A Method for Two-valuing of Printed Images

Issei Yamasaki*

Abstract

This paper deals with a method for setting of the threshold value when printed images are transformed into two-valued black and white patterns. A 'model' stroke which has gray levels is constructed, considering the allowable limit of strokewidth variations. It is theoretically clarified that the threshold value is the average of the values of black and white portions. The article includes discussions on the reliability of bases for the threshold value.

1. Introduction

Printed Images are usually processed as two-valued black and white patterns in optical character recognition. Since printed Images are essentially black and white patterns, and two-valued patterns can be easily processed. Several investigations in the field of threshold value setting have been carried out [1].

Scanned signals have gray levels. A threshold value to transform non black and white patterns into two-valued patterns, has been determined based on expriences. The median of the brightness of black portion and that of white part is adequate to the threshold value. Whether the threshold value should be true average or should be an average biased to black or to white, have not been clarified.

This paper investigates into a method for threshold value setting. First, a "model" stroke which has gray levels is introduced. Second, it is studied that the method to transform the model stroke into a two-valued pattern. Third and finally, we discuss the reliability of the bases for the threshold value.

2. A Model of a Blurred Stroke

This paper first appeared in Japanese in Joho-Shori (Journal of the Information Processing Society of Japan), Vol. 16, No. 5 (1975), pp. 419~425.

* Information Science Division, The Electrotechnical Laboratory

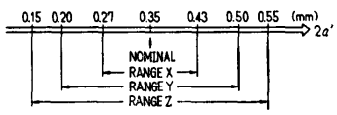

fig. 1 Three ranges of strokewidth variations.

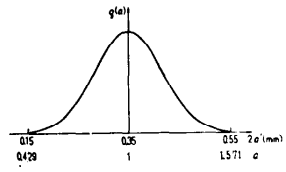

fig. 2 The probability that stroke with width of a' appears.

A stroke $h(x)$ with normalized width of $2a$ centered at the origin of a spacial coordinate x is written as:

$$h(x) = \begin{cases} 1 , & |x| < a \\ 0 , & |x| > a \end{cases} ; \qquad (1)$$

where the spacial coordinate x is normalized with the nominal strokewidth ($2a_0' = 0.35$ mm) [2], and black portion is 1 and white portion is 0.

The stroke $h_0(x)$ with nominal width ($a = 1$) is

$$h_0(x) = \begin{cases} 1 , & |x| < 1 \\ 0 , & |x| > 1 \end{cases} . \qquad (2)$$

Allowable variations of strokewidth in optical character recognition are shown in fig. 1.

We consider that the probability that strokewidth of $2a'$ will appear obeys the normal distribution as shown in fig. 2. Almost all (99.73%) of strokewidths appeared are in the range Z. Very few strokes appear outside of the range Z. The probability density $g(a)$ of the stroke where normalized width of $2a$ will appear, is written as:

$$g(a) = (1/(\sqrt{2\pi}\,\rho)) \exp[-(a-a_0)^2/(2\rho^2)] ; \qquad (3)$$

where $a_0 = 1$, and $\rho = 0.190$.

Now, we study a model of stroke. We consider that superimposed strokes which appear according to eq. (3) is an actual stroke; namely the actual stroke is expressed as a probability of black will happen at a point x. A model stroke $g(x)$, then, is written as:

$$g(x) = \int_{|x|}^{\infty} g(a)\,da = \frac{1}{\sqrt{2\pi}\rho} \int_{|x|}^{\infty} e^{-\frac{(a-a_0)^2}{2\rho^2}}\,da ; \qquad (4)$$

where strokes are expressed as eq. (1). It is easily proved that the model stroke $g(x)$ is an even function, and is symmetric with respect to the point (1, 0.5) in the region of $0 < x < 2$. Fig. 3 shows an outline drawing of the model stroke $g(x)$.

3. A Method of Setting Threshold Value

3.1 Correspondence between the Model Stroke and a Two-Valued Stroke

We determine a width $2a$ of a stroke $h(x)$ which resembles the model stroke $g(x)$ in the sense of the similarity. Table 1 shows the similarity between the balck and white stroke $h(x)$ with didth of $2a$ and the model stroke $g(x)$ calculated for several values of a.

The model stroke is similar to a two-valued stroke with width of $2a = 2.04$:

$$a = 1.02 \quad .$$

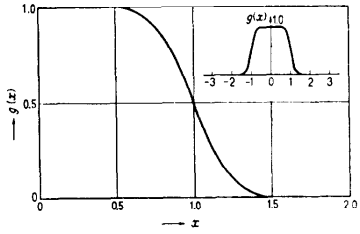

fig. 3 A "model" stroke $g(x)$

table 1 Similarity Between the Model Stroke and a Rectangular Type Stroke

STROKEWIDTH	SIMILARITY
0.97	0.97598992
0.98	0.97690250
0.99	0.97760907
1.00	0.97811148
1.01	0.97841217
1.02	0.97851408
1.03	0.97842070
1.04	0.97813601
1.05	0.97766449

3.2 Threshold Value Setting

We transform the model stroke $g(x)$ into a two-valued pattern by threshold value of $1/k$ times the maximum value $g(0)$ of the model stroke. The character k is termed a threshold coefficient.

In order to transform the model stroke $g(x)$ into the black and white stroke with width of $2a = 2.04$, the value of the threshold coefficient should be

$$k = 2.18301 \quad .$$

Because $g(1.02) = 0.458084$, and $g(0.00) = 1$ (cf. table 2).

It is desiarable in practical usage that the threshold coefficient is simple.

Now, we examine errors of the strokewidth and the similarity, when the threshold coefficient k is chosen as $k = 2.00$. The width of two-valued stroke, in this case, becomes unity: $X = 1$, from table 2.

Therefore, the relative error to the theoretical value ($X = 1.02$) is:

$$(1.00 - 1.02) \times 100/1.02 = -2 \ (\%)$$

The value of the similarity between the model stroke and the two-valued stroke

table 2 Variation of the Model Stroke $g(x)$ in the Neighbourhood of $x = 1.00$

X	$g(x)$
0.97	0.562730
0.98	0.541917
0.99	0.520987
1.00	0.500000
1.01	0.479013
1.02	0.458084
1.03	0.437270
1.04	0.416628
1.05	0.396214
1.06	0.376081
1.07	0.356280

which is derived from theoretical threshold ($k = 2.1831$), is 0.97851408. On the other hand, the value of the similarity between the model stroke and the two-valued stroke which is derived from simple threshold value ($k = 2.00$), is 0.97811148. The error becomes

$$(0.97851408 - 0.97811148) \times 100/0.97811148 = -0.04\ (\%)\ .$$

There is little difference whether the threshold coefficient is determined theoretically or not. Therefore, it may be reasonably concluded that the threshold coefficient k is able to select as:

$$k = 2.00\ .$$

4. Reliability of the Bases for Threshold Value Setting

The method of setting threshold value described above uses the minimum and the maximum reflected value found in the area where a character is printed. It should be confirmed experimentally whether these minimum and maximum values are exceptional or not.

We investigate the stability of these values by using actual data. Actual data are OCR-A numerals printed with a line printer equipped with new inked ribbon of 2 meters long. Printed sheets are sampled every 5,000 lines. Sampled sheets are scanned by a CRT flying spot scanner, and video signal is sampled at every intervals of 0.06 and 0.12 mm in horizontal and vertical directions, respectively. The number of sampling points of a character is 1365 (= 39 × 35). The levels of sampled signal are 64 (= 2^6). The total number of data is 140 (= 10 × 2 × 7).

Fig. 4 shows a frequency distribution of reflected values found in the area where a character is printed. This frequency distribution curve has two peaks. One corresponds to the reflection of black portion and the other corresponds to reflectance of paper white.

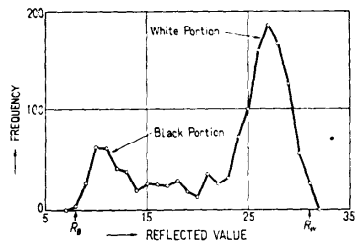

fig. 4 A frequency distribution of reflected value in the domain where a character pattern appears.

Now, we study the reflection by using the print contrast signal (PCS) [4].

Print contrast signal PCS_p at a point p is defined as follows:

$$PCS_p \quad (R_w - R_p)/R_w$$

where

R_w is the maximum reflected value found in the area where a character is printed;

R_p is the reflected value at point p.

PCS takes 0 through 1; 0 for the brightest point, and 1 for the darkest (true black) point.

We investigate the relation between the values of PCS_{peak} and PCS_{avg}.

PCS_{peak} is defined as follows:

$$PCS_{peak} = (R_w - R_B)/R_w ;$$

where

R_B is the minimum reflected value found in the area where a character is printed. PCS_{avg} is defined as the arithmetic average value of PCS measured over the darkest 80% of the stroke within the minimum character outline limit [4].

Variations of PCS_{avg} and PCS_{peak} due to printed lines are shown in fig. 5.

There is a positive correlation between PCS_{avg} and PCS_{peak}. The correlation coefficient between PCS_{peak} and PCS_{avg} is 0.933. The correlation coefficient between averaged PCS_{peak} and averaged PCS_{avg} is 0.983. It may be concluded that PCS_{peak} is linearly dependent to PCS_{avg}. The reliability of PCS_{peak}, therefore, is nearly the same as that of PCS_{avg}.

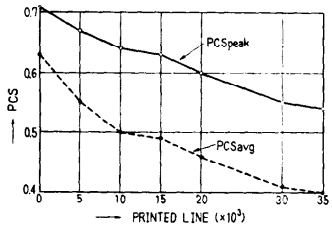

fig. 5 Variation of PCS_{peak} and PCS_{avg} due to printed line.

From the fact described above, we may conclude that the minimum and the maximum reflected value found in the area where a character is printed, can be used as bases for threshold value.

5. Conclusions

It has been clarified that the threshold value is the average of the minimum and the maximum reflected values found in the area where a character is printed. The threshold value derived coincides with the threshold value based on experiences.

It is confirmed that the minimum and the maximum reflected value, which are used as bases for threshold value, are sufficiently reliable.

Proposed threshold value can be used for print quality evaluation [5] and for inexpensive OCR.

References

[1] for example Bartz, M. R.: "The IBM 1975 Optical Page Reader Part II: Video Thresholding System", IBM J. Res. and Develop., 12, 5, pp. 354 - 363 (Sep. 1968)

[2] JIS C 6250 - 1970: "Alphanumeric Character Sets for Optical Character Recognition" (June 1970); equivalent to ISO/R 1073 (May 1969)

[3] Yamasaki, I. and Iijima, T.: "On Character Image Sampling", Trans. IECE (C), 51-C, 9, pp. 428 - 429 (Sep. 1968); available in English in ECJ, same date, pp. 146 - 147

[4] ANSI: "Proposed American National Standard Character Set and Print Quality for Optical Character Recognition" (Sep. 1970)

[5] Yamasaki, I. and Iijima, T.: "Print Quality Evaluation of a Large Number of Data", Jour. IPS. Japan, 13, 8, pp. 525 - 532 (Aug. 1972); available in English in Info. Proc. in Japan 13, pp. 7 - 12 (1973)