한글 문자 영상에서의 정보량 및 엔트로피의 분포

이 성 환

충북대학교 자연과학대학 전자계산학과
충북 청주시 개신동 산 48 (우편번호 360-763)

요약

본 논문은 다양한 활자체 및 크기의 한글 문자 영상에서의 정보량 및 엔트로피의 분포에 관한 연구이다. 12 종류의 서로 다른 활자체 및 크기의 한글 문자 영상이 실험에 사용되었으며, 사용 빈도수가 높은 520 자의 한글 문자 영상에 대하여 정보량과 엔트로피를 측정하였다. 실험 결과의 분석을 통하여 정보량과 엔트로피의 측정치는 문자의 구조적 형태에 따라 변하지만 활자체에는 무관하며, 대부분의 정보량이 문자의 가장자리 부분에 위치함을 알 수 있었다.

1. 서론

한글 문자는 다양한 형태의 정보를 포함하고 있다. 따라서 한글 문자 영상에서의 정보의 분포를 알 수 있다면 한글 문자 영상의 코딩 및 탐색은 물론 한글 문자의 인식을 포함하여 한국어 정보 처리에 큰 도움이 되리라 생각한다. 정보란 여러 개의 대상(object)들로 부터 한 대상을 구별할 수 있는 상태에 관한 지식으로써 상태에 관한 불확실성(uncertainty)을 감소시키는 것이며, 엔트로피는 무질서의 정도를 나타내는 척도로써 원래는 열역학적 엔트로피와는 다른 관점에서 연구가 시작되었으나 결국 수학적인 표현은 서로 일치하며 상호 밀접한 관계를 갖고 있다[1].

한글 문자 인식에 관한 연구는 지난 20여 년 간 꾸준히 진행되어 왔으나, 다양한 활자체 및 크기의 한글 문자 인식에 관한 연구는 별로 연구된 바 없으며, 특히 한글 문자 영상에서의 정보의 분포에 관한 연구 결과는 발표된 바 없다. 따라서 본 연구에서는 한글 문자 영상에서의 정보의 분포를 파악하려는 노력의 일환으로써 12 종류의 서로 다른 활자체 및 크기의 한글 문자 영상에 대하여 정보량(information content)과 엔트로피(entropy)를 추출한 다음, 이들을 한글 문자 인식의 관점에서 분석하였다.

이 논문은 1989년도 문교부 지원 한국학술진흥재단의 신진교수 학술연구조성비에 의하여 연구되었음.

2. 정보량 및 엔트로피 추출 알고리즘

Shannon은 그의 정보 이론(information theory)에 관한 논문[1]에서 정보량과 엔트로피를 다음과 같이 정의하였다.

n개의 사건으로 구성된 다음과 같은 확률 분포가 존재할 때,

$$P_i \geq 0 \quad (i = 1, 2, \cdots, n), \tag{1}$$

$$\sum_{i=1}^{n} P_i = 1 \tag{2}$$

정보량 I와 엔트로피 H는 다음과 같다.

$$I = -\sum_{i=1}^{n} \log_2 P_i \tag{3}$$

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{4}$$

지면 관계상 위의 이론에 대한 상세한 설명 및 해석은 생략하므로, 보다 자세한 내용은 참고 문헌 [1]을 참고하기 바란다.

임의의 문자 영상은 다음과 같이 2차원 좌표계로 표현될 수 있다.

$$f(x, y) = \begin{cases} 0 & \text{(배경의 하얀 화소)} \\ 1 & \text{(문자 부분의 검정 화소)} \end{cases} \tag{5}$$

단, $1 \leq x, y \leq M$.

이 때, 문자의 무게중심 $(\bar{X}, \bar{Y})$는 다음과 같이 정의된다.

$$\begin{cases} \bar{X} = m_{10} / m_{00} \\ \bar{Y} = m_{01} / m_{00} \end{cases} \tag{6}$$

단, 문자의 기하학적 모멘트 $m_{p,q} = \sum_{x=1}^{M} \sum_{y=1}^{M} x^p y^q f(x,y)$.

$(\bar{X}, \bar{Y})$를 좌표계의 원점으로 할 때(즉 $x_0 = \bar{X}$, $y_0 = \bar{Y}$), 문자 영상은 다음과 같이 표현될 수 있다.

$$f'(x, y) = \begin{cases} 0 & \text{(배경의 하얀 화소)} \\ 1 & \text{(문자 부분의 검정 화소)} \end{cases} \tag{7}$$

단, $-m \leq x \leq m$, $-n \leq y \leq n$ 이며, $m = \mathrm{Max}(\bar{X}, M - \bar{X})$, $n = \mathrm{Max}(\bar{Y}, M - \bar{Y})$.

이 때, 한 화소에 대한 검정 화소의 출현 빈도수 C[i, j]는 다음과 같다.

$$C[i, j] = \sum_{k=1}^{N} f_k'(i, j) \tag{8}$$

단, $-m \leq i \leq m$은 i번째 행, $-n \leq j \leq n$은 j번째 열, N은 문자의 총 갯수.

따라서, 검정 화소의 총 갯수 N_c는 다음과 같으며

$$N_c = \sum_{k=1}^{N} \sum_{i=-m}^{m} \sum_{j=-n}^{n} f_k'(i, j) \tag{9}$$

한 화소에 대한 검정 화소의 출현 확률 P[i, j]는 다음과 같이 정의될 수 있다.

$$P[i, j] = C[i, j] / N_c \tag{10}$$

단, $\sum_{i=-m}^{m} \sum_{j=-n}^{n} P[i, j] = 1$

이 때, 문자 영상에서 한 화소에 대한 정보량 I[i, j]와 엔트로피 H[i, j]는 다음과 같다.

$$I[i, j] = -\log_2 P[i, j] \tag{11}$$

$$H[i, j] = P[i, j] \cdot I[i, j] \tag{12}$$

따라서, 문자 영상 전체에 대한 정보량 I와 엔트로피 H는 다음과 같다.

$$I = \sum_{i=-m}^{m} \sum_{j=-n}^{n} I[i, j] \tag{13}$$

$$H = \sum_{i=-m}^{m} \sum_{j=-n}^{n} H[i, j] \tag{14}$$

3. 실험 및 결과 분석

3.1 실험 환경

실험 환경은 SUN 3/60 워크스테이션 상에서 C 언어로 구현하였다. 실험에 사용된 한글 문자의 집합은 한글 기계화 연구소에서 발표한 한글 문자의 잦기 순서[2]의 상위 520 자로 구성하였다. 이 520 자에 대한 누적 사용 빈도율은 약 98% 이상이다[2].

실험에 사용된 한글 활자체는 Qnix 레이저 빔 프린터(QLBP)에서 사용되는 크기가 다른 3 종류의 명조체(mh4, mh5, mh6)와 3 종류의 고딕체(gh4, gh5, gh6)를 포함하여 신문 명조체, 명조체, 고딕체, 공작체, 디나루체, 궁서체 등 12 활자체로써 (즉, 실험에 사용된 한글 문자는 총 6,240 자), 인치당 300 화소의 해상도를 갖는 광학 문자 입력 장치인 Microtek MSF 300GS 스캐너로 입력되었다.

이러한 다양한 활자체 및 크기의 문자 영상들을 처리하기 위해서는, 우선 입력 영상을 표준 크기로 정규화하여야 하는데, S가 크기 변환율(scaling factor)이라고 할 때, $S \leq 1$ 인 경우는 다대일(many-to-one) 사상이 존재하므로 별다른 문제없이 크기 정규화를 수행할 수 있다. 그러나 $S > 1$인 경우는 일대일(one-to-one) 사상이 존재하여 크기 정규화 대상 영상에서의 화소들에 대한 연결성(connectivity)을 보장할 수 없기 때문에 잡음가지(noisy

branch)가 발생하거나 의미있는 부분이 손상될 수 있다.

이러한 문제점을 해결하기 위하여 본 연구에서는 참고문헌 [3]의 크기 정규화 알고리즘을 사용하여 입력 영상을 120x120의 표준 크기로 정규화하였다.

3.2 실험 결과

12 종류의 서로 다른 활자체 및 크기의 한글 문자 영상에 대하여 추출된 정보량과 엔트로피의 통계 결과를 그림 1과 그림 2의 그래프로 표현하였다. 그래프에서 진한 부분은 높은 값을, 연한 부분은 낮은 값을 의미하며 그래프 중앙의 흰 점은 무게중심을 나타낸다. 명조체에 대한 정보량의 분포(그림 1)와 엔트로피의 분포(그림 2.8)를 비교해 보면, 정보량이 높은 부분에서 엔트로피가 낮음을 쉽게 관찰할 수 있다. 표 1은 12 종류의 서로 다른 활자체 및 크기의 문자 집합에 대하여 몇 가지의 통계 결과를 보여준다.

표에서 사용된 몇몇 변수는 다음과 같이 정의되었다.

C_{max} = Max(C[i, j])

P_{max} = Max(P[i, j])

I_{max} = Max(I[i, j])

H_{max} = Max(H[i, j])

실험 결과에 대한 분석을 통하여 다음과 같은 결론을 얻을 수 있었다.

a) 12 종류의 서로 다른 활자체 및 크기의 한글 문자 영상에 대한 정보량의 분포는 매우 유사하였다. 이는 정보량의 분포가 문자의 구조적인 형태에 따라서는 변하지만, 활자체에는 크게 영향을 받지 않음을 의미한다.
b) 그래프로 표현된 정보량과 엔트로피의 분포를 살펴봄으로써 많은 양의 정보가 문자의 가장자리 부분에 위치함을 알 수 있었다. 이는 문자의 가장자리 부분에서 얻어진 특성들[4,5,6]이 한글 문자 영상의 인식 및 축약에 의미있게 사용될 수 있는 근거가 될 수 있다.

4. 결론

본 연구는 한글 문자 영상에서의 정보의 분포를 파악하려는 노력의 일환으로써 다양한 활자체 및 크기의 한글 문자 영상에 대하여 정보량과 엔트로피를 추출하였다. 실험 결과의 분석을 통하여 정보량과 엔트로피의 측정치는 문자의 구조적 형태에 따라 변하지만 활자체에는 무관하며, 대부분의 정보량이 문자의 가장자리 부분에 위치함을 알 수 있었다.

본 연구 결과를 기반으로 한, 다양한 활자체 및 크기의 한글 문서 인식에 관한 연구는 현재 진행 중에 있다[7].

감사의 글

본 연구의 실험을 도와준 충북대학교 전자계산학과 진 상헌 군에게 감사드린다.

표 1. 12 종류의 한글 문자 영상에 대한 통계값

	I	H	N_c	C_{max}	P_{max}	I_{max}	H_{max}
명조체(mh4)	2.106E+05	1.329E+01	2353825	445	1.891E-04	2.117E+01	2.338E-03
명조체(mh5)	2.286E+05	1.336E+01	2021427	406	2.009E-04	2.095E+01	2.467E-03
명조체(mh6)	2.178E+05	1.336E+01	2141310	409	1.910E-04	2.103E+01	2.359E-03
고딕체(gh4)	2.405E+05	1.354E+01	2068243	386	1.866E-04	2.098E+01	2.312E-03
고딕체(gh5)	2.433E+05	1.353E+01	1765476	353	1.999E-04	2.075E+01	2.457E-03
고딕체(gh6)	2.357E+05	1.355E+01	1385245	305	2.202E-04	2.040E+01	2.675E-03
신문 명조체	2.173E+05	1.333E+01	1196036	279	2.333E-04	2.019E+01	2.815E-03
명조체	2.137E+05	1.332E+01	1656050	380	2.295E-04	2.066E+01	2.774E-03
고딕체	2.195E+05	1.345E+01	1730524	345	1.994E-04	2.072E+01	2.451E-03
공작체	2.417E+05	1.357E+01	1757942	301	1.712E-04	2.075E+01	2.142E-03
디나루체	2.349E+05	1.356E+01	2181388	400	1.834E-04	2.106E+01	2.276E-03
궁서체	1.882E+05	1.312E+01	1489724	434	2.913E-04	2.051E+01	3.422E-03

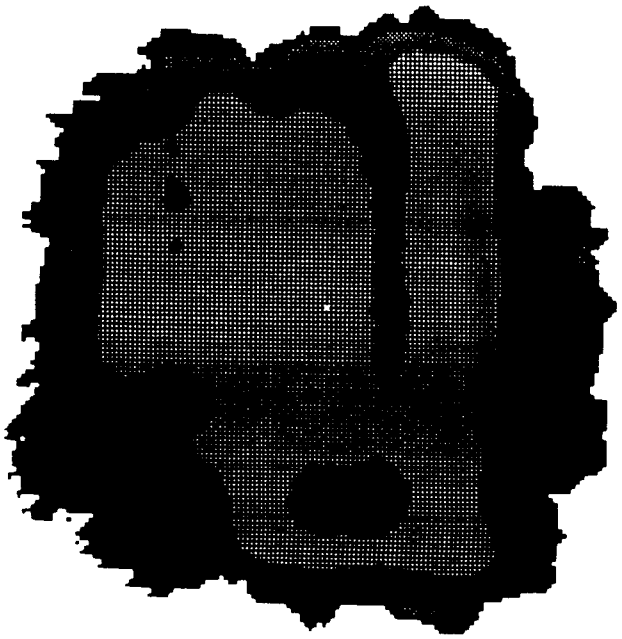

그림 1. 명조체의 정보량 분포

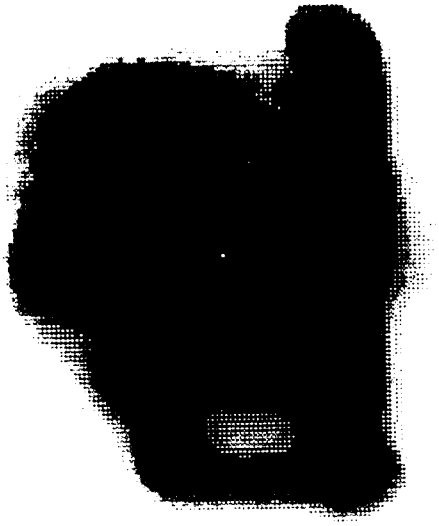

그림 2.1 QLBP mh4의 엔트로피 분포

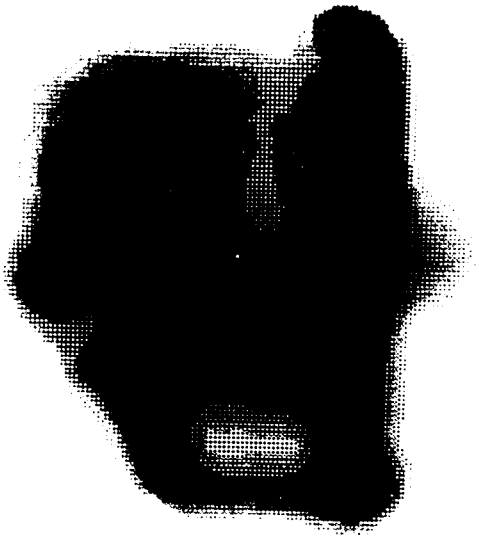

그림 2.2 QLBP mh5의 엔트로피 분포

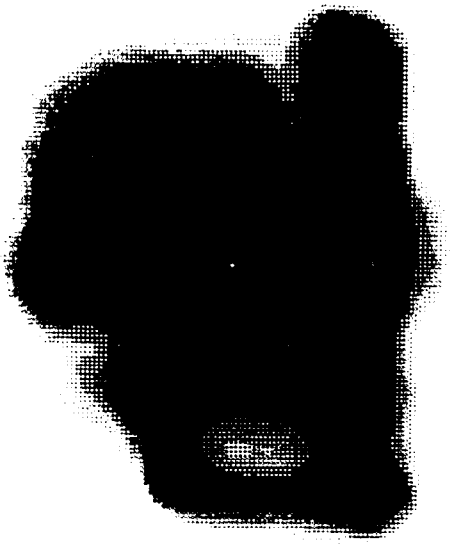

그림 2.3 QLBP mh6의 엔트로피 분포

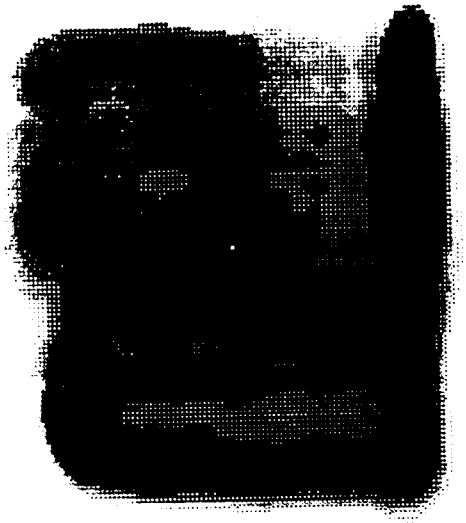

그림 2.4 QLBP gh4의 엔트로피 분포

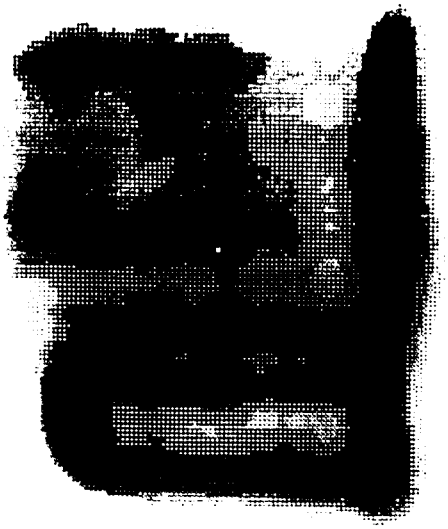

그림 2.5 QLBP gh5의 엔트로피 분포

그림 2.6 QLBP gh6의 엔트로피 분포

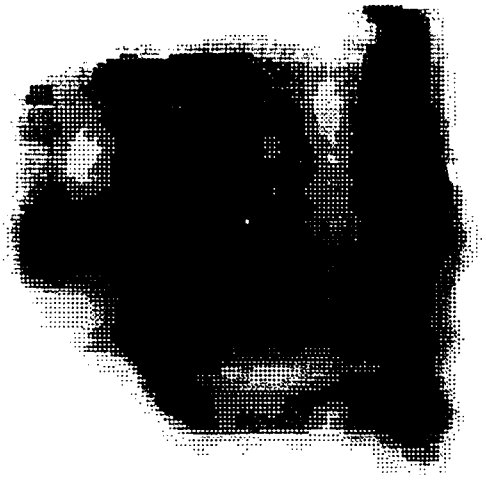

그림 2.7 신문 명조체의 엔트로피 분포

그림 2.8 명조체의 엔트로피 분포

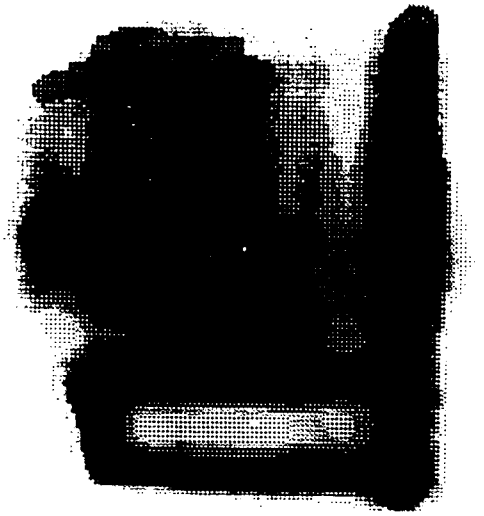

그림 2.9 고딕체의 엔트로피 분포

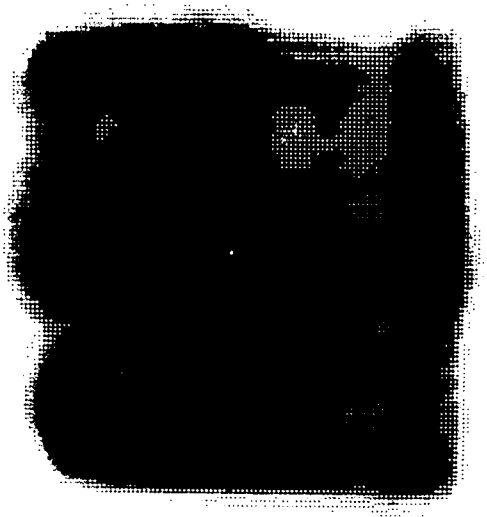

그림 2.10 공작체의 엔트로피 분포

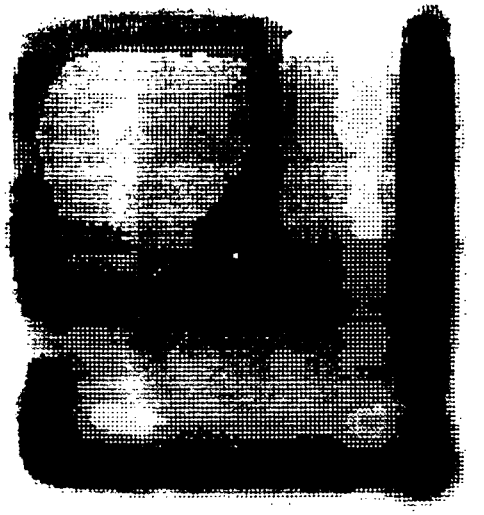

그림 2.11 디나루체의 엔트로피 분포

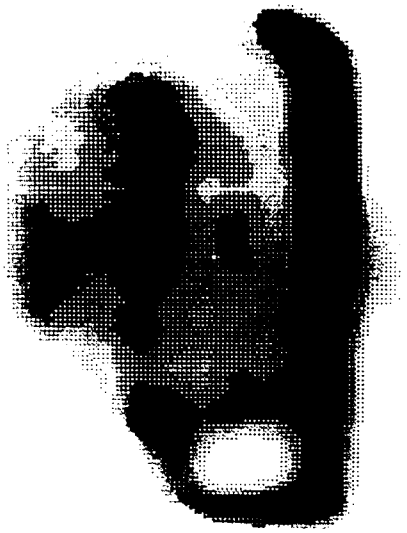

그림 2.12 궁서체의 엔트로피 분포

참고 문헌

1. C.E. Shannon, "A mathematical theory of communications," Bell System Technical Journal, vol. 27, 1948, pp. 379-423, pp. 623-656.

2. 한글 기계화 연구소, 한글 기계화 연구, 1975.

3. H.D. Cheng, Y.Y. Tang and C.Y. Suen, "VLSI architecture for size-orientation-invariant pattern recognition," Pattern Recognition, vol. 23, no. 10, 1990, pp. 1113-1130.

4. K. Yamamoto and S. Mori, "Recognition of handprinted characters by an outermost point method," Pattern Recognition, vol. 12, no. 4, 1980, pp. 229-236.

5. S. Mori, K. Yamamoto and M. Yasuda, "Research on machine recognition of handprinted characters," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 6, no. 4, 1984, pp. 386-405.

6. C.Y. Suen, M. Berthod and S. Mori, "Automatic recognition of handprinted characters - the state of the art," Proceedings of the IEEE, vol. 68, no. 4, 1980, pp. 469-487.

7. 이 성환, "다양한 활자체 및 크기의 한글 문자 인식을 위한 결정 트리의 최적 설계에 관한 연구," 1990. (발표 준비중)