획 밀도를 이용한 한영 구분

원 남 식[†] · 전 일 수[†] · 이 두 한[††]

요 약

다중 문자 환경의 문서인식 시스템에서 문자를 인식하기 전에 문자의 종류를 먼저 구분하는 것은 인식률의 향상에 중요한 요인이 된다. 각 나라의 문자는 그 문자마다 고유의 구성상의 다양한 특징을 가진다. 본 연구에서는, 문자를 구분하기 위한 방법으로 획 밀도 값을 이용하였고, 대상 문자는 한글과 영문으로 한정하였다. 다양한 형태의 활자가 사용되는 문서에 적용하기 위해 입력 데이터는 정규화 과정을 거친 후 처리되었다. 제안된 방법은 90% 이상의 높은 확률로 한영 구분이 가능함을 실험 결과로써 입증하였다.

Distinction of the Korean and English Character Using the Stroke Density

Nam Sik Won[†] · Il Soo Jeon[†] · Doo Han Lee[††]

ABSTRACT

It is an important factor to distinguish the kind of the character for increasing recognition rate before the character recognition in the document recognition system composed of the multi-font and multi-letters. All the letters of each country have a various unique characteristic in the each composition. In this paper, we used the stroke density as a method to distinguish the letter, and it has been adopted only Korean and English character. Input data is processed by the normalization to adopt multi-font document. Proposed method has been proved by the results of experiment the fact that the distinction probability of the Korean and English is more than 90%.

1. 서 론

국내외에서 문서인식에 관한 연구는 매우 활발하게 이루어지고 있으나, 대부분이 하나의 특정 문자에 국한 된 연구이다. 그러나 국내에서 사용되는 일반적인 문서 환경은 영문과의 혼용이 불가피한 상태이므로, 한글과 영문이 혼용된 문서를 인식하기 위한 연구는 수행되어야 할 과제가 된다. 현재 한영 구분에 관한 연구는 국내외에서 거의 수행되지 않고 있는 실정이나, 실제의 문서인식 시스템에서는 인식률 향상을 위하여 한글과 영문을 먼저 구분하여 인식하는 기능은 매우 요구되고 있다.

문서인식을 위한 문서 구조 분석은 상향식 방법과 하향식 방법으로 나누고, 상향식 방법으로는 연결 요소 분석 방법[1]과 인접 선분 밀도 분석 방법[2, 3]등이 있고, 하향식 방법으로는 런 길이 평활화 방법, 투영 윤곽 분석 방법, Fourier 변환에 의한 방법, 원형 정합에 의한 방법, 교차 횟수 분석 방법등이 있다[4, 5, 6]. 문서인식을 위한 전처리 과정으로서 정규화 과정이

† 정 회 원:경일대학교 전자계산학과
†† 정 회 원:경동전문대학 정보처리학과
논문접수:1996년 11월 21일, 심사완료:1997년 5월 22일

있다. 일반적인 정규화의 목적은 외접 다각형 또는 모멘트[7] 등을 이용하여 영상을 일정한 위치, 크기, 기울기 등을 갖는 영상으로 변환하는데 있다. 정규화 방법에는 선형 변환과 비선형 변환이 있다. 선형 변환은 선형성을 보존하는 특성이 있으며, 수학적인 표현이 쉬운 반면, 불규칙적이고 부분적으로 발생하는 형태 왜곡을 보상하기 위해 다양한 비선형 형태 정규화 방법이 제안되었다[6, 8, 9, 10, 11, 12].

문서 인식 시스템에서 문자의 인식율을 향상하기 위해서는 먼저 인식 대상 문자의 종류를 알므로써 보다 빠르고 정확한 인식이 가능해 진다. 현재 국내에서 사용되는 대부분의 문서들은 국문과 영문 그리고 한문이 혼용되어 사용되고 있으므로, 이들 문자를 먼저 구분한 후 인식하므로서 인식율을 보다 향상시킬 수 있다. 한글, 한자, 영문을 동시에 높은 확률로 정확히 구분하는 것은 매우 어려운 문제가 된다. 한글과 영문은 그 글자의 구성상의 특징이 서로 다르지만, 한문은 한글과 거의 유사한 형태의 수평, 수직, 경사 획 성분에 의해 구성되어 있기 때문이다. 그러나, 현재 일반적으로 사용되는 문서에서는 한글과 영문의 혼용이 가장 많이 이루어 지고 있는 실정이므로, 본 연구에서는 한글과 영문을 구분하는 하는 것을 연구 목적으로 하였다. 한영 구분에서 각 문자는 매우 다양한 형태의 활자로 구성되어 이를 완벽하게 구분할 수 있는 방법은 존재할 수 없으므로, 일정한 수준의 실용성 있는 범위에서의 한영 구분이 가능하면 된다. 또한 알고리즘을 수행하는데 소요되는 시간은 인식 시간에 대비하여 상대적으로 무시될 수 있는 범위라야 한다. 본 연구에서는 높은 확률로 구분이 가능하고, 적은 경비로 처리할 수 있는 획 밀도에 의한 한영 구분을 위한 방법을 제시하였다. 다양한 형태의 활자가 사용되는 문서에 적용하기 위해 입력 데이터는 선형 정규화 과정을 거친 후 처리하였다. 제안된 방법은 적은 경비로 90% 이상의 높은 확률로 한 영 구분이 가능함을 실험 결과로서 입증하였다.

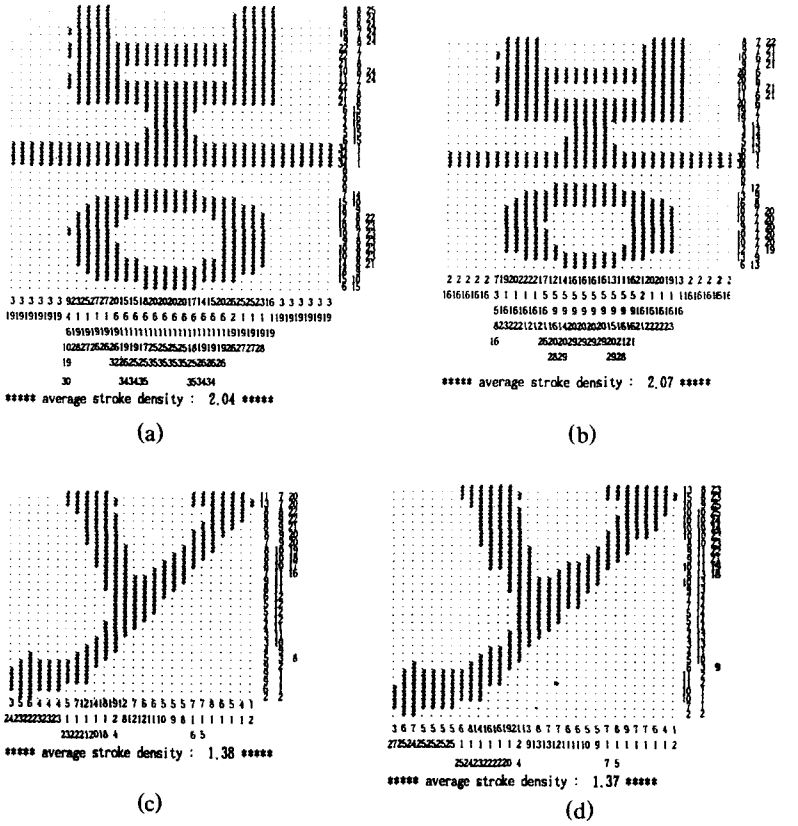
실험 방법은 많이 사용되고 있는 한글 폰트 5종에 대해서 각각 한글 완성형 2350자에 대하여 획 밀도 측정을 수행하였고, 영문 폰트는 일반성 있는 5종의 대문자, 소문자 폰트에 대해 News Week지에 게재된 영문을 이용하여 획 밀도 측정을 수행하였다. 연구 수행의 실험환경은 Pentium PC, Epson GT-9000 image scanner를 사용하였다.

2. 정규화

형태 정규화 방법은 선형 형태 정규화 방법과 비선형 형태 정규화 방법으로 분류되고[13], 비선형 형태 정규화 방법은 투영하는 특징의 종류에 따라 점밀도[12]를 이용하는 방법과 획 밀도를 이용하는 방법으로 구분될 수 있다. 획 밀도를 이용하는 방법은 교차 횟수(crossing count)에 바탕을 둔 획 밀도[10]를 이용한 방법, 획 간격(line interval)에 바탕을 둔 획 밀도[9]를 이용하는 방법, 그리고 내접원(inscribed circle)에 바탕을 둔 획 밀도[11]를 이용하는 방법으로 구분된다. 본 논문에서의 정규화는 선형 형태 정규화 방법에 의해 구현하였다.

대부분의 인쇄소에서 사용되는 활자들은 동일한 글씨체라도 서로 미세한 차이를 나타낼 수 있으므로, 이러한 활자들에 의해 인쇄된 문서를 인식하는데 일정한 통계 치의 적용은 오차의 범위를 더 크게 하는 결과가 될 수도 있다. 그러므로 문자 구성의 특징을 추출하기 위한 방법으로서, 통계적인 수치에 의해 처리하기 위한 방법에서는 정규화 과정이 필요하게 된다. 본 연구에서는 입력된 데이터에서 특징을 추출하는데 적당한 크기의 데이터로 처리하기 위해 정규화 과정에서 데이터의 크기를 가변으로 조절하여 정규화 된 데이터를 구하도록 하였다.

정규화 된 데이터의 크기에 관련된 것으로서 데이터가 크면 특징 데이터의 추출에는 유리하나 처리 시간이 문제가 되고, 크기가 적으면 처리 시간은 단축되나 정규화 과정에서 입력된 데이터의 일부가 소멸될 수도 있다. 문자의 구성 획이 복잡한 경우 획의 일부가 사라질 수도 있으므로, 정규화 된 데이터의 크기는 특징 데이터를 얻기 위한 적정한 크기로 결정되어야 한다. (그림 1)은 한글과 영문에서의 정규화된 결과를 보인다. 획 밀도를 얻기 위한 정규화 데이터의 크기는 30×30으로 하였다. 300dpi의 해상도로 데이터가 입력될 때, (그림 1)에서와 같이 정규화 된 결과에서 한글은 축소되었고, 영문은 확대되었음을 알 수 있다. (그림 1)에서 우측면과 밑면에 보이는 수치들의 의미는, 첫째 줄에 나타나는 수치는 수평 방향과 수직 방향으로 투영된 화소의 숫자를 나타내고,



(그림 1) 입력 데이터의 정규화 (a), (c) : 입력 데이터
(b), (d) : 정규화 데이터
(Fig. 1) Normalization of the input data (a), (c) : input data (b), (d) : normalization data

두 번째 줄부터는 수평, 수직 방향에서 각각 획이 시작되는 위치를 순서대로 나타낸다. 획이 시작되는 위치는 화소가 백화소에서 흑화소로 바뀌는 점이 된다. 그리고 평균 획 밀도를 나타내었다.

3. 획 밀도 측정에 의한 한 영 구분 알고리즘

한영 구분을 위한 방법으로서 한글과 영문에서 문자 구성 형태상의 차이를 분석하면 직선 획과 곡선 획을 사용하는 비율의 차이, 글자 크기의 차이, 문자를 구성하는 연결 요소 수의 차이와 획 밀도의 차이 등 여러 가지의 다양한 형태로 한영 구분을 위한 분석이 가능하다. 본 연구에서의 한영 구분을 위한 방법은 교차 횟수(N_c : crossing count)에 의한 획 밀도 측정값을 이용한 한영 구분 알고리즘을 제안하였다.

3.1 교차 횟수에 의한 획 밀도

교차 횟수(N_c)는 문서의 수평 또는 수직의 주사 방향에서, 백화소에서 흑화소로 바뀌는 점의 수 또는 흑화소에서 백화소로 바뀌는 점의 수를 각 주사선 별로 계산한 것이다. 본 논문에서의 교차 횟수에 대한 정의는 백화소에서 흑화소로 바뀌는 점의 수로 정의하였다. 그러므로 임의의 화소점에서의 수평 방향의 교차 횟수는 (식1)과 같고, 수직 방향의 교차 횟수는 (식2)와 같이 된다. 획 밀도는 교차 횟수를 수평 방향과 수직 방향으로 투영하여 이들 교차 횟수의 합을

주사한 화소 수로 나눈 값을 획 밀도(ρ)로 정의하였다. 그러므로 수평 방향 획 밀도 $\rho_h(i)$는 (식3)과 같고, 수직 방향 획 밀도 $\rho_v(j)$는 (식4)와 같이 표현된다. 평균 획 밀도 ρ_a는 수평 방향 획 밀도와 수직 방향 획 밀도의 합을 평균한 값으로 (식5)와 같으며, 정규화 이진 영상에서의 평균 획 밀도 ρ_{na}는 (식6)과 같다. 한 영 구분을 위한 임계 획 밀도 ρ_t는 많은 획 밀도 측정 데이터에 의해서 결정된다.

$f(i, j)$: I×J 크기의 입력 이진 영상(i=1 … I, j=1 … J)
f_n: m×n 크기의 정규화 이진 영상
N_c: 교차 횟수(crossing count)
ρ: 획 밀도(stroke density)
ρ_a: 평균 획 밀도(average stroke density)
ρ_t: 임계 획 밀도(threshold stroke density)
ρ_{na}: 정규화 이진 영상에서의 평균 획 밀도

임의 화소점에서 수평 방향 교차 횟수:

$$N_{hc}(i)=\sum_j \overline{f(i, j)} \cdot f(i, j+1) \tag{1}$$

임의 화소점에서 수직 방향 교차 횟수:

$$N_{vc}(j)=\sum_i \overline{f(i, j)} \cdot f(i+1, j) \tag{2}$$

수평 방향 획 밀도:

$$\rho_h(i)=\frac{\sum_i N_{hc}(i)}{I} \tag{3}$$

수직 방향 획 밀도:

$$\rho_v(j)=\frac{\sum_j N_{vc}(j)}{J} \tag{4}$$

평균 획 밀도:

$$\rho_a=\frac{\rho_h(i)+\rho_v(j)}{2} \tag{5}$$

정규화 된 이진 영상에서의 평균 획 밀도:

$$\rho_{na}=\frac{\rho_h(m)+\rho_v(n)}{2} \tag{6}$$

3.2 한 영 구분 알고리즘

입력된 문자 이미지 $f(i, j)$를 정규화 한다. 그리고 정규화 된 이미지 데이터 $f_n(m, n)$와 한 영 구분을 위한 임계 획 밀도 ρ_t를 사용하여 한글과 영어를 구별한다.

한 영 구분 알고리즘
{ 입력: $f(i, j)$, ρ_t
  출력: 한 영 구분 }
Convert $f(i, j)$ to $f_n(m, n)$;
Calculate ρ_{na};
if $\rho_{na} > \rho_t$ then 한글
  else 영문;

4. 획 밀도 측정

획 밀도 측정을 위한 객관성 있는 데이터로서 한글은 한글 완성형 2350자를 명조, 신 명조, 고딕, 궁서, 샘물체등의 5가지 폰트에 대하여 각각 실험하였고, 영문은 News Week지에서 문장을 발췌하여 이를 명조, 신명조, 고딕, 산세리프, 이탤릭체에 대하여 각 2500자에 대해 획 밀도를 측정한 후, 그 결과를 비교 분석하였다.

4.1 한 영문 각종 활자체에서의 획 밀도 분포

모든 문자들은 각 각 독특한 획 구성 형식을 가지므로 각 문자의 획 밀도 특성은 서로 다르게 나타나므로, 이러한 획 밀도 특성을 이용함으로서 문자의 구분이 가능하게 된다. 한글은 획의 구성이 영문에 비해 비교적 복잡하므로 획 밀도 값이 높게 나타나므로 획 밀도 값에 의해서 한글과 영문을 구분할 수 있다. 한글에서는 대부분의 글자의 획 밀도(ρ)는 ρ_{na}값이 1.9이상의 높은 값에 분포되어 있으므로, 〈표 1〉에서는 임계 획 밀도값 ρ_t보다 ρ_{na}값이 적은 글자 수를 나타내었다. 영문에서는 대부분의 글자의 ρ_{na}값의 분포가 1.8미만의 낮은 값에 분포되어 있으므로 〈표 2〉와 〈표 3〉에서는 임계 획 밀도값 ρ_t보다 ρ_{na}값이 큰 글자 수를 나타내었다. 〈표 4〉에서 〈표 6〉까지는 임계 획 밀도 값 ρ_t를 구하기 위한 것으로서, 〈표 4〉에서는 한글에 대한 획 밀도에 따른 구분 율을 나타내었고,

〈표 5〉와 〈표 6〉에서는 영문에 대한 획 밀도에 따른 구분 율을 백분율로 각 각 나타내었다. 영문 소문자에서는 획 밀도 분포가 〈표 3〉에서와 같이 ρ_t값의 범위가 0.1 정도의 적은 범위에서 큰 변화를 보이므로 〈표 3〉의 데이터로는 한 영 구분을 위한 정확한 임계 획 밀도 ρ_t를 구할 수 없으므로, 변화가 가장 심한 소구간의 범위만을 확대하여 〈표 7〉에서 〈표 11〉까지 나타내었다. 한글은 〈표 1〉에서 Kf1은 명조, Kf2는 신 명조, Kf3은 고딕, Kf4는 궁서, Kf5는 샘물체 등의 5가지 폰트에 대하여 각각 실험하였고, 〈표 2〉는 영문 대문자로서 Euf1은 명조, Euf2는 신 명조, Euf3은 고딕, Euf4는 산세리프, Euf5는 이탤릭체에 대하여 각각 실험하였고, 〈표 3〉은 영문 소문자로서 Euf1은 명조, Euf2는 신 명조, Euf3은 고딕, Euf4는 산세리프, Euf5는 이탤릭체에 대하여 각각 실험하였다.

〈표 1〉 한글 2350자에 대한 획 밀도 분포($\rho_t > \rho_{na}$)
〈Table 1〉 The stroke density distribution of 2350 characters in the korean($\rho_t > \rho_{na}$)

ρ_t / font	1.6	1.7	1.8	1.9	2.0	2.1
Kf1	18	41	92	165	302	497
Kf2	14	27	55	98	194	305
Kf3	9	23	56	103	205	356
Kf4	40	107	241	456	746	1014
Kf5	110	206	354	574	857	1156

〈표 2〉 영문 대문자 2500자에 대한 획 밀도 분포($\rho_t < \rho_{na}$)
〈Table 2〉 The stroke density distribution of the upper case 2500 characters in the english($\rho_t > \rho_{na}$)

ρ_t / font	1.7	1.8	1.9	2.0	2.1	2.2
Euf1	1168	661	397	91	56	14
Euf2	1220	990	782	211	60	37
Euf3	1115	701	145	2	0	0
Euf4	463	171	8	0	0	0
Euf5	610	150	73	59	27	2

〈표 3〉 영문 소문자 2500자에 대한 획 밀도 분포($\rho_t < \rho_{na}$)
〈Table 3〉 The stroke density distribution of the lower case 2500 characters in the english($\rho_t > \rho_{na}$)

ρ_t / font	1.8	1.9	2.0	2.1	2.2	2.3
Elf1	630	339	47	0	0	0
Elf2	668	398	61	36	9	0
Elf3	326	166	13	4	0	0
Elf4	231	24	2	0	0	0
Elf5	512	191	38	8	0	0

〈표 4〉 한글 임계 획 밀도에 대한 구분율(%)
〈Table 4〉 Distinction percentage by the threshold stroke density in the korean character(%)

ρ_t / font	1.6	1.7	1.8	1.9	2.0	2.1
Kf1	99.2	98.2	96.1	93.0	87.2	76.5
Kf2	99.5	98.8	97.7	95.8	91.7	87.1
Kf3	99.6	99.0	97.6	95.6	91.3	84.9
Kf4	98.3	95.4	89.7	80.6	68.3	56.9
Kf5	95.3	91.2	84.9	75.6	63.5	50.8

〈표 5〉 영문 대문자에서의 임계 획 밀도에 대한 구분율(%)
〈Table 5〉 Distinction percentage by the threshold stroke density in the upper case english character(%)

ρ_t / font	1.7	1.8	1.9	2.0	2.1	2.2
Euf1	53.3	73.6	84.1	96.4	97.7	99.4
Euf2	51.2	60.4	68.7	91.5	97.6	98.5
Euf3	55.4	72.0	94.2	99.9	100	100
Euf4	81.5	93.2	99.7	100	100	100
Euf5	75.6	94.0	97.1	97.6	98.9	99.9

〈표 6〉 영문 소문자에서의 임계 획 밀도에 대한 구분율(%)
〈Table 6〉 Distinction percentage by the threshold stroke density in the lower case english character(%)

ρ_t / font	1.8	1.9	2.0	2.1	2.2	2.3
Elf1	74.8	86.4	98.1	100	100	100
Elf2	73.3	83.0	97.6	98.6	99.6	100
Elf3	87.0	93.4	99.5	99.8	100	100
Elf4	91.8	99.3	99.9	100	100	100
Elf5	79.5	92.4	98.5	99.7	100	100

4.2 획 밀도값에 의한 한영 구분예

실행 예로서 (그림 2)에 임의의 자소가 분리된 한영 데이터로서 획 밀도 측정에 의한 한영 구분의 수행 예를 나타내었다. 한영 구분을 위한 임계 획 밀도는 1.9로서 구분하였다.

4.3 획 밀도 분포에 대한 고찰

서로 다른 종류의 문자는 획 밀도 값에 의해서 구분될 수 있다. 이를 구분하기 위한 획 밀도 값을 임계 획 밀도(ρ_t)라 하고, 입력 데이터의 평균 획 밀도는 정규화 한 이진 영상에서의 평균 획 밀도(ρ_{na})로 표시한다. 한글과 영문의 획 밀도 특성은 한글이 높게 나타나므로, 임계 획 밀도 값이 설정된 상태에서의 한영 구분은 $\rho_t < \rho_{na}$이면 한글, $\rho_t > \rho_{na}$이면 영문으로 구분

〈표 7〉 Elf1에 대한 임계 획 밀도 분포와 구분율
〈Table 7〉 The threshold stroke density distribution and distinction percentage of the Elf1

ρ_t	1.8	1.9	1.92	1.93	1.94	1.95	2.0	2.1	2.2	2.4
$\rho_t < \rho_{na}$	630	339	256	197	110	91	47	0	0	0
구분율(%)	74.8	86.4	90.8	92.1	95.6	96.4	98.1	100	100	100

〈표 8〉 Elf2에 대한 임계 획 밀도 분포와 구분율
〈Table 8〉 The threshold stroke density distribution and distinction percentage of the Elf2

ρ_t	1.8	1.9	1.92	1.94	1.96	2.0	2.1	2.2	2.3	2.4
$\rho_t < \rho_{na}$	668	398	345	250	175	61	36	9	0	o
구분율(%)	73.3	83.0	86.2	90.0	93.0	97.6	98.6	99.6	100	100

〈표 9〉 Elf3에 대한 임계 획 밀도 분포와 구분율
〈Table 9〉 The threshold stroke density distribution and distinction percentage of the Elf3

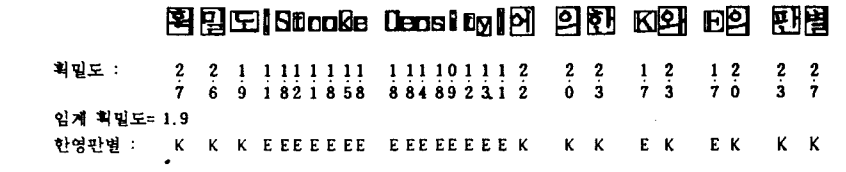
ρ_t	1.8	1.84	1.85	1.9	1.93	1.95	2.0	2.1	2.2	2.4
$\rho_t < \rho_{na}$	326	255	226	166	97	56	13	4	0	0
구분율(%)	87.0	89.8	91.0	93.4	96.1	97.8	99.5	99.8	100	100

〈표 10〉 Elf4에 대한 임계 획 밀도 분포와 구분율
〈Table 10〉 The threshold stroke density distribution and distinction percentage of the Elf4

ρ_t	1.8	1.82	1.84	1.85	1.86	1.9	2.0	2.1	2.2	2.4
$\rho_t < \rho_{na}$	231	170	126	71	59	24	2	0	0	o
구분율(%)	91.8	93.2	95.0	97.2	97.6	99.0	99.9	100	100	100

〈표 11〉 Elf5에 대한 임계 획 밀도 분포와 구분율
〈Table 11〉 The threshold stroke density distribution and distinction percentage of the Elf5

ρ_t	1.8	1.82	1.84	1.86	1.88	1.9	1.95	2.0	2.1	2.2
$\rho_t < \rho_{na}$	512	478	451	368	272	191	90	38	8	0
구분율(%)	79.5	80.9	82.0	85.3	89.2	92.4	96.4	98.5	99.7	100



(그림 2) 획 밀도를 이용한 한영 구분
(Fig. 2) Distinction of the korean and english character using the stroke density

한다. 그러므로 한영 구분을 위한 적당한 ρ_t값이 설정되어야 한다.

실험 결과로서 글자체에 따라서 획 밀도 값이 다르게 나타남을 알 수 있다. 한글에서는 〈표 1〉에서와 같이 명조, 신 명조, 고딕체에서 획 밀도 특성이 양호하게 나타나고, 궁서체와 샘물체에서는 획 밀도 특성이 나쁘게 나타남을 알 수 있다. 그러나 이러한 글씨체는 일반적인 문서에서는 거의 사용되지 않고 있다. 명조, 신 명조, 고딕체에서는 〈표 4〉에서와 같이 임계 획 밀도가 1.9에서 90%이상의 높은 확률로 구분할 수 있음을 알 수 있다. 영문에서는 〈표 2〉와 〈표 3〉에서와 같이 획 밀도가 1.9보다 적은 값에 대부분이 분포하고 있으며, 글씨체에 따라서 획 밀도 값의 변화가 심하게 나타남을 알 수 있다. 영문에서는 고딕체, 산세리프체, 이탤릭체가 획 밀도 분포 특성이 우수하게 나타났으며, 명조체와 신 명조체에서 획 밀도 분포 특성이 비교적 나쁘게 나타남을 알 수 있다. 획 밀도 측정에 의한 실험 결과 한글의 획밀도 분포는 1.9에서 2.7정도 나타났으며, 영문에서는 1.2에서 1.9 정도의 범위로 나타났다.

5. 결 론

혼용된 문서 인식 시스템에서 문자를 인식하기 위해서는 인식 대상 문자의 종류를 먼저 파악한 후, 적합한 알고리즘에 의해서 인식하는 것은 문자의 인식률과 인식 속도 향상에 큰 영향을 가져올 수 있다. 그러나 이러한 문자 구분을 위한 알고리즘의 수행에 많은 경비를 치르게 되면 문자 인식 속도에 영향을 미치게 되므로, 적은 경비에 의해 경제적인 범위의 확률로 판단되면 실용성 있는 방법이 될 수 있다. 본 연구에서는 문자의 구성상의 특징을 구분하기 위한 방법으로는 획 밀도 값을 이용하였고, 대상 문자는 한글과 영문으로 한정하였다. 그리고 다양한 형태의 활자가 사용되는 문서에 적용하기 위해 입력 데이터는 정규화 과정을 거친 후 처리되었다. 문서인식의 전처리 과정에서 완전한 문자의 구분은 사실상 불가능하므로, 어느 정도 높은 확률로 구분되면 실용 가능성 있는 방법이 된다. 제안된 방법은 빠른 수행시간으로 처리되며, 90% 이상의 높은 확률로 한 영 구분이 가능함을 실험 결과로서 입증하였다.

추후 연구과제로서는, 한영 구분하는데 획 밀도값 외에 구분 확률을 높이기 위한 보조적인 수단으로서, 각 문자의 독특한 구성상의 특징을 이용하여 획 밀도 측정에서 잘못 구분되는 경우를 극복할 수 있는 연구가 수행되어야 한다.

참 고 문 헌

[1] L. A. Fetcher and R. K. Kusutri, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, 1988, pp. 910-918.

[2] O. Iwaki, H. Arakawa, "A Segmentation Method Based on Office Document Hierachical Structure," Proc. 1987 IEEE Int. Conf. on Systems, Man and Cybernetics,1987, pp. 759-763.

[3] K. Kubota, O. Iwaki and H. Arakawa, "Document Understanding System," Proc. 7th Int. Conf. on Pattern Recognition, 1984, pp. 612-614.

[4] T. Akiyama and N. Hagita, "Atomated Entry System for Printed Documents," Pattern Recognition, Vol. 23, No. 11, 1990, pp. 1141-1153.

[5] J.-C. Oriot, D. Barba and J.-C. Salome, "Address Block Location Method Based on Transition Analysis Approach: Design and Evaluation on Flats Objects," Proc. 1st Int. Conf. on Document Analysis and Recognition, Saint-Malo, France, Sep. 1991, pp. 665-673.

[6] 이성환, "문자 인식," 1994, pp. 89-113.

[7] R. G. Casey, "Moment Normalization of Handprinted Character," IBM Journal of Research and Development, Vol. 14, Sep. 1970, pp. 548-557.

[8] Z. C. Li, T. D. Bui, Y. Y. Tang and C. Y. Suen, *Computer Transfrmation of Digital Images and Patterns,* World Scientific Press, 1989.

[9] J. Tsukumo and H. Tanaka, "Classification of handprinted Chinese Characters Using Nonlinear Normalization Methods," Proc. 9th Int. Conf. on Pattern Recognition, Rome, Italy, Nov. 1988, pp. 168-171.

[10] H. Yamata, T. Saito and K. Yamamoto, "Line Density Equalization-A Nonlinear Normalization for Correlation Method," Trans. IECE of Japan, Vol. J67-D, No. 11, Nov. 1984, pp. 1379-1383.

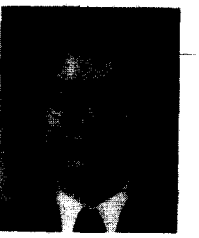
[11] H. Yamata, K. Yamamoto and T. Saito, "A Nonlinear Normalization for Handprinted Kanji Character Recognition-Line Density Equalization," Pattern Recognition, Vol. 23, No. 9, pp. 1023-1029.

[12] Y. Yamashita, K. Higuchi, Y. Yamata and Y. Haga, "Classification of Handprinted Kanji Characters by the Structured Segment Matching Method," Pattern Recognition Letters, Vol. 1, No. 5, pp. 475-479.

[13] 이성환, 박정선, "대용량 필기체 문자 인식을 위한 비선형 형태 정규화 방법의 정량적 평가," 대한전자공학회논문지, 제30권 B편 제9호, 1993년 9월, pp. 896-905.

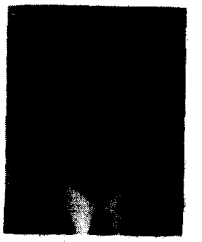
원 남 식

1974년 인하대학교 전자과 졸업(학사)
1984년 영남대학교 대학원 전자과 졸업(석사)
1996년 영남대학교 대학원 전산공학과 졸업(박사)
1976년~1978년 한국과학기술연구소 연구원
1978년~1981년 한국전자기술연구소 연구원
1981년~현재 경일대학교 전자계산학과 교수
관심분야: 문자인식, 세선화 알고리즘, 네트워크, 컴퓨터 그래픽스



전 일 수

1984년 경북대학교 전자공학과 (학사)
1988년 경북대학교 전자공학과 (석사)
1995년 경북대학교 전자공학과 (박사)
1984년~1985년 삼성전자(주)
1989년~현재 경일대학교 전자계산학과 부교수
관심분야: 문자인식, 데이터베이스



이 두 한

1987년 경북대학교 전자공학과 공학사
1991년 경북대학교 전자공학과 공학석사
1993년 경북대학교 컴퓨터공학과 박사과정 수료
1994년~현재 경동전문대학 전산정보처리과 조교수
관심분야: 객체지향 데이터베이스, 인공지능