論　文

文字集団の印字品質の数量表現*

山崎　一生**

Abstract

This paper is concerned with print quality evaluation of a large number of data. A computer simulation of evaluation for actual data shows that distributions of evaluation values do not obey the Gaussian distributions in the majority of cases. "Representative" which expresses a statistical value for a set of printed images, is proposed. The representative R is dfiened as: $R=M/N$, where M and N denote the 2nd moment about the origin and a total number of data, respectively. Four kinds of values of representative (darkness, strokewidth, noise factor, and centroid deviation) are strongly correlated each other. The quality for a set of priented characters can be expressed by any one kind of representative.

1. まえがき

光学文字認識において入力資料の印字品質を規定する方法に関して，国際標準化機構(ISO)を中心に約10年前から作業が進められている[1]．作業開始から10数年を経た今日，光学文字読取装置(OCR)の方は実用に入っているが，印字品質の測定に関しては統一的な確立された方法が存在しているとは言い難い[2]．

これまで，印字品質測定に関して行われてきた方法は，次に示す2つである．

ⅰ）紙面上の文字を拡大して標準文字ゲージを当てがい良否判定を行う．

ⅱ）印字の濃さの測定を行う．

上の2つの中で，ⅱ）は機械化することも可能で，大量データの処理を行い客観的なデータを得ることができる．他方，ⅰ）はヒトに頼らざるを得ず，大量に処理して統計的に印字品質を測定評価することは一般に困難であり，かつまた測定の再現性も乏しい．しかしながら，OCRの読取不良の原因究明等に際しては，ⅰ）の測定方法でも利用価値があり，実際に使用されてきた．

OCRの性能評価の場合のように，大量の印字データの品質を定量的に評価する必要があるとき適用できる方法として，著者らは先に大量印字データの品質評価法の原理を提案し[3]，その実用の可能性をも示した[4]．

本論文においては，個々の文字を評価して得られる4種類の評価量をそれぞれ如何に統計的に処理して，文字集団の評価値とするかという問題，更には，4種類の評価統計量を総合して，文字集団の印字品質を1つの値として数量化する問題について考察を加えるものとする．

2. 個々の文字に対して算出される評価量

大量印字データ品質評価方式において，個々の入力文字に対して算出される評価値は，次の4種類である．

ⅰ）印刷鮮明度の尖頭値

ⅱ）平均線幅

ⅲ）ノイズ成分

ⅳ）重心偏位

2.1 印刷鮮明度の尖頭値

文字が与えられる領域における最大，最小反射率をそれぞれ R_{max}，R_{min} とするとき，尖頭値 P は次式で定義される．

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

尖頭値 P は，紙面の明るさに対する印字の濃さの程

* A Quantitative Representation of Quality for a Set of Printed Images by Issei YAMASAKI (The Electrotechnical Laboratory)

** 電子技術総合研究所パターン情報部

度を表わす量で，0から1までの値を取る．淡く印字されていれば0に近い値，逆に濃く印字されていれば1に近い値をとる．(OCR 用に印字された資料の場合には，通常 P は 0.6〜0.8 なる値をとる．)

2.2 平均線幅

白黒2値に変換された入力文字図形の0次モーメントを M，入力文字に対応する標準文字図形の基準線幅時における0次モーメントを M_0 とするとき，平均線幅 w は，次式で定義される．

$$w \equiv M/M_0$$

平均線幅 w は，文字線の太さ(幅)に対応する量で，入力文字図形の線幅が，平均的に基準線幅と一致するとき1を取る．入力の平均線幅が基準線幅より太ければ1以上の値，逆に細ければ1以下の値を取る．

2.3 ノイズ成分

正規図形[3]に変換された入力標準両文字（図形両者の0次モーメントは等しいものとする）をそれぞれ g, g_0 とするとき，ノイズ成分 n は次式で定義される．

$$n \equiv 1-[(g, g_0)/(\|g_0\| \cdot \|g\|)]^2$$
$$= 1-(g, g_0)^2$$

ここで，$(\cdot, \cdot)$ は内積を，また $\|\cdot\|$ はノルムを表わす．

ノイズ成分 n は，入力文字図形の崩れの程度を表わす量で，0から1までの値を取る．入力と標準文字図形とが完全に一致すれば，n は0を取り，崩れの程度に従って1に近い値を取る．

2.4 重心偏位

入力標準両文字図形が最も良く重なり合う状態（類似度[5]最大）にした場合における両者の重心座標をそれぞれ $\boldsymbol{D}$, $\boldsymbol{D}_0$ とすると，重心偏位 $\boldsymbol{d}$ は次式で定義される．

$$\boldsymbol{d} \equiv \boldsymbol{D}-\boldsymbol{D}_0$$

また，重心間距離 d は，偏位 $\boldsymbol{d}$ から次式により得られる．

$$d \equiv \|\boldsymbol{d}\|$$

重心偏位は，入力文字図形と標準文字図形とが最も良く重なり合うような状態にしたとき，入力の重心が標準の重心を基準にして，どの方向にどの程度ずれているかを示す量である．

3. 評価値の分布

印字条件が明かな OCR 用入力資料を，文字データ集積装置[6]によって走査し，これを計算機シミュレーションによって評価し，結果の分析を行うものとする．

Table 1 Data Used for Simulation

Font	OCR-A Numeral "0"
Printing Unit	Lineprinter
Ribbon	NBC: Silk, 13 meters long.
Paper	OCR Paper
Smpling	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)

ここで用いる入力データの諸条件は，**Table 1** に示す通りである．

評価処理によって得られる4種類の評価量の頻度分布を **Fig. 1**(次頁参照)に示す．この図には，インクリボンがほとんど消耗していないとき，中程度に消耗したとき，かなり消耗したときの3段階を代表として示した．評価値の分布は，印刷鮮明度の尖頭値 (PCS_{peak}) を除き，正規分布からかなりはずれた分布となっている．

このような正規分布でないものにおいては，一般的な統計量（最大値，最小値，平均値，標準偏差等）では，その集団の数量表現として適切ではない．また，最小値，最大値，平均値，標準偏差なる4種類の数字を羅列することも煩雑である．そこで，印字品質評価において都合のよいと考えられる統計量として，次の4.で定義する「代表値」を導入するものとする．

4. 代表値による文字集団の印字品質の表現

代表値 R を次式で定義する

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

ここで，$\{x_i\}$ は入力データを，N はデータ総数を表わす．

代表値 R は，原点まわりの2次モーメントを総個数で除し，その平方根をとった量である．上で定義した代表値 R と平均値 M，標準偏差 S との間には，次の関係式が成立する．

$$R^2 = M^2 + S^2(1-1/N)$$

代表値 R は，入力データ（集合）の平均値と標準偏差とを組み合わせた量である．データ総数が大きい場合には，平均値と標準偏差とをそれぞれ2乗して加え，平方根をとった値にほぼ等しい．また，データの分布が正規分布に近く，標準偏差も小さい場合には，代表値 R と平均値 M とは，ほぼ等しい値をとること

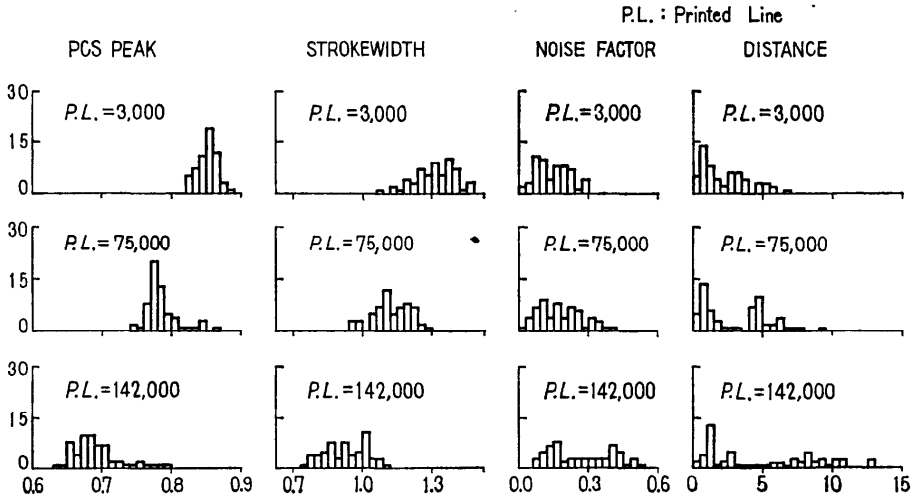

Fig. 1 Distribution of evaluation values.

Table 2 Results of Evaluation

** PCS PEAK **

PRINTED LINE	MEAN	MINIMUM	MAXIMUM	STANDARD DEVIATION	REPRE-SENTATIVE
1	0.852	0.825	0.890	0.014	0.852
2	0.848	0.819	0.875	0.013	0.848
3	0.842	0.814	0.884	0.015	0.842
4	0.821	0.785	0.861	0.017	0.821
5	0.824	0.786	0.860	0.018	0.824
6	0.814	0.773	0.860	0.021	0.814
7	0.854	0.819	0.894	0.015	0.854
8	0.786	0.741	0.862	0.024	0.786
9	0.786	0.751	0.856	0.025	0.786
10	0.777	0.731	0.850	0.029	0.777
11	0.756	0.707	0.852	0.036	0.757
12	0.735	0.676	0.817	0.031	0.736
13	0.732	0.687	0.810	0.029	0.733
14	0.712	0.625	0.837	0.040	0.713
15	0.692	0.631	0.793	0.034	0.693

** STROKEWIDTH **

PRINTED LINE	MEAN	MINIMUM	MAXIMUM	STANDARD DEVIATION	REPRE-SENTATIVE
1	1.285	1.050	1.423	0.082	1.287
2	1.229	1.075	1.362	0.071	1.231
3	1.210	0.996	1.364	0.071	1.212
4	1.178	0.968	1.300	0.077	1.180
5	1.167	0.991	1.303	0.076	1.170
6	1.156	0.965	1.319	0.083	1.159
7	1.210	1.042	1.345	0.080	1.213
8	1.102	0.922	1.248	0.079	1.104
9	1.074	0.845	1.209	0.079	1.077
10	1.024	0.806	1.178	0.075	1.027
11	1.023	0.764	1.166	0.094	1.028
12	1.002	0.800	1.129	0.077	1.005
13	0.944	0.791	1.089	0.081	0.948
14	0.957	0.759	1.113	0.084	0.961
15	0.907	0.726	1.084	0.092	0.912

** NOISE FACTOR **

PRINTED LINE	MEAN	MINIMUM	MAXIMUM	STANDARD DEVIATION	REPRE-SENTATIVE
1	0.145	0.023	0.295	0.071	0.161
2	0.141	0.028	0.323	0.069	0.157
3	0.142	0.034	0.273	0.067	0.157
4	0.159	0.001	0.356	0.085	0.180
5	0.154	0.034	0.331	0.071	0.169
6	0.173	0.046	0.362	0.073	0.187
7	0.176	0.034	0.400	0.094	0.200
8	0.179	0.026	0.394	0.094	0.202
9	0.199	0.023	0.395	0.116	0.230
10	0.205	0.028	0.422	0.112	0.233
11	0.213	0.037	0.459	0.119	0.244
12	0.202	0.043	0.470	0.106	0.228
13	0.253	0.069	0.541	0.116	0.278
14	0.223	0.062	0.457	0.104	0.246
15	0.271	0.071	0.520	0.133	0.301

** DISTANCE BETWEEN CENTROIDS **

PRINTED LINE	MEAN	MINIMUM	MAXIMUM	STANDARD DEVIATION	REPRE-SENTATIVE
1	2.3	0.1	6.8	1.7	2.8
2	2.3	0.1	6.8	1.8	2.9
3	2.3	0.1	5.2	1.7	2.8
4	2.5	0.1	7.1	1.9	3.1
5	2.5	0.2	6.3	1.7	3.0
6	2.7	0.1	6.5	1.9	3.3
7	2.8	0.2	8.6	2.2	3.6
8	3.1	0.2	9.2	2.4	3.9
9	3.5	0.5	8.7	2.6	4.4
10	3.8	0.3	9.3	2.6	4.6
11	3.7	0.2	10.5	2.9	4.7
12	3.6	0.3	9.0	2.5	4.4
13	4.1	0.5	9.7	2.8	5.0
14	4.0	0.2	9.9	2.9	4.9
15	5.0	0.3	12.7	3.7	6.2

となる.

先の **3.** において述べた文字データを評価した結果を統計処理した結果を **Table 2** に示す. また, 代表値をプロットしたグラフを **Fig. 2** (次頁参照) に示す. (リボンの消耗段階66000行台の所で異常な結果が見られるが, これはラインプリンタのリボンの反転機構の動作不良により, リボンが通常より多く巻き取られ, 濃く印字されたものと思われる. 原資料も確かに濃く印字されている.)

5. 評価量相互間の相関

先の **Fig. 2** から, 4種類の代表値の間には, 正あ

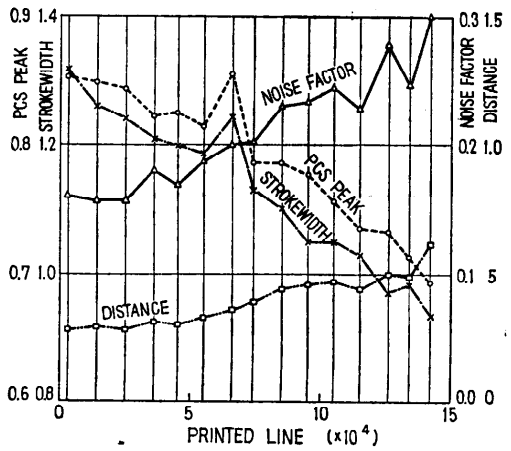

Fig. 2 Results of evaluation for a numeral of "0" (OCR-A)

るいは負の相関が存在することが読み取れる．代表値の間の相関係数を **Table 3** に示す．

この表から，代表値相互間の1次従属関係がかなり強いことが結論される．すなわち，統計的に見れば，次のことがいえる．濃く印字されていれば，その資料は印字品質が良い．逆に，淡く印字されていれば，その資料は印字品質が悪い．

このことから，文字集団の印字品質を言い表わすときには，4種類の評価量の中のいずれの代表値を用いても良いことが結論される．

さて，入力文字図形個々の評価値の間の相関係数を **Table 4** に示す．この表を見ると，ノイズ成分と重心間距離との間にはかなり強い1次従属関係が存在するが，これ以外の評価値の間の相関は，それ程高くないことが分かる．このことから，個々の文字を処理して得られる4種類の評価値は，入力文字図形の品質を表現する量として，それぞれ意味のある量であることが結論される．

Table 3 Correlation Coefficients Between Values of Representative

	PCS PEAK	STROKE WIDTH	NOISE FACTOR
STROKEWIDTH	0.977		
NOISE FACTOR	-0.913	-0.943	
DISTANCE	-0.928	-0.946	0.985

Table 4 Correlation Coefficients Between Evaluation Values

	PCS PEAK	STROKE WIDTH	NOISE FACTOR
STROKEWIDTH	0.726		
NOISE FACTOR	-0.194	-0.678	
DISTANCE	-0.144	-0.674	0.905

6. あとがき

大量印字データ品質評価方式によって得られる個々の文字に対する4種類の評価値を，それぞれ文字集団ごとにまとめる統計処理の方法について述べた．実データの評価シミュレーションの結果から，文字集団の評価値として，原点まわりの2次モーメントをもとにした代表値を用いることを提案した．また，4種類の代表値の間の正あるいは負の1次従属関係が強いことから，この中の1つの代表値を，文字集団の評価値とすることができることを結論した．この場合，機械化処理を行うことの比較的容易な印刷鮮明度の尖頭値を測定するのが良いと思われる．この印刷鮮明度の測定は，ISOの印字仕様案[7]にも取り入れられている．この意味で，印刷鮮明度の測定は，印字品質の評価において重要な働きをする測定項目であると言える．

謝辞 統計処理に関連して，電子技術総合研究所パターン情報部オートマトン研究室 磯道義典氏から有益なコメントを頂いた．ここで用いたデータの原資料は，日立製作所中央研究所の山本真司氏から御提供頂いたものである．記して感謝の意を表したい．日頃御指導御鞭撻を頂く東京工業大学工学部 飯島泰蔵教授，並びに本研究の機会を与えられた電子技術総合研究所パターン情報部長 西野博二氏に感謝したい．

参考文献

1) ISO/TC97/SC3/WG1/N56: Report of the Expert Groupe 'Printing' (1965)
2) JIS C 6253：光学文字認識のための印字仕様，日本規格協会 (1975)
3) 山崎，飯島：大量印字品質評価法，情報処理，Vol. 13, No.4, pp. 225～231 (1972)
4) ——：大量印字データの品質評価，情報処理，Vol. 13, No. 8, pp. 525～532 (1972)
5) ——：文字図形の標本化について，信学論誌 C, Vol. 51-C, No. 9, pp. 428～429 (1968)
6) 山崎，竹村：文字データ集積装置と文字図形データ，電総研彙報，Vol. 37, No. 9, pp. 828～841 (1973)
7) ISO Recommendation R1831: Printing Specifications for Optical Character Recognition, Ref. No.: ISO/R1831-1971 (E) (1971)

(昭和51年5月28日受付)
(昭和51年8月16日再受付)