ショートノート

文字読取り装置のための最適閾値設定アルゴリズム†

山形秀明†† 立川道義††† 佐藤 元†††
富田 豊†† 堀内敏夫††

OCR は，手書きまたは印字された文字の構成画素の濃度をある閾値で2値化したものについて文字を認識する装置である．文字濃度がページごとに異なるような原稿の認識を行う際に，認識率を向上させるためには，ページごとに最適閾値を求めることが必要となる．本研究は，特定の OCR について原稿から自動的に最大認識率を与える最適閾値を求めるアルゴリズムの開発を目的としている．2値画像の周辺長をコード数，面積を黒画素数と定義すると，(コード数)²/黒画素数が最大となる閾値は文字濃度が等しいページについては文字の種類に関係なく等しいことが実験的に確かめられたので，これを利用して OCR にとっての最適閾値を求めた．

1. まえがき

多くの文字読取り装置（以下 OCR と略す）は前段に原稿を読み取るイメージ・スキャナ部を持つ．イメージ・スキャナは原稿濃度をある閾値で2値化し，OCR に2値画像を送る．OCR はその2値画像の特徴量を抽出し，あらかじめ OCR 中に記憶してある各文字の特徴量と比較することで文字の認識を行っている[1]．このとき図1に示すように，2値化する際の閾

† Optimal Threshold Selection Algorithm for an Optical Character Reader by HIDEAKI YAMAGATA (Department of Instrumentation Engineering, Faculty of Science and Technology, Keio University), MICHIO TACHIKAWA, GEN SATO (Ricoh Research and Development Center), YUTAKA TOMITA and TOSHIO HORIUCHI (Department of Instrumentation Engineering, Faculty of Science and Technology, Keio University).

†† 慶応義塾大学理工学部計測工学科
††† (株)リコー中央研究所

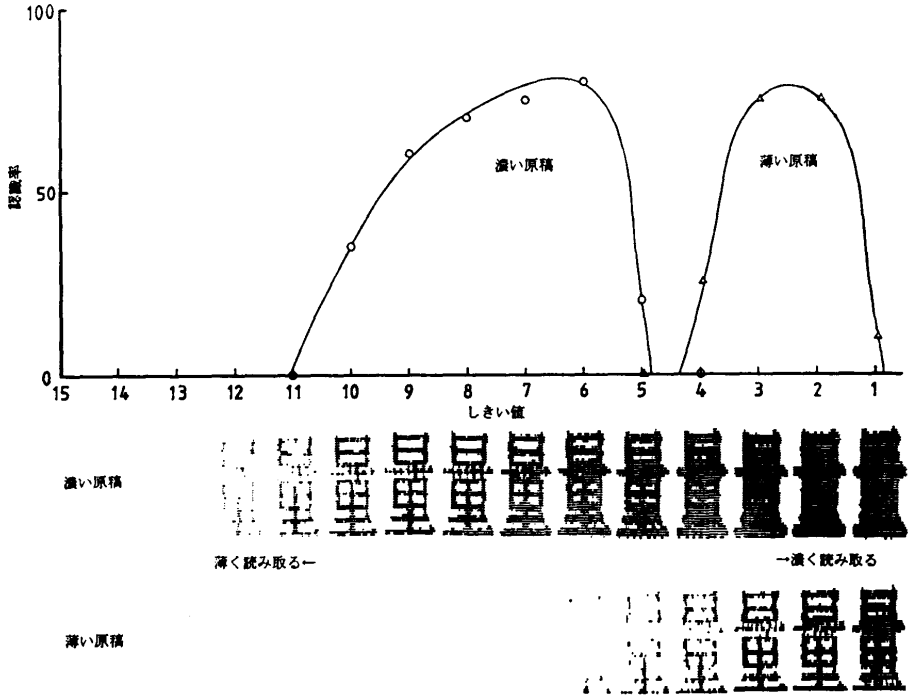


図 1 閾値による画像および認識率の変化
Fig. 1 Threshold dependency of an image and its recognition ratio.

値を変化させることにより，原稿を濃くあるいは薄く読み取ることができる．濃い原稿では閾値6で認識率が最大となり，薄い原稿では閾値2で最大となった．このように，認識率を向上させるためには，原稿濃度の違いによって閾値を変える必要があることがわかった．

ドット・プリンタによる原稿など，ページごとに濃度の異なる品質の低い原稿については，その原稿ごとに最適な閾値を求めることが必要不可欠である．この最適閾値を求める方法として，多値画像上の各画素の濃度ヒストグラムに現れる谷を利用した自動閾値選定法[2]や濃度微分による閾値選定法[3]，文字幅による閾値選定法[4]などが考案されている．しかしながらそれらの方法は背景と文字の分離に主眼を置いたものが多く，OCR の認識率などの観点から見ると不十分なものが多い．そこで本稿では，文献1）に示した認識アルゴリズムを用いた OCR に対して最適閾値を自動的に求めるアルゴリズムの開発を目的とし，実験を行った．

2. 実　　験

実験装置の概要を図2に示す．使用した OCR は前段にイメージ・スキャナ部が直結され，2値化閾値を手動で設定するが，ここでは 16 階調のイメージ・スキャナ部の出力をマイクロコンピュータに転送し，マイクロコンピュータ上で2値化し，さらに文字ごとの切出し，黒画素数，コード数の測定を行った．OCR ではマイクロコンピュータから送られてくる2値画像に対する文字認識を行う．イメージ・スキャナの読取り密度は 300 dpi で行った．原稿は表1に示す5通りの濃度のものを，汎用ドット・プリンタで作製した．また文字は表2に示す 45 文字を用いた．

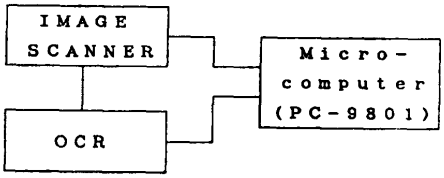


図 2 実験装置の概要
Fig. 2 Schematic diagram of the experimental system.

表 1 原稿の作製方法
Table 1 Method to make machine printed samples.

原稿濃度	印字された状態
a	新しいリボン，強い印字圧で印字した
b	新しいリボン，弱い印字圧で印字した
c	ある程度使用したリボン，中程度の印字圧で印字した
d	古いリボン，強い印字圧で印字した
e	古いリボン，弱い印字圧で印字した

表 2 サンプル文字
Table 2 Character samples.

数　字	2458
英　字	ABEJHM
平仮名	はにのまわだ
漢　字	在日来呼明　板平実強速 形谷意逢権　勲電薫舗鐘 量機慨槽論　驚臓艦蟻

3. 結　　果

使用した OCR の閾値が認識率に与える影響は，図

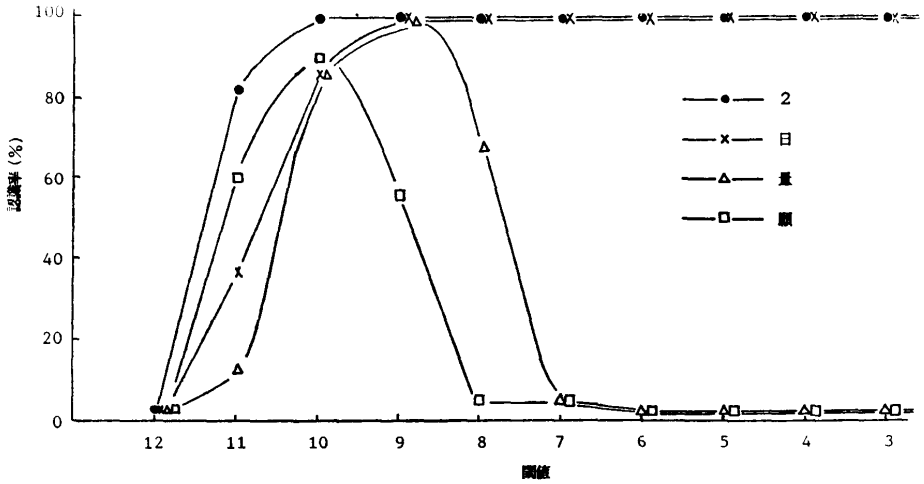


図 3 種々の文字に対する閾値による認識率の変化
Fig. 3 Threshold dependency of recognition ratio for various characters.

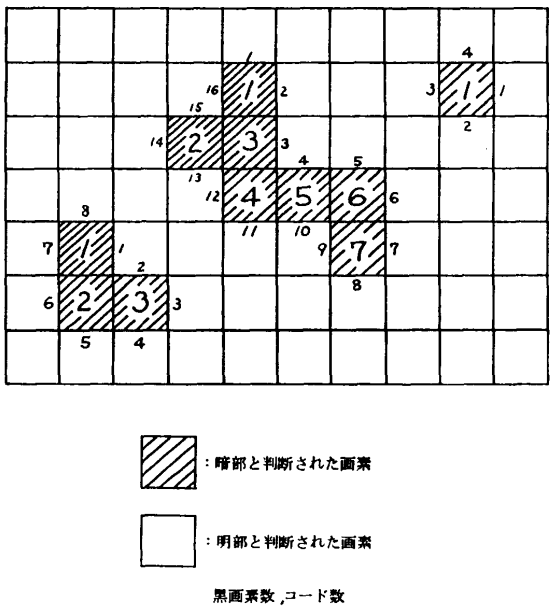


図 4 黒画素数とコード数の定義
Fig. 4 The definitions of the number of black pixels and the number of code.

3 に示すように「2」や「日」のような簡単な文字ではある閾値以下では 100% の認識が可能であり，許容閾値の幅が広い．しかし「量」や「願」のような複雑な文字では認識率は閾値の変化に対して上に凸となり，文字に依存する最適閾値とその許容幅が狭い．このことから閾値は文字によって最適な値が異なり，閾値の最適幅が狭い文字に対して最適値をどのように設定するかが最も厳しい課題である．

本研究では容易に測定できる変数の中から図 4 に示す黒画素数 b とコード数 c を選び，閾値の変化に対して b, c がどのように変化するかを見た．ここで b は2値画像における暗部の画素数であり，c は2値画像における暗部と明部の境界画素数である．その結果，b は図 5 の一例で示すように，閾値の減少に対して，原稿濃度に無関係に増加することが確かめられ，c は図 6 の一例で示すように，認識率と閾値の関係と同様に上に凸または閾値の減少に対して飽和することが明らかとなった．b および c を組み合わせて，閾値に対して上に凸になる関数 $f(b, c)$ が求まれば，convex fuzzy とみなすことができ，membership 関数最大となる閾値が定められる．この閾値と特定の認識アルゴリズムによる認識率との間に高い相関があれば，$f(b, c)$ はその OCR の閾値と変数とする認識率関数の写像と考えることができる．

そこで，b と c の種々の組合せについて閾値を変化させ各閾値での認識率との相関を調べた．その結果，表 2 にある実験に用いた文字に関しては，図 7 に示すように文字パターンの複雑さにかかわりなく，c^2/b の最大値をとる閾値はほぼ一致し，認識率の最大値をとる閾値との間に高い相関があることがわかった．結果を図 8 に示す．

図 8 の回帰直線により c^2/b が最大となる閾値から，認識率が最大となる閾値を推定したとき，今回実験に用いた延べ 225 文字中 150 文字について，推定した閾値と実際に認識率最大となる閾値が一致した．また他の文字についても，その差は ±1 以内であり，b^2/c が最大となる閾値に注目することで文字の種類に関係なく原稿濃度に応じた最適閾値を高い精度で求めることができた．

4. むすび

本稿では文字画像の特徴量として，コード数と黒画

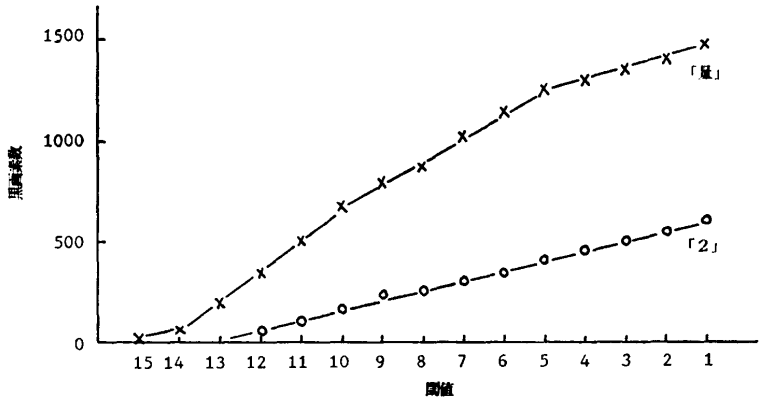


図 5 閾値に対する黒画素数の変化
Fig. 5 Threshold dependency of the number of black pixels.

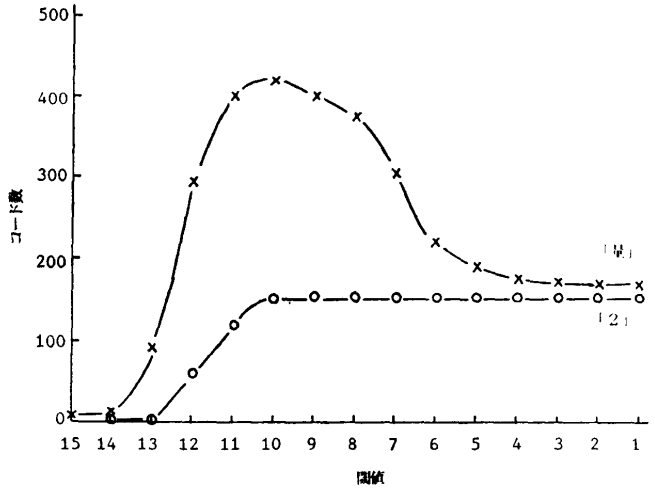


図 6 閾値に対するコード数の変化
Fig. 6 Threshold dependency of the number of code.

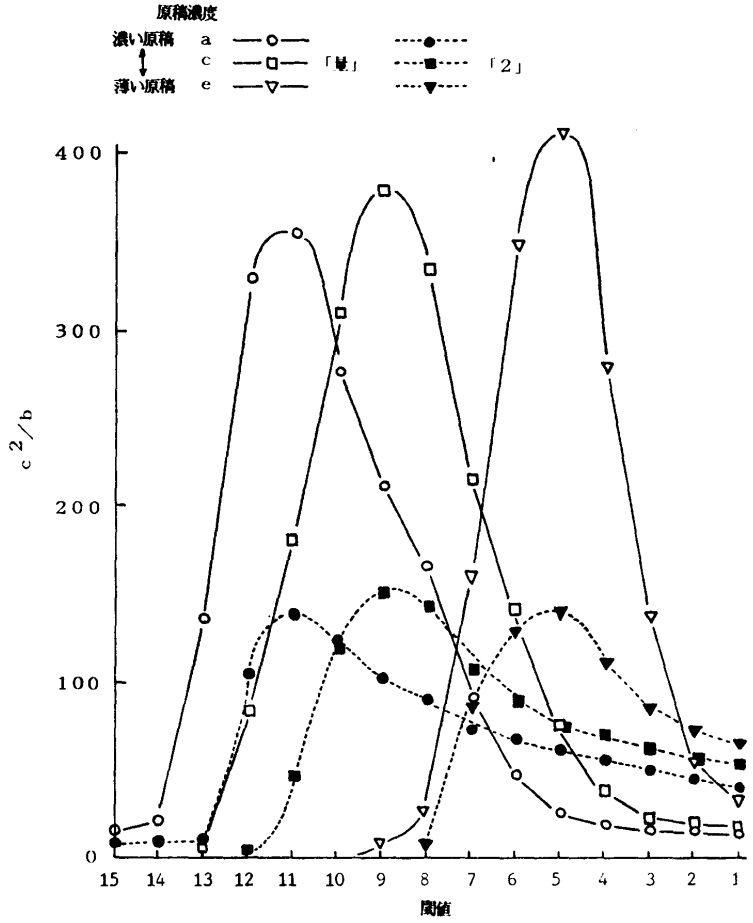


図 7 種々の原稿濃度に対する閾値と (コード数)²/黒画素数 の関係
Fig. 7 Threshold level vs (the number of code)²/the number of black pixels for various printing densities.

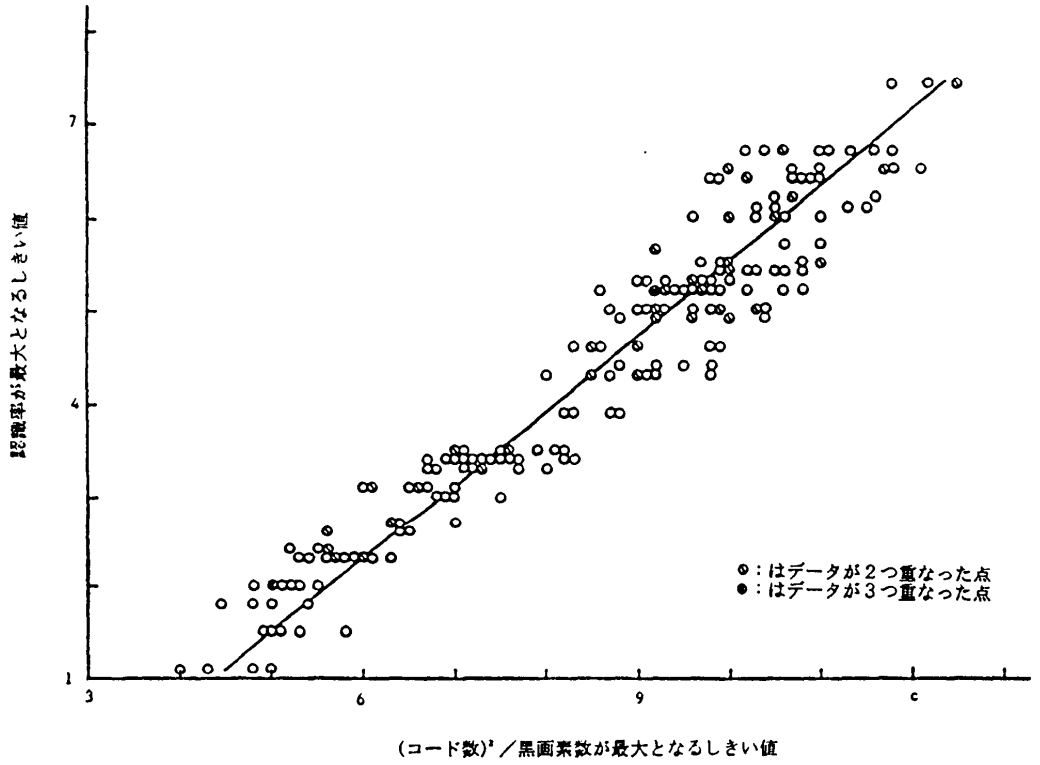


図 8 (コード)²/黒画素数 が最大となる閾値と認識率が最大となる閾値との関係
Fig. 8 Relationship between the threshold levels which give the maximum of (the number of code)²/the number of black pixels and the recognition ratio.

素数を定義し，(コード数)²/黒画素数が最大となる閾値を利用することで，文字種に関係なくその原稿濃度に応じて今回使用した OCR にとっての最適閾値を決定する手法を提案した．その結果，66% の文字について OCR の認識率が最大となる閾値を原稿から自動的に決定することができ，他の文字についてもその誤差は ±1 の範囲内であった．今後は実用化に向けて，測定時間の短縮方法などについて検討していく．

参 考 文 献

1) 立川，嶺脇：漢字認識における特徴量圧縮方式，第 35 回情報処理学会全国大会論文集，1H-4, pp. 1929-1930 (1987).
2) Prewitt, J. M. S. and Mendelsohn, M. L.: The Analysis of Cell Images, *Ann. N. Y. Acad. Sci.*, Vol. 128, pp. 1035-1053 (1966).
3) Weszka, J. S., Nagel, R. N. and Rosenfeld, A.: A Threshold Selection Technique, *IEEE Trans. Comput.*, Vol. C-23, No. 12, pp. 1322-1326 (1974).
4) 中野，小関，山本：低品質刻印文字の認識，信学技報，PRL 85-29, pp. 21-29 (1985).

(昭和 63 年 8 月 16 日受付)
(平成 元 年 4 月 11 日採録)

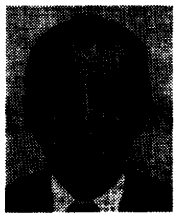
山形 秀明 (正会員)
1964 年生．1988 年慶応義塾大学理工学部計測工学科卒業．同年(株)リコー入社．中央研究所にて画像処理，文字認識の研究に従事．電子情報通信学会会員．

立川 道義 (正会員)
1958 年生．1982 年早稲田大学理工学部応用物理学科卒業．同年(株)リコー入社．中央研究所にて画像処理，文字認識の研究に従事．電子情報通信学会，S. I. D. 各会員．

佐藤 元 (正会員)
1949 年生．1973 年千葉大学理学部卒業．1980 年東京大学大学院理学系研究科鉱物学専攻博士課程修了．理学博士．国立極地研究所での隕石の研究を経て，1982 年(株)リコー入社．中央研究所にてパターン認識の研究に従事．日本鉱物学会会員．

富田 豊（正会員）

1949年生．1973年慶応義塾大学工学部計測工学科卒業．1975年同大大学院工学研究科計測工学専攻修士課程修了．同年(株)東芝入社．1977年慶応義塾大学医学部助手，1981年同大理工学部助手，現在助教授．工学博士．測定科学，生体計測の研究に従事．日本人間工学会，計測自動制御学会，応用物理学会各会員．



堀内 敏夫

1925年生．1947年慶応義塾大学工学部機械工学科卒業．同年慶応義塾大学工学部助手，専任講師，助教授を経て現在教授．工学博士．測定科学，測定工学に関する研究に従事．応用物理学会，計測自動制御学会，日本エム・イー学会，日本音響学会等各会員．