A Quantitative Representation of Quality for a Set of Printed Characters

Issei YAMASAKI*

1. Introduction

Optical character recognition (OCR) systems have been established for several years as an effective and reliable means of data-entry for computer systems. Specifications of print quality for OCR systems have been discussed at the International Organization for Standardization (ISO) for more than ten years [1]. It is unable to say, however, that there exists a unified and established measurement method of print quality.

Conventional measurements concerning print quality are as follows:

i) Judging good or bad by magnifying a character and comparing it with the measuring gauge showing the minimum and maximum character outline limits (COL).

ii) Measuring the density of inking.

The measurement ii) can be performed by devices (for example, reflectometers); therefore, we can obtain objective data by inspecting a large number of printed images. The measurement i), on the other hand, should be performed by an examiner, so a lot of characters can not be inspected, and, also, the measured values are poor in reproducibility. Nevertheless, the measurement i) is useful and has been employed in practical applications, when we look into the cause of poor reading of an OCR system.

The author and T. Iijima have proposed a method for print quality evaluation of a large number of printed characters [2] and have shown that this method can be applied in practical usage [3]. In this paper, we discuss two subjects. First, in what way we statistically process evaluation values, and how to obtain a statistical figure for a set of printed characters. Second, we study a method for expressing the print quality of a set of printed images in the form of one numerical quantity by integration of four kinds of statistical figures.

This paper first appeared in Japanese in Joho—Shori (Journal of the Information Processing Society of Japan), Vol. 18, No. 3 (1977), pp. 253~256.

* Information Science Division, The Electrotechnical Laboratory

2. Print Quality Parameters

Evaluation values computed for each input character are the following:

1) The peak value of the print contrast signal (PCS), P, which is defined by

$$P \equiv (R_{max} - R_{min})/R_{max} ,$$

where R_{min} and R_{max} denote the minimum and maximum reflectances found within the region R were a character pattern appears, respectively. The quantity P represents the density of inking relative to the background of paper, and takes the values zero to unity.

2) The average stroke-width W, which is defined by

$$W \equiv M/M_0 ,$$

where M and M_0 denote the zeroth moments of a two-valued input pattern and its standard pattern with the nominal stroke-width, respectively. The quantity W corresponds to an average stroke-width.

3) The noise factor n, which is defined by

$$n \equiv 1 - s^2 ,$$

where the symbol s denotes the similarity between an input pattern g and its standard pattern g_0. The similarity s is defined as

$$s \equiv (g,g_0)/[||g_0||\cdot||g||] ,$$

where $(\cdot,\cdot)$ and $||\cdot||$ denote the inner product and the norm, respectively. It is assumed that these two patterns have an equal zeroth moment and are transformed into normalized forms [2]. The quantity n represents a degree of deformation, and takes the values zero to unity.

4) The centroid deviation $\mathbf{d}$, which is defined by

$$\mathbf{d} \equiv \mathbf{D} - \mathbf{D}_0 ,$$

where $\mathbf{D}$ and $\mathbf{D}_0$ denote the coordinates of centroids of an input pattern and its standard pattern, respectively, when the input and its standard patterns overlap at a best fit position. The quantity $\mathbf{d}$ represents the positional deviation of the centroid of the input pattern measured from the centroid of the standard pattern as a basis, when the input and standard patterns are at the best fit position. The centroid deviation $\mathbf{d}$ is converted into a centroid distance d for statistical analysis. The centroid distance d is defined by

$$d \equiv ||\mathbf{d}|| .$$

3. Distributions of Evaluation Values

The documents for testing are scanned by a CRT flying spot scanner, and the character on the documents are recorded on magnetic tape. The scanned data recorded on magnetic tape are evaluated by a program on a large-scale digital computer. Printing conditions of the documents are shown in Table 1.

Frequency distributions of four kinds of evaluation values are shown in Fig. 1. Fig. 1 illustrates the distributions obtained in three grades of ribbon life; upper, middle and lower parts of the figure show examples of new, medium and considerably exhausted ribbons, respectively. These distributions do not follow the normal distribution, except the peak value of PCS. Primitive statistical values——the mean, the minimum, the maximum and the standard deviation——are not suitable for the numerical expression of the population which does not follow the Gaussian distribution. It is also complicated to enumerate these four kinds of statistical figures.

Table 1 Data Used for Simulation

Font	OCR-A numeral "0"
Printing Unit	Lineprinter
Ribbon	Silk, 13 meters long
Paper	OCR Paper
Sampling	Four lines are sampled at approximately every 10000 lines from a printed pile of 142000 lines. Fifteen numeral of "0" are scanned in a line.
Total No.	900 (= 15 × 4 × 15)

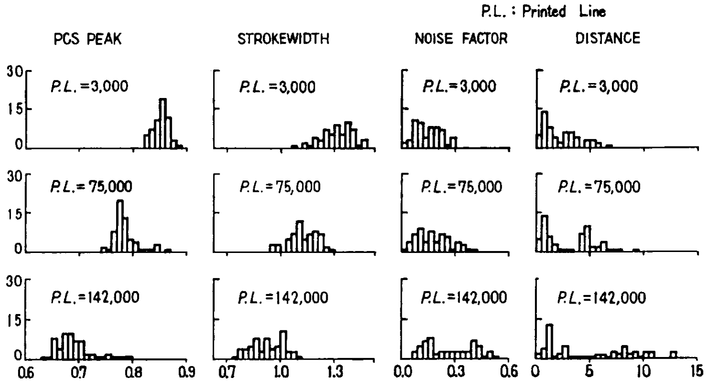

Fig. 1 Distribution of evaluation values.

4. Representation of Print Quality for a Set of Printed Characters

Now, we introduce a quantity which is considered to be suitable for a statistical figure of a set of printed characters. The quantity Z termed "representative measurement" is defined by

$$Z \equiv \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} ,$$

where $\{x_i\}$ denotes input data, and N is the total number of data. The representative measurement Z is defined as the square root of the quantity of the second moment around the origin divided by the total number. The following relation exists among the representative measurement Z, the mean M, and the standard deviation V:

$$Z^2 = M^2 + V^2 (1 - 1/N) .$$

The representative measurement Z is the quantity combining the mean and the standard deviation of input data. The representative measurement Z is nearly equal to the square root of the squared sum of the mean and the standard deviation, if the total number of data is large. If the distribution of the input data follows the Gaussian distribution, and if the standard deviation takes a small value, then the representative measurement Z takes the value very close to the mean.

Fig. 2 shows the variation of the representative measurement of each evaluation value obtained in the foregoing simulation in Section 3. We see an abnormal result

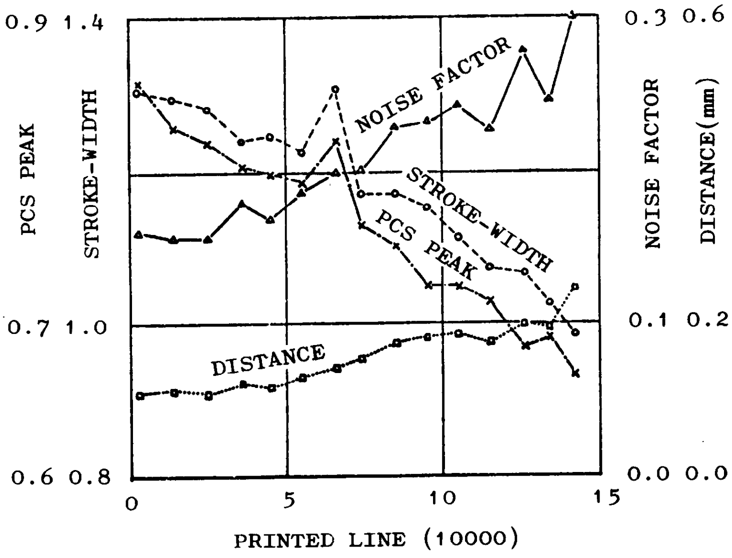

Fig. 2 Results of evaluation for a numeral of "0" (OCR-A).

at the printed line 6600 in this figure. This abnormal result would be caused by the malfunction of the reversing mechanism of the ribbon. The documents printed at about 6600 lines actually have heavily printed characters.

5. Correlations Among Values of Representative Measurements

It is observed from Fig. 2 that four kinds of representative measurements correlate with each other. Correlation coefficients among these values of representative measurements are shown in Table 2. These four kinds of parameters are linearly dependent upon each other. We may draw in the statistical point of view the following conclusions from the facts described above: If a document is printed heavily, print quality of the document is high grade; on the other hand, if a document is printed lightly, print quality of the document is low grade. Therefore, it may be possible to express the print quality of a set of printed characters by one kind of parameter.

Incidentally, correlation coefficients among evaluation values of each input pattern are shown in Table 3. We see from this table the following: Although the noise factor and the centroid deviation are linearly dependent upon each other, the correlations among the other evaluation values are small. It can be said that the four kinds of evaluation values obtained by processing each input character; therefore, are meaningful as quantities representing print quality of an input character.

Table 2 Correlation Coefficients Among Values of Representative Measurements

	PCS Peak	Stroke-Width	Noise Factor
Stroke-Width	0.977		
Noise Factor	-0.913	-0.943	
Distance	-0.928	-0.946	0.985

Table 3 Correlation Coefficients Among Evaluation Values

	PCS Peak	Stroke-Width	Noise Factor
Stroke-Width	0.726		
Noise Factor	-0.194	-0.678	
Distance	-0.144	-0.674	0.905

6. Conclusion

A statistical method to integrate four kinds of evaluation values in the form of one numerical quantity have been studied. The representative measurement which is obtained from the second moment around the origin is proposed from the fact that the distributions of evaluation values do not follow the Gaussian distributions in the majority of cases. It is concluded that the print quality of a set of printed characters can be expressed by one kind of representative measurement, since these four kinds of parameters are linearly dependent upon each other. In this case it is better to measure the peak value of PCS, since PCS can be relatively easily measured by devices. It would be said that the measurement of PCS, which is mentioned in the specification of print quality recommended by ISO [4], is the most important measuring parameter of print quality.

References

[1] ISO/TC97/SC3/N56: "Report of Expert Group 'Printing'" (1965)

[2] I. Yamasaki and T. Iijima: "A Method for Print Quality Evaluation of a Large Number of Data," Jour. of Information Processing Society of Japan, vol.13, pp.225-231 (April 1972); available in English in Information Processing in Japan, vol.12, pp.119-125 (1972)

[3] ——: "Print Quality Evaluation of a Large Number of Data," ibid, vol.13, pp.525-532 (August 1972); available in English in ibid, vol.13, pp.7-12 (1973)

[4] ISO Recommendation R1831: "Printing Specifications for Optical Character Recognition," Ref. No.: ISO/R1831-1971 (E) (1971)