Quality Evaluation of Character Image Database and Its Application

Hiroyuki Hase
Graduate School of Engineering, University of Fukui
3-9-1 Bunkyo Fukui city, 910-8507, Japan
haseh@u-fukui.ac.jp

***Abstract*--This paper presents a quantitative evaluation method for hand-written character image database. In fact, this research was done in 1992. However, recently, a large amount of character images has been collected from school children, the quality of these character data have been evaluated by this method. In this paper, firstly, some other evaluation methods are introduced, and their drawbacks are pointed out. Then, a method using entropy are introduced. Our method based on entropy is also presented. We call this variation metric "variation entropy". This metric has two kinds of aspects. One is a absolute evaluation of variation, the other is a relative evaluation of variation. The former can be quantified by "variation entropy for a unit boundary length(VEUB)", and the latter can be quantified by "variation entropy for a unit area(VEUA)". The properties of these two metrics are complementary. Lastly, two variation entropies are applied to the standard kanji character database and a database collected from school children.**

Keywords   database evaluation; hand-printed characters

I. INTRODUCTION

Hand-printed characters are different in size, shape and/or position etc. depending on the writer. We call this variation "character variation". When we try to design a character recognizer, a character image database is definitely necessary. Objective evaluation of the recognizer can be possible using the database. However, high quality database tends to brings high accuracy, low quality database brings low accuracy. Therefore, the quality evaluation of the database is absolutely necessary.

Some quality evaluation methods for character image database have ever been proposed. They are classified into three categories. One is a statistical topological analysis such as the number of strokes or holes, aspect ratio, the position of gravity, the number of black pixels, and so on.[1]. This method is heuristic and loses character image information. The second is a method which quantifies the distance or distortion from a standard pattern[2]. However, the selection of the standard pattern is subjective. It loses objectivity. The third is based on information criterion. This is a method which calculates the average amount of information. This can be described by entropy. The problem of this method is how to take the average. If we would make a mistake the way of averaging, the method will be meaningless[3]. For example, consider the following formula.

$$H = -\frac{1}{M}\left\{\sum_{i}^{M}(P_1(x_i)\log P_1(x_i) + P_0(x_i)\log P_0(x_i))\right\} \quad (1)$$

$P_1(x_i)$ is the occurrence ratio of black pixel at x_i, $P_1(x_i) = 1 - P_0(x_i)$. M is the number of total pixels. This formula seems to be appropriate. However, when the image size M is enlarged without changing character size, H will be smaller. Therefore, H is not appropriate for the evaluation of character valuation.

The author proposed the variation entropy based on information criteria[4]. This idea was called as relative evaluation. Furthermore, in another paper, an absolute evaluation method was proposed [5]. They have good characteristics of easy calculation and standard pattern unnecessary.

In fact, this basic research was done in 1992. However, this is the first paper written in English. Also, we try to explain two types of variation entropy in the unified way in this paper. Furthermore, recently, a large amount of character images has been collected from school children, the quality of these character data have been evaluated by our method.

In this paper, the relation between relative or absolute aspect of variation will be discussed from a consistent viewpoint. Next, two metrics are applied to Japanese character databases and another database gathered from school children. Lastly, I will give some suggestions for the usage of two evaluation metrics.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.284

II. VARIATION ENTROPY

Let *N* binary character images that consist of 0/1 be $n^{(\alpha)}(x,y)$• • •1•2,…,*N*• : Now, the center of each image is matched and they are piled up, the following blurred image $f(x,y)$ can be created (Figure 1).

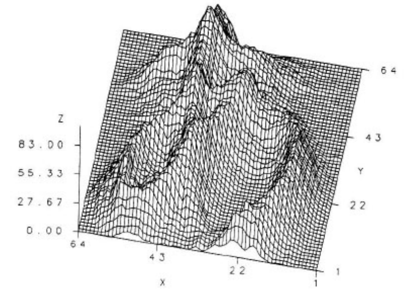
$$\bullet\ \bullet\ \bullet f(x,y)=\sum_{\alpha=1}^{N} n^{(\alpha)}(x,y)\bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\bullet\bullet\bullet(2\bullet$$

This is a Japanese Hiragana character. We can see blurred image due to handwriting valuation. We can calculate the entropy from this 2D distribution:

$$I=-\iint_{Y,X}\frac{f(x,y)}{S}\log\frac{f(x,y)}{S}dxdy \qquad (3)$$

where, $S=\iint_{Y,X} f(x,y)dxdy$, *X*,*Y* are all area of the image. Here, let the average area for one character image be *b*, then $S=b\times N$. Eq.(3) can be rewritten as;

$$I=-\frac{1}{b}\iint_{Y,X}\frac{f(x,y)}{N}\log\frac{f(x,y)}{N}dxdy\bullet+\bullet\log\bullet b$$



(a) ETL8B

Figure 1 Blurred image of Japanese Hiragana character "• •" written by 100 people.

Let the first term be H^A.

$$.\ H^A=-\frac{1}{b}\iint_{Y,X}\frac{f(x,y)}{N}\log\frac{f(x,y)}{N}dxdy \qquad (4)$$

This H^A has the following properties.

A) When all *N* character images are the same. ($f(x,y)=N\bullet/\bullet$), then H^A becomes 0.

B) H^A is getting smaller when the image has smaller blur. The minimum value of H^A is 0.

C) H^A does not change even when all images are scaled up or down with the same rate.

D) When $f(x,y)$ of all pixels are the same, H^A takes the maximal value.

Especially, the property C) is important. This means that H^A doesn't depend on the resolution of the image. This is because a character size depending on resolution is absorbed in the second term *b*. Therefore, the first term is considered as the amount of variation. We called it "variation entropy".

As you see in Eq.(4), This is normalized by the average character area *b*. From this meaning, we call it "variation entropy for a unit area". Now, consider Figure 2. (a) is a blurred circle with diameter 60 pixels which is blurred by the 2D Gaussian function with the standard deviation 4.0. (b) has been scaled down (a) by the half of resolution. (c) is a blurred circle with diameter 30 pixels which is blurred by the same Gaussian function with (a).

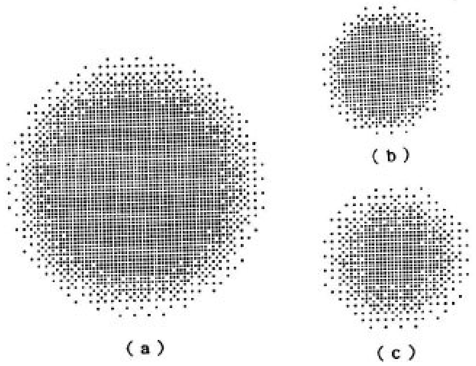


Figure 2 Illustrations of blurred circles

According to the property C), H^A's of (a) and (b) should be the same. The simulation results for these examples are shown in Table 1.

Table 1 Evaluation of blurred circles

	(a)	(b)	(c)
Diameter	60	30	30
Std. of Gaussian	4.0	2.0	4.0
H^A	0.362	0.361	0.743

From Table 1, we can see that the amount of variation (a) and (b) are almost the same. However, someone may think that the amount of variation of (a) and (c) should be the same because they were blurred by the same function. What is the contradiction?

The answer is that there are two kinds of views in variation evaluation. That is, one is that (a) and (b) have the same variation, the other situation is that (a) and (c) are the same.

III. QUANTIFICATIONS OF TWO VIEWS OF VARIATION

In Eq.(4), the numerator

$$-\iint_{Y,X} \frac{f(x,y)}{N} \log \frac{f(x,y)}{N} dxdy \qquad (5)$$

is the total amount of blur of the image $f(x,y)$. Eq.(4) is integrated along x and y. However, to calculate the total amount of blur, it is possible to integrate along the boundary line and along the normal direction to the boundary of the averaged character (see Figure 3).

So, let us integrate it along the boundary *s* and along the normal direction *r* at the boundary *s* of the averaged character. Then Eq.(5) can be rewritten as;

$$-\iint_{Y,X} \frac{f(x,y)}{N} \log \frac{f(x,y)}{N} dxdy$$
$$= -\iint_{S,R(s)} \frac{f'(r,s)}{N} \log \frac{f'(r,s)}{N} drds \qquad (6)$$

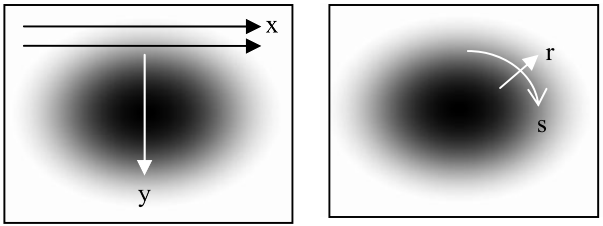


Figure 3 Integration routes to get the total amount of variation

Where $R(s)$ represents the integration range along the perpendicular direction to the tangent at the boundary *s*. Let a small amount of variation for a unit boundary length at s be $h_L(s)ds$, $h_L(s)$ can be described as Eq.(7).

$$h_L(s) \equiv -\int_{R(s)} \frac{f'(r,s)}{N} \log \frac{f'(r,s)}{N} dr \qquad (7)$$

Now, letting the average length of boundary be *L,* and letting $\underset{s}{E}[h_L(s)]$ be H^L, The total amount of variation can be described as follows..

$$\int_S h_L(s)ds = H^L \times L$$

Where, we call H^L "variation entropy for a unit boundary length". H^L and H^A are related by Eq.(8).

$$H^A \times b = H^L \times L \qquad (8)$$

Therefore, H^L can be easily calculated by Eq.(8).

Calculating H^L using simple patterns in Figure 2, the result is shown in Table 2.

Here, "variation entropy for a unit boundary length" satisfies the following properties.

Table 2 Two kinds of views for variation

	(a)	(b)	(c)
Diameter	60	30	30
Std. of Gaussian	4.0	2.0	4.0
H^A	0.362	0.361	0.743
H^L	5.423	2.706	5.572

E) When all *N* character images are the same, H^L becomes 0.

F) Increasing blur as keeping the number of black pixels, H^L is getting bigger.

G) When the image size is changed by *k* times in both horizontal and vertical directions, H^L becomes • times.

H) When $f(x,y)$ of all pixels are the same, H^L becomes the maximum.

IV. APPLICARION TO CHARACTER DATABASE

A. For the Standard Database of Japanese Characters

ETL8B and ETL9B are famous databases of hand-printed Japanese characters. These were made in 1980's by the National Institute of Electro-Technical Laboratory to develop OCR technology. ETL8B includes 956 categories of Japanese characters and 152,960 samples, each character is written in the frame of 10×10(mm) and is digitized by 64×63 pixels[6]. ETL9B includes 3036 categories of characters and 607,200 samples, each character is written in the frame 8×9(mm), and is digitized by 64×63 pixels[7].

Firstly, the calculation of variation entropy for a unit area H^A and a recognition test for Hiragana 71 categories are carried out after character normalization in position, size, or shape. This is because we have a prediction that a database with smaller variation will show a better recognition rate. The standard pattern was made from half of character samples for each category as learning. The other half samples were used as test sample by the similarity method. In this experiment, four types of normalization such as (1) the center of image is matched, (2) gravity is matched, (3) normalization in size of 64×64 pixels and (4)stroke density equalization[8] were applied. The experimental results are shown in Table 3.

As you see in Table 3, the recognition rate tends to go up when the variation entropy becomes small in the both databases. Therefore it was shown that the variation entropy for a unit area indicates the quality of the database.

Table 3 Comparison between variation entropy and recognition rate

(a) ETL8B

Normali-zation method	ETL8B H^A	Recognition rate (%)		
		Test data	Learning data	Total
Center	1.885	73.07	81.30	77.18
Gravity	1.823	79.27	85.21	82.24
Image size	1.632	80.31	85.92	83.11
Line density	1.471	81.58	87.01	84.30

(b) ETL9B

Nomali-zation method	ETL9B H^A	Recognition rate (%)		
		Test data	Learning data	Total
Center	1.716	67.47	76.23	71.85
Gravity	1.528	77.80	83.47	80.63
Image size	1.351	76.93	83.16	80.04
Line density	1.202	80.48	86.20	83.34

However, comparing the results for ETL8B and ETL9B, the variation entropy H^A of ETL8B is bigger than that of ETL9B though the recognition rate of ETL8B is better than that of ETL9B. This is a contradiction against our prediction. This reason seems to come from the difference of data collection. That is, a character is written in the frame of 10×10(mm) in ETL8B, a character is written in the frame of 8×9(mm) in ETL9B. Therefore, the average character area of ETL9B was relatively bigger than that of ETL8B. In fact, the average character width of ETL8B was 3.435 and 4.544 for ETL9B. It is thought that this fact caused this contradiction. When the variation entropy for a unit boundary length is calculated. The results are shown in Table 4

From Table 4, we can see that the variation of ETL9B is bigger than that of ETL8B. That is, when applying variation entropy for a unit boundary length will be appropriate in the case of different condition in data collection.

Table 4 Variation entropy for a unit boundary length

Normalization method	H^L of ETL8B	H^L of ETL9B
Center	3.334	3.577
Gravity	3.237	3.504
Image size	3.337	3.855
Line density	3.111	3.383

B. For Database by School children

According to these considerations, we applied the variation entropy for characters written by children. A large amount of hand-printed characters were collected from three primary schools and two junior high schools in Japan to evaluate the quality difference for the grade. Japanese primary school has 6 grades and junior high school has 3 grades. The character categories are 71 HIRAGANA and 71 KATAKANA characters, each type of data was written in different paper(in Figure 4). Each character was written within 1cm×1cm area. The total number of papers is 3547, and the total number of characters is 251,837.

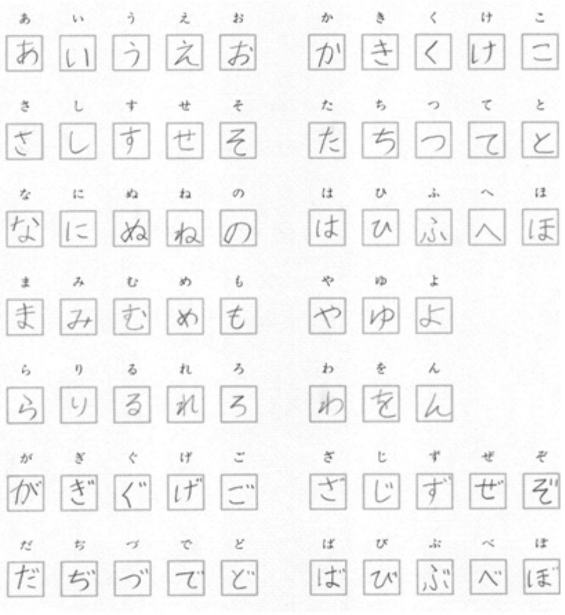

あ い う え お か き く け こ
さ し す せ そ た ち つ て と
な に ぬ ね の は ひ ふ へ ほ
ま み む め も や ゆ よ
ら り る れ ろ わ を ん
が ぎ ぐ げ ご ざ じ ず ぜ ぞ
だ ぢ づ で ど ば び ぶ べ ぼ

Figure 4 Example of character data(HIRAGANA)

All characters were cut out from the data sheet, and collected for each grade and each category. In this case of data collection, all data were written under the same condition so that the variation entropy for a unit area (VEUA) was used for this analysis. VEUA was calculated using 100 character data for every categories, every grades. When the character image was piled up, the gravity was matched. Figure 5 is an evaluation result for each grade. From 1 to 6 on horizontal axis are the grades of primary school, and J1 to J3 are the grades of junior high school. A data point is the average of all HIRAGANA, KATAKANA categories for a grade. The vertical bar is the standard deviation. The dotted line is the regression line.

As you can see, the VEUA is decreasing for upper grades. This means that the writing skill of upper grade students becomes good.

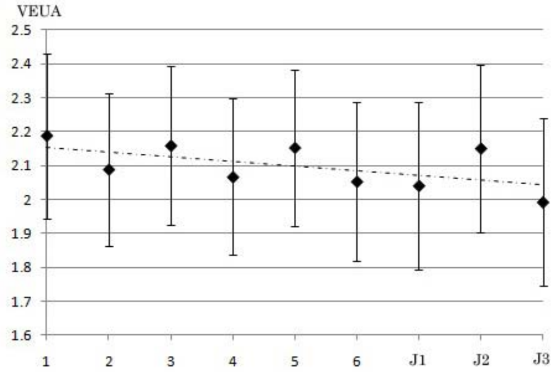


Figure 5 Analysis result

V. Summary

This paper presented that there were two points of views in variation evaluation of character database. One is a variation evaluation for a unit area (VEUA) which is thought of as a relative evaluation. This metric H^A does not depend on the resolution. The other is a variation evaluation for a unit boundary length (VEUB) which is thought of as an absolute evaluation. This metric H^L is proportional to the resolution. The former evaluates that the variations of (a) and (b) in Figure 2 are the same, the latter evaluates that the variations of (a) and (c) are the same. However, there is no situation that (b) and (c) have the same variation. Next, the both metrics were applied to two standard Japanese character databases. From the results, it became obvious that there exists strong relation between the variation entropy and the recognition rate. From this meaning, we will be able to say that the variation entropy is reflected the quality of character image database. Although these results was presented in two papers individually [4,5], this paper tried to explain two types of variation entropy in the unified way.

The usage of two types of variation entropies depend on the condition of data collection. That is, we suggest, when character data collected under the different condition are evaluated, the variation entropy for a unit boundary length(VEUB) will be appropriate, on the other hand, when character data collected under the same condition are evaluated, the variation entropy for a unit area(VEUA) will be appropriate.

Furthermore, a large amount of data written by school children was evaluated. From the result, we found that the writing skill of upper grade students is getting better.

This basic research was done in 1992. However, this is the first paper written in English. And the analysis result for school children characters is the first report in this paper.

Recently, some similar metrics applied this method are presented [9].

ACKNOWLEDGEMENT

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (C) (23500116).

REFERENCES

[1] T.Saito etc.; Analysis of handwritten character database, Technical report of Electro-technical laboratory, 42,5, 385-434, 1978.(in Japanese).

[2] .R.Shinghal and C.Y.Suen; A method for selecting constrained handprinted character shapes for machine recognition, IEEE Trans. PAMI-4,1,74-78, 1982.

[3] S.Mori, H.Yamada, T.Saito and T.Miyagawa, Some variation analyses of handprinted characters, IPSJ, 18,8,814-821, 1977.(in Japanese)

[4] H.Hase,M.Yoneda,, M Sakai and J.Yoshida; Evaluation of handprinting variation of characters using variation entropy, Trans. of IEICE. J71-D, 6,1048-1056, 1988.(in Japanese)

[5] M.Yoneda,H.Hase,M.Sakai; A consideration on the evaluation of character variation, Trans. of IEICE, D-II, J75-D-II,1,103-110, 1992.(in Japanese)

[6] T.Saito,H.Yamada,K.Yamamoto; Technical Report of National Institute of Electro-technical laboratory, 45,1&2,49-77,1981.

[7] T.Saito,H.Yamada and K.Yamamoto, On the database ETL9B of handprinted characters in JIS Chinese characters and its analysis, Trans.of IEICE, J68-D,4,757-764,1985.(in Japanese).

[8] H.Yamada, K.Yamamoto and T.Saito, A nonlinear normalization method for handprinted kanji character recognition -line density equalization-, Pattern Recognition, 23, 9, 1023-1029,1990.

[9] F.Tajima; Effect of gradation processing method on evaluation of hand-written character variation with directional line-element entropy, Trans of IEICE, D-II,J84-D,9, 2167-2172,2001.(in Japanese)