체 'f'를 쓸 때 나타나며 <그림 1>에서는 (f)가 그 예이다

2.4. 루프(loop)

영어 필기를 할 때 생기는 둥근 모양의 폐곡선을 루프리고 정의하였다 필기체 'l'이나 'k'등을 쓸 때 나티나는 전역적 특성이다

2.5. 수직바(vertical var)

수직바는 루프의 특수한 경우이다. 필기시에 하향획과 상향획이 붙어 일직선과 가까운 루프기 만들어질 때가 있는데, 이때에 생기는 일직선에 가까운 루프를 수직바라고 정의하였다

2.6. 뾰족점(cusp)

필기기 상향획에서 하향획으로 바뀔 때, 필기 획의 위 부분 끝이 뾰족해 질 때기 있다 이런 뾰족한 획을 뾰족점이라 정의하였다.

2.7. 혹(hump)

필기가 상향획에서 하향획으로 바뀔 때, 필기 획의 위 부분 끝이 뭉툭해 질 때가 있다. 이런 뭉툭한 획을 혹이라 정의하였다

3. 전역적 특성 탐색 및 단어사전 필터링

3.1. 기준선 찾기

영어 필기에서 기준선을 찾는 문제는 상당히 까다로운 문제이다 필기 된 단어 자체가 기준선이 모호할 수 있기 때문이다. 기준선을 찾을 때 기본적으로 적용되는 가정은 각 하향획들의 최소 y 좌표점(y minima)이 기준선 부분에 집중적으로 존재할 것이라는 것이다. 또 이와 비슷하게 하향획들의 최대 y 좌표점(y maxima)이 두 번째 기준선(half line)에 집중적으로 존재할 것이라고 가정하고 있다 이러한 가정하에 주어진 필기 입력에서 최대 y 좌표점과 최소 y 좌표점들을 이용하여 기준선을 찾아낸다.

3.2. 하향획의 유형 찾기

기준선 4개를 모두 찾으면 하향획에 대해서 각각의 유형을 얻을 수 있다. 하향획의 유형을 얻는 방법은 첫 번째 기준선과 두 번째 기준선 사이의 중간선을 구하고, 세 번째 기준선과 네 번째 기준선 사이의 중간선을 구한뒤에 이 선들을 기준으로 구하게 된다.

```
downstroke[] = GetDownstroke().
yminima[] = GetYminima().
yminima_num = GetYminimanum()
for(i = 0. i< 2. i++) {
   if(yminima_num > 0) {
      gap =
         ((yminima[max] - yminima[0]) /
         downstroke_num) *1 5.
      if(downstroke_num == 1) {
         base line = yminima[0].
      } else if(downstroke_num >= 2) {
         seed1 = yminima들중 가장 큰 값.
         seed2 = yminima들중 가장 작은 값.
         if( seed1,seed2의 차 ( gap) {
            baseline =  yminima의 평균
         }
         DoClustering() {
            do {
               yminima들을 seed1과 seed2중 가까운 것으로 분류한다
               점의수가 많은 seed를 선택
               점의수가 적은 seed에 속한 ymin ma를 비림.
            } while(선택된 seed값에 변화가 있는동안)
         }
      }
   }
}
baseline = 선택된 seed.
y maxima에 대하여 위의 과정을 다시 수행.
halfline = 선택된 seed.
```

<알고리즘 1> base line을 찾는 방법

3.3. 루프(loop), 수직바(vertical bar) 찾기

루프(loop)는 필기 시에 선이 시작점에서 멀어졌다가 가까워질 때 생긴다 시작점부터 시작해서 임계전뒤의 점들을 검색하여 시작점과 한계값 이상 가까워지면 루프로 판정한다 특정한 수 이상의 점들을 검색해도 루프가 발견되지 않으면 실패로 판정하고 다음 점에 대하여 루프를 찾는다 루프를 찾았을 때, 그 루프가 지나치게 넙작하면 그 루프는 수직바로 판정한다

```
range = (vmaxima - ymin ma)/8.
i = 0.
while(point[i] != 끝) {
   start_point = point[i].
   for(j = 15. j ( 50. j++) {
      dist = start_point와 point[j]의 거리.
      if(dist (= range) {
         loop 발견.
         if( loop의 긴지름 ) (loop의 작은지름*6)) {
            bar 발견.
         }
      }
   }
}
```

<알고리즘 2> loop와 vertical bar를 찾는 방법

3.4. 뾰족점(cusp), 혹(hump) 찾기

뾰족점(cusp)과 혹(hump)은 획이 상향에서 하향획으로 바뀔 때 나타나는 특성이다. 상향획에서 하향획으로 바뀔 때 그 부분이 뾰족하면 뾰족짐(cusp) 뭉툭하면 혹(hump)으로 한다

3.5. 필터링(filtering)

전역적 특성을 이용하여 단어 사전을 감축하는 작업이 필터링이다. 하향획의 종류를 필터링 조건으로 할 때에는 입력으로부터 하향획의 유형을 나타내는 문자열을 추출하고, 이 문자열을 <표 1>을 근거하여 사전에서 얻어낸 하향획 유형 문자열과 비교하여 필터링을 수행한다.

```
for(i = 2, i< point_num-2, i++) {
   x1 = point[i-2].x,
   y1 = point[i-2].y,
   x2 = point[i+2].x,
   y2 = point[i+2].y,
   max = 현재점과 점1과의 거리,
   if(점1,점2의 거리 < max) {
      cusp 발견,
   } else {
      hump 발견,
   }
}
```

<알고리즘 3> cusp, hump를 찾는 방법

루프, 수직바, 뾰족점 및 혹에 대해서는 필터링을 하기 전에, 각 단어에 대해 전역적 특성의 개수의 정보를 기지고 있는 단어 사전을 준비한다. 이 사전에 포함되는 내용은 단어와 각 전역적 특성들이 나타날 수 있는 최대값 및 최소값이다. 사전의 값들은 알파벳 각각의 전역적 특성들의 출현 평균 및 분산을 구한뒤 이들 값을 더하거나 빼서 얻는다.

필기 입력에 대해서 구한 전역적 특성의 개수와 단어 사전의 평균 출현개수를 비교하여 필터링작업을 수행하게 된다. 전역적 특성의 개수가 최소 값과 최대 값 범위에 포함된다면 그 단어는 선택되고, 그렇지 않다면 제외된다.

글자	유형	글자	유형	글자	유형	글자	유형
a	a	h	la	o	a, aa	v	a, aa
b	l,la	i	a	p	ga	w	aa, aaa
c	a	j	g	q	ag, aga	x	aa
d	al	k	la, laa	r	a, aa	y	ag
e	a, aa	l	l	s	a, aa	z	a, ag
f	l, f, fa	m	aaa	t	l		
g	al	n	aa	u	aa		

<표 1> 영어 소문자 알파벳의 하향획 유형

4. 실험 및 결과 분석

이 실험에서 사용한 사전의 단어의 총 개수는 22,433개이다. 입력으로 사용한 데이터는 모두 단이 사전에 있는 단어들로서 6인이 필기한 3,287개의 단어이다. 하향획의 유형, 루프, 수직바, 뾰족점, 혹들에 대한 각각의 필터링 결과가 <표 2>에 나와 있다.

전역적특성	감축율(%)			오류율(%)		
	σ*1	σ*2	σ*3	σ*1	σ*2	σ*3
점 및 수평바	14.19			32.64		
하향획의 유형	2.82			33.68		
루프	56.81	84.83	95.27	35.65	7.94	1.33
수직바	64.16	95.91	98.22	30.05	2.00	0.48
뾰족점	49.34	68.54	97.75	44.93	23.48	0.39
혹	66.03	89.03	94.44	18.55	0.88	0.30

<표 2> 전역적 특성의 필터링 결과

σ*2는 알파벳의 전역적 특성 출현 빈도수 평균과 표준편차를 구한 뒤, 각 알파벳의 전역적 특성 출현빈도 최소값과 최대값을 구할 때 표준편차의 2배를 했다는 의미이다. σ*3은 세배를 의미한다. 세 경우에 대하여 하향획의 유형에 의한 필터링의 결과가 같게 나오는 이유는 하향획의 유형에 대해서는 출현 횟수가 아니라 출현 순서를 조사하기 때문에 표준편차의 영향을 받지 않기 때문이다. 실험 결과에서 사전 감축률이 기대했던 것보다 좋지 않은 이유는 다음과 같이 분석된다. 하향획의 유형을 이용하는 경우, 기준선을 먼저 찾게 되는데, 입력 데이터 자체가 기준선이 모호한 경우에는 잘못 필터링될 확률이 높아진다. 루프, 수직바, 뾰족점, 혹들에 대해서는 필터링의 방법에 있어서 단순히 전역적 특성의 출현 수를 이용하였는데, 이 방법이 단어들의 전역적 특성에 대한 차이점을 충분히 적용시켜주지 못한 것 같고, 뾰족점이나 혹을 찾는 방법이 하향획을 기준으로 찾게 되어 있는데, 이것도 역시 단순히 개수를 세는 것으로는 의미가 없는 필터링 방법이었던 것 같다. 오류율에 대해서는 전역적 특성을 찾아낼 때 사용하는 여러 인수들의 조정이 부족했던 것 같다.

5. 결론 및 향후 과제

영어 필기 단어 인식문제에 있어서 널리 사용되는 외부 지식인 단어 사전을 줄이는 방법에 대해서 연구를 진행하였다. 본 논문에서는 영어 단어 필기에 나타나는 여러 가지 전역적 특성을 살펴보았고, 그 중 하향획의 유형, 루프, 수직바, 뾰족점 그리고 혹을 이용하는 방법을 적용시켜보았다. 필터링 방법에 있어서는 단어 내에서 전역적 특성들이 나타나는 개수를 이용하였다. 보다 더 나은 성능을 위하여 두 개 이상의 전역적 특성의 조합, 필터링 방법의 확장, 다른 전역적 특성의 적용 등에 대한 연구가 필요하다.

참고 문헌

[1] 이성환, 문자인식 - 이론과 실제, 홍릉과학출판사, 서울, 1993.

[2] C.C Tappert, C.Y Suen and T Wakahara, "The state of the art in on-line handwriting recognition," IEEE Trans on Pattern Analysis and Machine Intelligence, Vol 12, No 8, 1990, pp. 787-808. 1990.

[3] J-Y. Ha, S.-C. Oh and J-H Kim, "Recognition of Unconstrained Handwritten English Words with Character and Ligature Modeling," in International Journal of Pattern Recognition and Artificial Intelligence, Vol 9, No 3, June 1995, pp 535-556

[4] C Farouz, M Gilloux and J-M. Bertille, "Handwritten word Recognition with Contextual Hidden Markov Models," in 6th Int. Workshop on Frontiers in Handwriting Recogniton.(Taegon, Korea), pp.133-142, 1998

[5] 최동원, "온라인 영어 필기의 전역적 특성을 이용한 사전으로부터의 후보단어 선택", 석사학위논문, 한국과학기술원 전산학과, 1995.

[6] C.A. Higgins and D.M Ford, "A new segmentation method for cursive script recognition," in 2nd Int Workshop on Frontiers in Handwriting Reccognition, (Chateau de Bonas, France), pp 241-252 1991

[7] E.R.Brocklehurst and P.D Kenword, "Preprocessing for Cursive Script Recognition, "NPL Report DITC 132/88, November 1988