엔트로피 변화를 이용한 문자 영상 데이타의 변형량

김 은 정, 김 대 환, 방 승 양
포항공과대학 전자계산학과

A Variation Measure of Character Image Data Using Entropy Changes

E. J. Kim, D. H. Kim & S. Y. Bang
Dept. of Computer Science, POSTECH

요 약

본 논문에서는 문자 인식을 위해 수집된 문자 영상 데이타들의 변형 정도를 측정하는 변형량의 필요성과 변형량이 가져야 할 조건들을 알아본다. 지금까지 연구된 5가지 변형량들이 이 조건을 모두 만족시키지는 않음을 보이고 이 조건을 만족시키는 새로운 변형량, 평균 엔트로피 변화량을 제안한다. 이 변형량은 여백이나 문자의 두께에 무관하며 같은 문자 뿐만 아니라 다른 문자 간에도 비교할 수 있는 특성을 가진다.

I 서론

문자인식에서 어떤 문자데이타를 선택하느냐는 중요한 문제이다. 인식 실험을 하기 위해서는 한 문자에 대한 영상이 여러개 있어야 한다. 이 영상들의 집합을 데이타라 할때 만일 문자 데이타에 하나의 영상만 존재하거나 여러개의 영상이더라도 같은 영상들이라면 이 데이타에는 변형이 존재하지 않는다. 그러나 이와같은 데이타는 그리 유용하지 않으며 일반적으로 데이타에는 같지 않은 영상들이 여러개 있기 때문에 데이타에는 변형이 존재하게 된다. 또, 변형의 정도는 인식률에 많은 영향을 준다. 그러므로 인식 결과의 객관화와 인식 실험을 위한 문자데이타의 선택의 기준을 제시하기 위해서는 변형을 측정하는 변형량이 필요하다.

문자 데이타의 변형량에 관한 연구는 Shinghal & Suen[2], Hase et al.[3], Yoneda et al.[5] 등에 의해 이루어졌다. 그러나, 변형에 대한 일반적인 정의가 없기 때문에 변형을 어떻게 측정하느냐도 어려운 문제이다. 그리고 변형을 비교하는 것도 같은 문자에 대해서만이 아니라 다른 문자에 대해서도 비교할 수 있어야 하며 변형에 무관한 요소들 즉 문자를 제외한 여백의 양이나, 문자의 두께 등이 다른 데이타에 대해서도 변형의 양만 같으면 같다고 볼 수 있는 변형량이 필요하다. 이러한 연구는 1차원의 일반적인 데이타에 대해서는 이미 연구되어 발표[7]하였으나, 2차원에서의 증명은 미흡하였다. 문자 영상은 기본적으로 2차원이므로 2차원 영상 데이타에 적용할 수 있는 변형량 연구가 필요하여, 본 연구에서의 모든 증

명과 제안은 2차원 영상 데이타를 대상으로 한다. II장에서 이 논문에서 사용되는 기본적인 용어들을 정의하고 변형량의 조건들을 제시하고,III장에서 기존에 연구된 변형량들이 이 조건들을 어느 정도 만족시키고 있는지 알아본다. IV장에서 변형량의 조건을 모두 만족하는 평균 엔트로피 변화량을 제안하고 타당성을 검증한다.

II 문자 데이타의 변형량

먼저 기본적인 용어를 정의하자. 2차원 **영상**(image)의 각각의 원소를 **점**(pixel)이라고 한다. 점은 0 또는 1의 이진 값을 가진다. 영상의 크기가 $X \times Y$이면 점의 갯수 $N = X \times Y$이다. **과**(class)는 영상의 집합(set)으로 정의하며 같은 과에 속하며 같은 갯수의 점을 가진 영상의 집합을 **데이타**(data)라고 한다. **데이타베이스**(database)는 데이타의 집합으로 정의하며 같은 과의 영상이 두개 이상의 데이타에 동시에 속하지 않는다.

필기체 숫자 데이타베이스를 예로 들어 설명하자. 이 데이타베이스는 240명에게서 0부터 9까지의 숫자를 수집하였다. 각 숫자는 100×100의 크기로 컴퓨터에 저장된다. 여기서 영상은 컴퓨터에 저장된 100×100개의 점들이며 영상은 $N = 100 \times 100 = 10000$개의 점을 가진다. 각 점은 이진 값을 가지며 0은 흰 점, 1은 검은 점을 나타낸다. 과는 '0'부터 '9'까지의 숫자이고 데이타는 한 과의 240개의 영상들이므로 데이타 내의 영상의 갯수 $M = 240$이 된다. 이들 10개의 데이타를 묶은 것이 데이타베이스로 데이타베이스에는 10개의 과에 2400개의 영상이 들어있다.

만일 어떤 과에 표준 영상이 존재한다면 하나의 영상만으로도 표준 영상과의 차이를 측정함으로써 변형량을 정의할 수 있다. 그러나 일반적으로 표준 영상은 존재하지 않으며 여기서는 표준 영상이 없는 경우만 다루므로 하나의 영상에 대해서는 변형량이 존재하지 않는다. 그러므로 변형량이 존재하는 가장 작은 단위는 데이타가 된다.

데이타 내의 m번째 영상의 (x,y)번째 점(binary point)의 값은 다음과 같다.

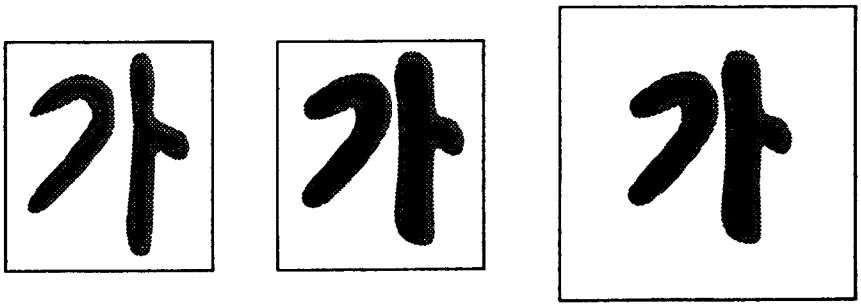
$$b_m(x,y) \in \{0,1\} \tag{1}$$

$\forall x,y, b_{m_1}(x,y) = b_{m_2}(x,y)$ 일때 영상 m_1과 m_2는 **같다**라고 한다.
데이타에서 M개의 영상중에서 (x,y)번째에 검은 점이 나타난 **빈도수**(frequency)는 다음과 같이 정의한다.

$$d(x,y) = \sum_{m=1}^{M} b_m(x,y) \tag{2}$$

$d(x,y)$은 $0 \le d(x,y) \le M$인 실수 값을 가진다. $d(x,y) = 0$인 점을 **여백**이라 하고 $d(x,y) = M$인 점을 **정점**이라 정의한다. 그외 $0 < d(x,y) < M$인 점을 **변형점**이라 정의한다.
M개의 영상중에서 (x,y)번째에 검은 점이 나타날 **확률**(probability)은 다음과 같이 정의한다.

$$p(x,y) = \frac{1}{M} d(x,y) \tag{3}$$



(1) (2) (3)

그림 1: 변형량이 같은 세 문자 데이타의 농도 영상

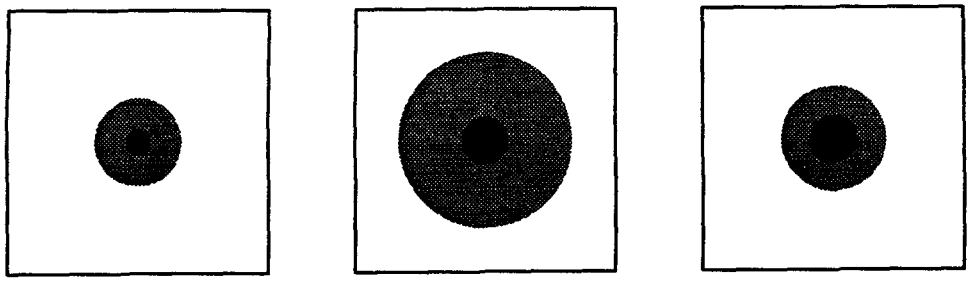
$p(x,y)$의 범위는 $0 \leq p(x,y) \leq 1$ 이다. 데이타의 **복잡도**(complexity)는 $d(x,y) \neq 0$인 (x,y) 점의 갯수로 정의한다.
앞으로 데이타는 M개 각각의 영상의 집합으로 다루지 않고 이를 중첩한 **농도 영상** (density image)하나로 다루겠다. 따라서 데이타의 변형량과 농도 영상의 변형량은 같은 의미로 사용한다.

문자 데이타의 변형을 나타내는 양을 변형량이라 하자. 그러나 문자 데이타에서 무엇이 변형이며, 이 변형을 어떻게 측정하는가하는 문제는 상당히 어려우며 이 의문에 명확하게 해답을 내릴 수 없다. 또 어쩌면 그것은 그 데이타를 쓰는 목적에 따라서 다를 수 있다. 그러나 한 문자에 대하여 수집한 여러 영상들이 정확하게 같지 않다면 그들 사이에 변형은 존재하며, 영상들이 많이 차이 나느냐 적게 차이 나느냐에 따라 변형의 크고 작음이 있다. 따라서 본 논문에서는 변형이라는 것이 존재한다는 가정하에 이 변형을 측정하는 변형량이 어떠한 조건을 만족해야 하는지 알아본다.

첫째 서로 다른 데이타베이스의 같은 문자 데이타에 대하여 비교할 수 있어야 한다. 데이타베이스가 다르다는 의미는 데이타를 구성하는 영상의 크기, 영상의 갯수가 다르다는 것이다. A 데이타베이스의 '가' 데이타가 100×100 크기로 1000개의 영상을 가지고 있고, B 데이타베이스의 '가' 데이타가 256×256 크기로 100개의 영상을 가지고 있다고 할때 변형량이 영상의 크기, 영상의 갯수에 의존한다면 이 두 데이타는 비교할 수 없다. 이를 위해서는 최소한 변형량의 최소값과 최대값을 영상의 크기, 영상의 갯수에 무관한 상수로 놓아 변형량이 가질 수 있는 값의 범위를 일정하게 만들 필요가 있다.

만일 문자 데이타를 구성하는 영상들이 모두 같다면 이 데이타의 변형은 전혀 없다. 즉 데이타를 구성하는 영상들이 모두 같을 때 변형량은 최소값을 가져야 한다. 또 문자 데이타를 구성하는 영상들이 무작위로 들어오는 경우 변형은 최대가 되므로 이때 변형량은 최대값을 가져야 한다.

영상의 갯수가 M개 일때 농도 영상의 각 점의 값을 농도라 하면 농도는 0부터 M까지의 값을 가진다. 그림 1은 문자 '가'에 대한 세개의 데이타의 농도 영상을 보여준다. 농도가 0인 점을 여백이라 하고 그림에서 흰 부분으로 나타난다. 농도가 M인 점을 정점이라 하고 그림에서 검은 부분으로 나타난다. 농도가 0도 아니고 M도 아닌 점을 변형점이라 하고 그



(1) (2) (3)

그림 2: 여러가지 변형의 농도 영상

림에서 회색 부분으로 나타난다. 그림에서 (2)와 (3)을 보면 (3)은 (2)의 여백만 증가시키고 나머지는 그대로이다. 이 두 농도 영상을 비교해 보면 변형의 관점에서 확실히 두 데이타의 변형량은 같아야 함을 알 수 있다. 즉 변형량은 여백의 증가에 무관해야 한다.

또 그림 1의 (1)과 (2)의 농도 영상을 비교해 보면 여백을 제외하면 이 두 농도 영상은 정점의 양만 다르다는 것을 알 수 있다. 같은 사람이 쓰더라도 굵기가 다른 펜으로 문자를 쓸 때 이와 같은 정점의 양만 다른 경우가 생긴다. 그런데 어떤 펜을 쓰는 지에 따라 변형량이 달라진다면 곤란하다. 따라서 문자의 굵기에 무관한 변형량을 원한다면 변형량은 정점의 증가에 무관해야 한다.

변형량이 영상의 크기에 무관해야 한다고 해서 어떤 농도 영상의 변형량과 이 농도 영상을 선형으로 확대시킨 새로운 농도 영상의 변형량이 같아야 한다는 말은 아니다. 그림 2를 보자. (1)의 농도 영상을 두배 선형 확대시킨 농도 영상이 (2)이다. 여기에 (3)의 농도 영상과 비교해 보자. (3)은 (1)에서 정점만 추가한 것으로, 정점을 추가하여 (2)의 정점과 완전하게 같게 만들었다. 즉 (2)와 (3)의 정점이 완전히 일치하므로 문자 데이타의 경우는 동일한 문자가 된다. 그러나 변형점의 분포는 다르다. 같은 문자의 농도 영상이며 변형점의 분포가 다른데도 변형량은 같을 수 없다. 즉 (2)의 변형량이 (3)의 변형량보다 커야 한다. (1)과 (3)은 정점의 양만 다르므로 변형량은 같고 (2)는 (3)보다 변형량이 커야 하므로 (2)는 (1)보다 변형량이 커야 한다. 따라서 농도 영상을 선형으로 확대시켰을 때 변형량은 커져야 한다는 결론을 얻을 수 있다.

다음, 같은 데이타베이스 내에 있는 다른 문자를 비교하려면 어떤 조건이 필요한지 알아보자. 문자가 달라짐으로써 변형량에 미치는 요인 중 주된 요인은 문자의 복잡도이다. 문자 데이타의 복잡도를 농도 영상에서 여백을 제외한 점의 갯수라 한다면 '느', '기' 등은 복잡도가 낮은 문자이며, '빨', '뀁' 등은 복잡도가 큰 문자이다. 문자의 복잡도를 고려하지 않고 변형량을 측정한다면 항상 복잡도가 큰 문자가 변형량이 클 것이다. 따라서 변형량은 복잡도에 의존하지 않아야 다른 문자간에도 비교할 수 있다. 그림 3의 두 농도 영상을 보자. 만일 변형량이 이 두 농도 영상에 대하여 같게 놓을 수 있다면 복잡도에 무관한 변형량이 될 수 있을 것이다. 그림 3의 (1)에서 원 중심을 지나는 세로 단면의 1차원 농도 영상은 (3)과 같다. (2)의 농도 영상에는 (3)과 같은 단면이 여러개 중복되어 있다. 따라서 같은 농도 영상이 여러개 겹쳐져 있더라도 같은 변형이라 볼 수 있으며 복잡도에 무관한 변형량이 될 수 있

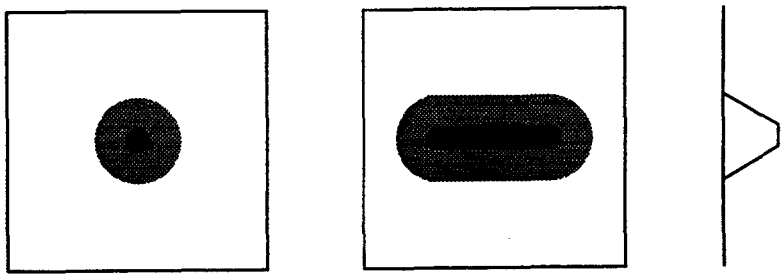


그림 3: 복잡도는 다르지만 변형량은 같은 두 데이타의 농도 영상

다. 문자 데이타의 변형량이 되려면 이와같은 조건들을 만족해야 한다.

III 변형량들의 비교 분석

1 평균 엔트로피 (H)

점 (x,y)의 **엔트로피**는 다음과 같이 정의한다.

$$h(x,y) = -p(x,y)\lg p(x,y) - (1-p(x,y))\lg(1-p(x,y)) \tag{4}$$

여기서 $\lg x = \log_2 x$ 이다. $h(x,y)$의 범위는 $0 \le h(x,y) \le 1$ 이 된다.

평균 엔트로피 (Average Entropy)는 각 점의 엔트로피의 평균으로 정의하며 다음과 같이 나타낼 수 있다.

$$H = \frac{1}{N}\sum_{x=1}^{X}\sum_{y=1}^{Y} h(x,y) \tag{5}$$

평균 엔트로피는 각 점의 엔트로피의 평균이므로 서로 다른 점의 엔트로피는 독립적이라는 가정이 있다.

2 분포 엔트로피 (I)

분포 엔트로피도 엔트로피의 개념을 사용하지만 앞의 평균 엔트로피와는 다른 관점에서 적용하였다.

$$D = \sum_{x=1}^{X}\sum_{y=1}^{Y} d(x,y) \tag{6}$$

$$q(x,y) = \frac{d(x,y)}{D}, \quad (D \neq 0) \tag{7}$$

이때 $0 \le q(x,y) \le 1$이며 $\sum_{x=1}^{X}\sum_{y=1}^{Y} q(x,y) = 1$이다.

분포 엔트로피 (Distribution Entropy)는 다음과 같이 정의한다.

$$I = -\sum_{x=1}^{X}\sum_{y=1}^{Y} q(x,y)\lg q(x,y) \tag{8}$$

분포 엔트로피는 데이타의 농도가 전체적으로 어떻게 분포되는가를 측정한 값이다.

3 변동 엔트로피 (H^A)

$$b = \frac{D}{M} \tag{9}$$

라 할때 **변동 엔트로피** (Variation Entropy)는 다음과 같이 정의한다.

$$H^A = I - \lg b = -\sum_{x=1}^{X}\sum_{y=1}^{Y} \frac{d(x,y)}{D}\lg\frac{d(x,y)}{M} \tag{10}$$

변동 엔트로피는 분포 엔트로피를 기본으로 하며 데이타의 확대 축소에 무관하도록 선형 변환에 불변하게 만들었다.

다음은 변동 엔트로피가 가지는 성질이다.

(1) 동일 영상을 M개 중첩했을 때 H^A는 최소값 0을 가진다.

(2) D가 일정하다면 I가 증가함에 따라 H^A도 증가한다.

(3) H^A의 최대값은 $\lg M$이다.

(4) 영상을 가로 세로로 일정하게 선형 확대 축소하여도 H^A는 변화없다.

4 단위 윤곽선 변동량 (H^L)

L : 평균 주위장, 문자 데이타의 경우 문자의 평균 윤곽선의 길이[5] (11)

단위 윤곽선 변동량 (Unit Length Entropy)은 다음과 같이 정의된다.

$$H^L = \frac{H^A \cdot b}{L} = \frac{1}{L}\cdot\sum_{x=1}^{X}\sum_{y=1}^{Y}\frac{d(x,y)}{M}\lg\frac{d(x,y)}{M} \tag{12}$$

단위 윤곽선 변동량은 윤곽선이 단위 길이당 변형량을 표시하므로 문자의 선 굵기에 무관하다.

다음은 단위 윤곽선 변동량의 성질이다.

(1) 동일 영상을 M개 중첩했을 때 H^L는 최소값 0을 가진다.

(2) M, N, L이 일정하다면 I가 증가함에 따라 H^L도 비례하여 증가한다.

(3) H^L의 최대값은 $-\frac{b}{L}\lg\frac{b}{N}$이다.

(4) 영상을 가로 세로로 k배 선형 확대할 때 H^L은 k배 증가한다.

5 분산 계수 (D_m)

원형(Template)을 다음과 같이 정의한다.

$$T_m = \{(x,y)|p(x,y) > m\} \ (0 \leq m \leq 1) \tag{13}$$

이때 **분산 계수**(Dispersion Factor)를 다음과 같이 정의한다.

$$D_m = \frac{\sum_{(u,v)\notin T_m} p(u,v) \cdot \min_{(x,y)\in T_m}[R(u-x)^2 + (v-y)^2]}{\sum_{(x,y)\in T_m} p(x,y)} \tag{14}$$

이때 $T_m = \phi$이면 D_m은 정의되지 않는다.

여기서 데이타의 크기가 $X \times Y$라 할때 **가중치** $R = \frac{Y}{X}$이다[2].

분모는 정규화 요소로서 데이타의 크기에 어느 정도 무관하도록 만든 것이다. 분자는 원형에 속하지 않는 점에서 원형까지의 거리의 합이며 이 거리의 합을 변형의 척도로 본 것이다. D_m은 T_m의 크기에 상당히 의존하는 값을 갖게 되므로 m은 분산 계수에서 핵심적인 요소이다.

위에서 지면 관계상 자세한 증명은 생략했으나, 기존의 변형량들을 비교 분석한 결과 모든 조건을 만족하는 변형량은 없음을 밝혔다. 이 모두를 만족하는 변형량을 다음 장에서 제안하겠다.

IV 평균 엔트로피 변화량 (V_d)

이 장에서는 새로운 변형량인 평균 엔트로피 변화량을 제안하고 평균 엔트로피 변화량이 어떤 특성을 가지는지 살펴본다.

데이타에 한점만이 존재할 때는 엔트로피가 변형량으로서 충분하다. 그러나 여러개의 점이 존재할 때는 변형량은 단순히 엔트로피로는 부족하므로 다른 요인을 추가해야 한다. 왜냐하면 엔트로피는 점들간이 독립적이라고 가정하므로 엔트로피 자체 만으로는 독립적이지 않은 점들의 변형량을 설명할 수 없기 때문이다.

한 점에 대한 변형을 다음과 같이 나타낼 수 있다.

엔트로피 변화량(Entropy Difference)을 다음과 같이 정의한다.

$$v(x,y) = \frac{h(x,y)}{\alpha \cdot \max\{|h(x+1,y) - h(x,y)|, |h(x,y+1) - h(x,y)|\} + 1} \tag{15}$$

점 (x,y)에서의 엔트로피 변화량은 그 점의 엔트로피에 비례하고 엔트로피 차이에 반비례함을 나타낸다. 분모의 "1"은 분모가 0이 안 되도록 만들어 주기 위한 것이다. 이로 인해 $v(x,y) \leq h(x,y)$이 된다. 분모의 α는 **확장 계수** (multiplying factor)로서 $v(x,y)$의 범위를 결정한다. 따라서 $v(x,y)$의 범위는 $\frac{h(x,y)}{\alpha+1} \leq v(x,y) \leq h(x,y)$가 되며 α가 크질수록 $\frac{h(x,y)}{\alpha+1}$는 0으로 가까와 진다. 그러나 α를 너무 크게 하면 $v(x,y)$이 $h(x,y)$보다는 α에 크게 의존해 버리므로 적당한 값을 선택해야 한다. 본 논문에서는 α를 100으로 두었다.

C를 다음과 같이 놓는다.

$$\mathbf{C} = \{(x,y)|h((x,y)) \neq 0\}, \quad |\mathbf{C}| = \mathbf{C}\text{의 원소 갯수} \tag{16}$$

평균 엔트로피 변화량 (Average Entropy Difference)는 다음과 같이 정의한다.

$$V_d = \begin{cases} \frac{1}{|\mathbf{C}|}\sum_{x=1}^{X}\sum_{y=1}^{Y} v(x,y) & (|\mathbf{C}| \neq 0) \\ 0 & (|\mathbf{C}| = 0) \end{cases} \tag{17}$$

$(x,y) \notin \mathbf{C}$인 점에서는 $v(x,y) = 0$ 이므로 평균 엔트로피 변화량을 다음과 같이 정의할 수도 있다.

$$V_d = \frac{1}{|\mathbf{C}|}\sum_{(x,y)\in\mathbf{C}} v(x,y) \tag{18}$$

여기서 평균 엔트로피 변화량은 모든 변형점들의 엔트로피 변화량의 평균이다.

다음은 평균 엔트로피 변화량의 성질이다.

(1) $|\mathbf{C}| = 0$ 일때, 즉 $\forall(x,y), h(x,y) = 0$ 일때 V_d는 최소값 0을 가진다.

즉 같은 영상의 중첩일때 V_d는 최소값을 가진다.

(2) $\forall(x,y) \in \mathbf{C}, v(x,y) = 1$ 일때 V_d는 최대값 1을 가진다.

$v(x,y) = 1$인 경우는 $h(x,y) = 1$ 그리고 $|h(x+1,y) - h(x,y)| = |h(x,y+1) - h(x,y)| = 0$ 일때이다. 즉 $\forall(x,y) \in \mathbf{C}, p(x,y) = \frac{1}{2}(d(x,y) = \frac{M}{2})$ 일때 이므로 변형점에서 무작위로 분포될 때 변형량은 최대가 된다.

(3) 평균 엔트로피 변화량은 여백과 정점의 증가 감소에 불변이다.

왜냐하면 여백이나 정점이 변하더라도 $\mathbf{C}$는 불변이므로 평균 엔트로피 변화량은 변하지 않기 때문이다.

(4) 데이타의 선형 확대에 평균 엔트로피 변화량은 증가한다.

어떤 데이타의 평균 엔트로피 변화량이 $V_d(XY$개의 점$)$이며 농도 분포가 $\{d(x,y)\}$이다. 다른 데이타의 평균 엔트로피 변화량이 $V_d'(X'Y'$개의 점, $X' = kX$, $Y' = mY)$이며 농도 분포가

$$\begin{aligned}
d'(kx-k+1, my) &= d'(kx-k+2, my) = \cdots \\
&= d'(kx, my) = d(x,y) \\
&\vdots \\
d'(kx-k+1, my-m+2) &= d'(kx-k+2, my-m+2) = \cdots \\
&= d'(kx, my-m+2) = d(x,y) \\
d'(kx-k+1, my-m+1) &= d'(kx-k+2, my-m+1) = \cdots \\
&= d'(kx, my-m+1) = d(x,y)
\end{aligned}$$

이다. 즉 V_d'은 V_d를 가로로 k배, 세로로 m배 선형 확대시킨 평균 엔트로피 변화량이다.

$$V_d' = \frac{1}{C'}\sum_{x=1}^{X'}\sum_{y=1}^{Y'} v'(x,y)$$

$$= \frac{1}{km|C|}\left[\sum\sum_{(x,y)\in C} v(x,y) + (k-1)\sum\sum_{(x,y)\in C} v'_x(x,y)\right.$$
$$\left.+(m-1)\sum\sum_{(x,y)\in C} v'_y(x,y) + (k-1)(m-1)\sum\sum_{(x,y)\in C} h(x,y)\right]$$
$$> \frac{1}{km|C|}\left[\sum\sum_{(x,y)\in C} v(x,y) + (k-1)\sum\sum_{(x,y)\in C} v(x,y)\right.$$
$$\left.+(m-1)\sum\sum_{(x,y)\in C} v(x,y) + (k-1)(m-1)\sum\sum_{(x,y)\in C} v(x,y)\right]$$
$$= V_d$$

이때, $v'_x(x,y)$와 $v'_y(x,y)$는 다음과 같다.

$$v'_x(x,y) = \frac{h(x,y)}{\alpha(h(x,y+1)-h(x,y))+1} \geq v(x,y)$$

$$v'_y(x,y) = \frac{h(x,y)}{\alpha(h(x+1,y)-h(x,y))+1} \geq v(x,y)$$

따라서 $V'_d > V_d$이다.

(5) 데이타의 농도 분포를 중복하여 늘릴때 평균 엔트로피 변화량은 변하지 않는다.

어떤 데이타의 평균 엔트로피 변화량이 V_d(N개의 점)이며 농도 분포가 $\{d(x,y)\}$ 이다. 다른 데이타의 평균 엔트로피 변화량이 V'_d($N'=kN$개의 점)이며 농도 분포가 $\{d'(1,y)=\cdots=d'(1+(k-1)X,y)=d(1,y),\cdots,d'(X,y)=\cdots=d'(kX,y)=d(X,y),\ y=1,\cdots,Y\}$ 이다. 이것은 농도 분포를 k개 만큼 중복한 것이다.

$V'_d = \frac{1}{|C'|}\sum_{(x,y)\in C'} v'(x,y)$
이때 $|C'| = k|C|$, $v'(x,y)=\cdots=v'(x+(k-1)X,y)$ 이므로
$V'_d = \frac{1}{k|C|}(k\sum_{(x,y)\in C} v'(x,y))$
이때 $v'(x,y)=v(x,y)$ 이므로
$V'_d = \frac{1}{k|C|}(k\sum_{(x,y)\in C} v(x,y)) = \frac{1}{|C|}\sum_{(x,y)\in C} v(x,y) = V_d$
따라서 $V'_d = V_d$이다.
즉 데이타의 농도 분포의 중복에 평균 엔트로피 변화량은 변하지 않는다.

V 결론

본 논문에서는 문자 데이타의 변형량이 가져야 할 조건들을 제시하였다. 문자 데이타의 변형량은 영상의 크기, 영상의 갯수에 무관해야 하므로 최소값, 최대값은 상수여야 한다. 최소

값이 될 조건은 데이타 내의 모든 영상이 같은 때 이며 최대값이 될 조건은 모든 영상들이 무작위로 들어올 때이다. 변형량은 여백의 변화에 변화없어야 하며 문자의 굵기에 무관하게 하기 위하여 변형량은 정점의 변화에 변화없어야 한다. 데이타의 선형 확대에 변형량은 커져야 하며, 문자의 복잡도에 무관하게하기 위하여 변형량은 여러 단면의 중복에 변화없어야 한다.

기존의 변형량들이 이와같은 조건들을 모두 만족시키지는 않으므로 이들을 모두 만족시킬 수 있는 평균 엔트로피 변화량을 제안하였다.

참고 문헌

[1] C. Y. Suen, M. Berthod, S. Mori, "Automatic recognition of handprinted characters-The state of the art", *Proc. IEEE*, Vol. 68, pp. 469~487, 1980, 4.

[2] R. Shinghal, C. Y. Suen, "A Method for Selecting Constrained Hand-Printed Character Shapes for Machine Recognition", *IEEE Tran. on PAMI*, Vol. PAMI-4, No. 1, pp. 74~78, 1982, 1.

[3] H. Hase, M. Yoneda, M. Sakai, J. Yoshida, "Evaluation of Handprinting Variation of Characters Using Variation Entropy", 일본 전자정보통신학회논문지 *D*, Vol. J71-D, No. 6, pp. 1048~1056, 1988, 6.

[4] K. Toraichi, R. Mori, I. Sekita, K. Yamamoto, H. Yamada, "Handprinted Chinese Character Database", *Computer Recognition and Human Production of Handwriting*, World Scientific Publ. Co., pp. 131~148, 1989.

[5] M. Yoneda, H. Hase, M. Sakai, "A Consideration on the Evaluation of Character Variation", 일본 전자정보통신학회논문지 *D-II*, Vol. J75-D-II, No. 1, pp. 103~110, 1992, 1.

[6] 이성환, "다양한 활자체 및 크기의 한글 문자 영상에서의 정보량 및 엔트로피의 분포", 한국정보과학회 논문지, Vol. 19, No. 2, pp. 133~139, 1992, 3.

[7] 김대환, 방승양 "필기체 문자 영상 데이타의 변형량의 비교 분석", 한국정보과학회 학술 발표 논문집, pp. 961~964, 1992, 가을.