パターン認識としての漢字の識別

1. まえがき

文字認識はいわゆるパターン認識のなかでももっとも古くから研究され，そして実用化の進んだ分野である．この技術を利用したOCR (Optical Character Reader，光学式文字読取装置)は大量の文字データを高速・正確に入力できる機器として，情報処理に関係するさまざまなところで使われている．最近では手書きと活字の英字，数字，記号，片仮名を読み取ることのできる小型OCRが市販されるまでに至っており，印刷漢字を読み取ることのできるOCRもごく最近開発された[1]．本稿ではこの漢字OCRについて認識原理，性能仕様，将来の応用などを紹介する．

2. 印刷漢字の認識方法

日本語は少なくとも2,000種以上の漢字を用いて表記され，個々の文字パターンの構造は英数字などと比べて格段に複雑である．多数の複雑な文字パターンを読取対象とする漢字OCRにおいては，認識処理の能率化を行なうための大分類法と個々の文字パターンを正確に識別することのできる強力な文字認識原理が必要である．開発された漢字OCRでは，大分類と個別認識を組み合わせた2段階の認識法が採用されている．

2.1 漢字パターンの大分類

漢字パターンを大分類することは漢和字典の文字の配列にも使われている．同一の偏や旁を有する文字が1カ所にまとめられているので，偏や旁によって検索する文字の属する類をまず定め，つぎにこの類の中から画数によって所用の文字を素早く引き当てることができる．これと同様のことが実用的な装置規模と速度で漢字を認識する場合にも必要である．機械によって容易に抽出することのできる特徴を発見し，これを用いて複数の類似文字群に漢字パターンを分類する技術，入力文字がどの類に属するかを先の特徴を用いて正確に決定し，答となるべき文字をこの類(候補文字群)の中から捜す技術が必要である．

漢字パターンの大分類法としては複雑指数と四辺コードと呼ばれる二つの方法が開発された[1][2][3]．複雑指数は文字を構成する線分が全体としてどの程度こみ入っているかを示す指標である．漢字パターンが主として縦または横の線分で構成されることから式(1)，(2)に示すように横方向と縦方向についての文字線密度によって定義される．

$$\text{横方向の複雑指数：}\quad c_x=l_y/\sigma_x \qquad (1)$$

$$\text{縦方向の複雑指数：}\quad c_y=l_x/\sigma_y \qquad (2)$$

上式においてl_xおよびl_yは，それぞれ横方向および縦方向の文字線の長さの和をあらわしている．斜線は傾きの程度によって，横線成分と縦線成分に分解されてl_x，l_yに加算される．σ_xとσ_yは文字パターンの横方向および縦方向への拡がり量（文字パターンの2次モーメントから求めた分散量)である．lとσはともに長さの次元をもっているので複雑指数は無次元の数量である．したがって，漢字パターンの大きさが相似的に変化しても，複雑指数の値は影響されずに一定であるという性質がある．

図1は漢字(＋印)，ひらがな(○印)，および英

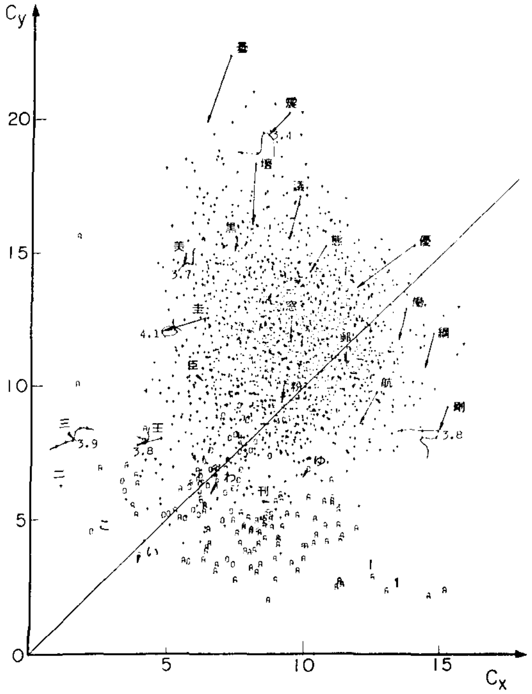


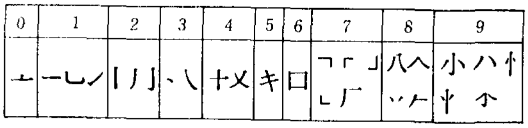
図 1 複雑指数座標軸上での漢字パターンの分布図

大小文字と数字記号(A印)から成る約 2, 300 字について，c_x と c_y を計算して c_x-c_y 平面上にプロットしたものである．たとえば「曇」や「震」の文字のように横線の多い文字は縦方向の複雑指数 c_y が大きくこの図の最上部に位置している．逆に「剛」の文字は縦線が多く，c_x が大きいので，この図の最右端に位置している．またひらがなや英字は漢字に比べると簡単な文字であるので，原点に近い所に分布している．そして全体としてながめると概ね平坦な分布であることがわかる．

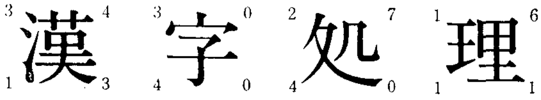
複雑指数を用いると漢字パターンの大分類と候補文字選択を次のようにして行なうことができる[3]．

(1) 図 1 を正方格子で区切り，格子点の周囲に存在する文字を集めて類似文字群として記憶する．

(2) 入力文字パターンの複雑指数を測定して，複雑指数平面上での位置を定める．そしてもっとも近傍に存在する格子点を見出し，この格子点で代表される類似文字群を検索し候補



(a) 部分図形に対する数字コード



(b) 四角号碼の例

図 2 四角号碼法の説明図

文字群とする．

下記の例は格子点の値とそれに対応した類似文字群である．

(c_x= 8，c_y=19)：震，義，置，選，費，集，壇，層……

(c_x=10，c_y=12)：窓，森，権，鉄，歴，廃，惑，明……

(c_x= 9，c_y= 6)：ゆ，ば，ぬ，あ，外，印，伯，材……

四辺コードは現代中国で行なわれている漢字の分類法(四角号碼法)とよく似た方法である．四辺コードの前に四角号碼法を紹介しよう．この方法は**図 2** に示すように漢字パターンの四隅の部分パターンを10種類に分類して 0 ～ 9 のコードに対応させることにより，漢字を 4 桁の数字コードであらわし字典を引く方法である．偏や旁を知らない人でも容易に使うことができる非常に能率的な方法である．四角号碼法は漢字パターンの分類に役立つ情報が文字の四隅に集中していることを示唆するものであるが，四辺コードはこの性質を数量的に調べるための実験を通じて発見された．**図 3** はこの実験の結果を示すものである．約 2, 000 種の漢字パターンを縦横50点でサンプルし，各点でのエントロピーを求め，この値に雑音に対しての安定係数を掛け合わせてプロットしたものである．黒丸の大きさが大分類に役立つ情報の大きさに比例している．文字の四隅というよりは上下左右の四辺に情報が集中していることがはっきりと示されている．

四辺コードはこのような検討結果から**図 4** のように文字の上下左右に矩形の検査領域を設け，この領域の中に存在する文字線の量（この場合は文

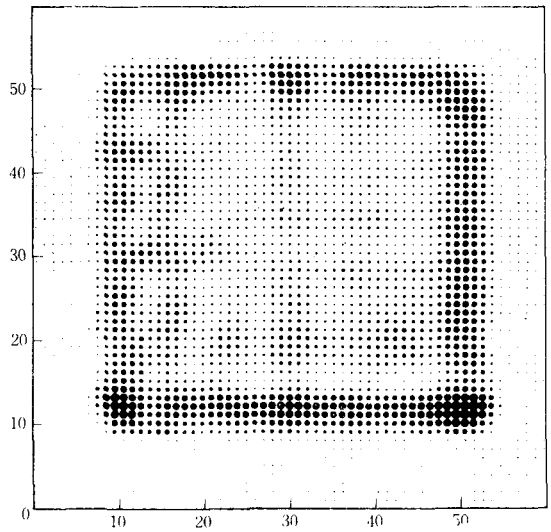


図 3　大分類情報の多い領域の分布図

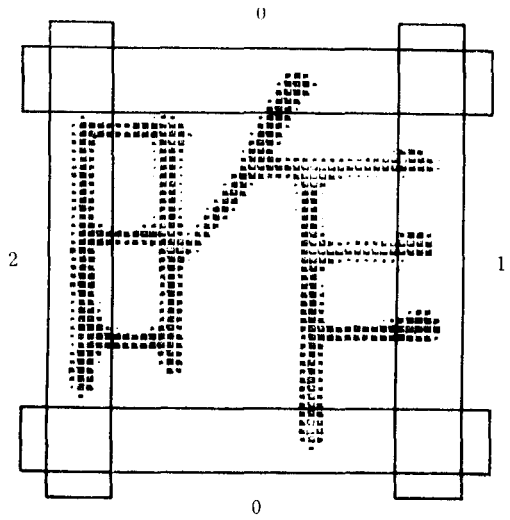


図 4　四辺コード化の方法

字線部に該当するサンプル点数の和）を0，1，2の3段階に量子化する方法である．「昨」の文字の場合は左辺の領域には長い垂直線が含まれているのでコードは2，上辺と下辺の領域にはほとんど文字線が存在しないのでコードは0，右辺はこの中間であるのでコードは1となる．結局この文字は「2010」にコード化される．四辺コード化の方法はきわめて簡単であり，機械によって容易にコードを決めることができる．またその分類効率は四角号碼のそれと同等以上であることも確認されている［2］．

四辺コードを用いた漢字パターンの大分類と候補文字選択のやり方［3］はつぎの通りである．

(1)　各文字種ごとに四辺コードを求め，同一のコードを有する文字を集めて類似文字群として記憶する．

(2)　入力文字パターンの四辺コードを求めて，このコードに対応した類似文字群を検索し候補文字群とする．

下記の例は四辺コードとそれに対応した類似文字群である．

(0210)：倍，任，住，佐，借，倹，枚，舎……

(1210)：進，道，達，途，連，速，遂，走……

(2222)：図，団，因，国，困，囲，園，閑……

複雑指数は文字全体の構造に関係し，四辺コードは文字の周囲の形状にのみ関係する．両者は統計的に独立であるので，図5に示すようにそれぞれの方法によって得られた候補文字群の中から，共通して含まれる文字を残すと，より効率のよい候補文字選択が行なえる．2,000字種を認識対象とした場合に候補文字数の平均は68字である．結局，大分類とそれにもとづく候補文字選択の方式を採用することにより，答となるべき文字を捜す

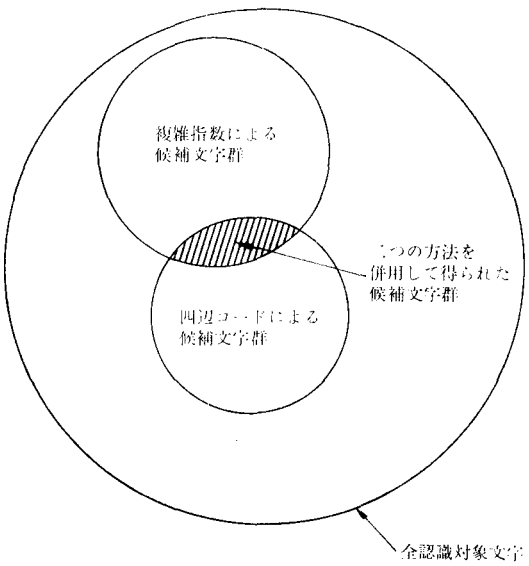


図 5　複雑指数と四辺コードを併用した候補文字の選択法

範囲を約1/30(68/2,000)に狭めることができる．

2.2 漢字パターンの個別認識

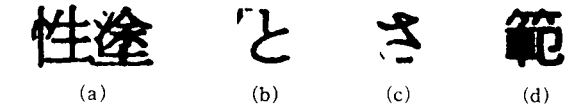
文字を認識するということは，読み取り対象となる文字の種類毎に標準のパターンもしくは特徴を用意し，これと入力文字パターンとを比較照合してもっとも類似した文字を発見し，その文字の種類名を出力することである．漢字認識の場合には入力文字の候補となる文字群をまず選び，つぎにこの中の個々の文字の認識を行ない，一つの文字を答として決定することである．

印刷漢字を正確に認識することのできる文字認識原理としては，複合類似度法[1][2]と混合類似度法[3]と称する方法が用いられている．これらの方法は一般にパターン・マッチング法とよばれているものの一種であって，入力文字パターンと標準文字パターンとを重ね合わせ（パターン・マッチング），両者の類似性を類似度とよばれる尺度であらわすことで特徴づけられる．類似度法が実用的なものであるためには，**図6**に示すように位置ずれ，汚れ，かすれ，つぶれなどさまざまな雑音が混入した低印字品質文字であっても安定な類似度出力が得られること，つまり類似度値が雑音に対して影響を受けにくいことが重要である．

複合類似度法と混合類似度法のキー・ポイントは，文字パターンに含まれている雑音成分が入力文字パターンとは独立ではなく強い相関関係をもっていることに着目し，両者の位相関係を解析することによって雑音に対して安定な認識方式を確立した点にある．

いま，未知入力文字パターンを縦横N点でサンプルし，各点でのインクの濃淡の値を要素としてもつ$N\times N$次元のベクトルgであらわすこととする．従来の方法（単純類似度法）では読み取り対象文字毎に雑音の加わっていない理想的な文字パターンfを想定し，これを標準パターンに用いてfとgとの類似度Sを次式で定義する．

$$S=(f,\ g)/\|f\|\cdot\|g\| \qquad (3)$$



(a)　(b)　(c)　(d)

図 6　雑音の混入した低印字品質文字パターンの例
(a)位置ずれ　(b)汚れ　(c)かすれ　(d)つぶれ

ここで$(f,\ g)$は二つのベクトルの内積を，fとgはそれぞれのベクトルのノルム（長さ）をあらわしている．(3)式の幾何学的な意味はfとgとが$N\times N$次元の空間で張る角度θの余弦である．両ベクトルの方向が一致したとき$(\theta=0)S$は最大値1をとり，このとき両パターンはもっとも類似しているとみなされる．θの増加につれてSの値は減少し，その分だけパターン間の相違があると判断される．n種類の文字を読取対象とした場合には，

$$S_i=(f^i,\ g)/\ f^i\cdot g\ (i=1,\cdots,n) \qquad (4)$$

を計算し，最大値を与えるS_jと次最大値を与えるS_kを求める．S_jおよびS_jとS_kとの差が基準値を超えている場合にはS_jを与える文字を入力文字パターンの答であるとする．この方法は簡単ではあるが，入力文字パターンと雑音成分との関係がどのようなものであろうとただ一つのパターンfで文字種を代表させるという無理がある．雑音の量が大きくなると類似度値は急激に減少するので正確な識別を行なうことはできない．

複合類似度法ではさまざまな雑音が混入した多数の現実のパターン（標本パターン）を用いて類似度を定義する．すなわち入力パターンgと全標本文字パターン$\{f_i\}$との単純類似度の二乗平均値を計算する．

$$S^*=\left[\frac{1}{m}\sum_{i=1}^{m}\frac{(f_i,\ g)^2}{\|f_i\|^2\cdot\|g\|^2}\right]^{\frac{1}{2}} \qquad (5)$$

ここでmは標本パターンの数である．

$\{f_i\}$の分布をより少ない次元の空間に写像して最良近似を求める方法はパターン認識の分野ではKL変換としてよく知られている．この方法によると，$\{f_i\}$の分布を特徴づける特徴ベクトル成分の組$\{\phi_j\}$を固有値問題を解くことによって求めることができる．$\{\phi_j\}$のうちの主要ないくつかを

用いて(5)式は近似的に次式のように書き改められる．

$$S^*=\left[\sum_{j=1}^{l}\frac{\lambda_j}{\lambda_1}\frac{(\phi_j,\ g)^2}{\|g\|^2}\right]^{\frac{1}{2}} \qquad (6)$$

ここでλ_jはϕ_jに対応した固有値である．一般的にはmが数百，数千のオーダーであるのに対し，lは3程度で充分であることが実験を通じて確められている．**図7**の(a)，(b)，(c)はこのようにして求めた漢字「玉」の標準パターンϕ_1，ϕ_2，ϕ_3を図示したものである．各パターンは縦横15点であらわされた濃淡図形である．黒く塗りつぶした円は正の値をとる部分，破線の円は負の値をとる部分である．それぞれの円の大きさが各点の正または負の値に比例する．

複合類似度法の幾何学的な意味は，入力パターンgを特徴ベクトル$\{\phi_j\}$で張られる空間に射影した成分g_0とgとのなす角の余弦で両者の類似性を評価することである．すなわち入力パターンgに含まれる雑音成分のうち，文字パターン集合$\{f_i\}$に固有の雑音成分を除去し，残ったものが評価されることになる．入力パターンに混入する雑音を含めて評価する単純類似度法に比較して，雑音に対して安定な認識法となっているのである．

漢字パターンを全体として見ると，「微—徴」，「土—士」，「衷—哀」，「玉—王」など相互にきわめて類似したパターンが2～3%の割合で存在する．混合類似度法はこれらの文字の識別に適した方法で，複合類似度法をもとに理論が組み立てられている．文字パターン集合$\{f_i\}$に類似した文字種(たとえば「玉」に対して「王」の文字)の平均パターンをhとしたとき，hから$\{f_i\}$の成分を差引いて$\{\phi_j\}$と直交したベクトルψをつくる．

$$\psi=\frac{h-\sum_{j=1}^{3}(h,\ \phi_j)\phi_j}{\left[\|h\|^2-\sum_{j=1}^{3}(h,\ \phi_j)^2\right]^{1/2}} \qquad (7)$$

混合類似度はこのψ(図7(d))を用いて次式で定義される．

$$S'=\left[\frac{\sum_{j=1}^{3}\frac{\lambda_j}{\lambda_1}(\phi_j,\ g)^2-\mu(\psi,\ g)^2}{\|g\|^2}\right]^{\frac{1}{2}} \qquad (8)$$

分子の第2項は入力文字パターンgが$\{f_i\}$の分布内にあればほとんど0の値となり，そうでない場合は大きな値となって類似度を減少させる効果がある．いま，入力文字パターンが「玉」であるとしよう．複合類似度では「玉」と「王」の類似度が接近して読取拒否となる場合であっても，混合類似度では「王」の類似度が先の第2項の働きで小さくなるので，両者をはっきりと区別することが可能となる．試作された漢字OCRでは，複合類似度と混合類似度が組み合わされて用いられている．

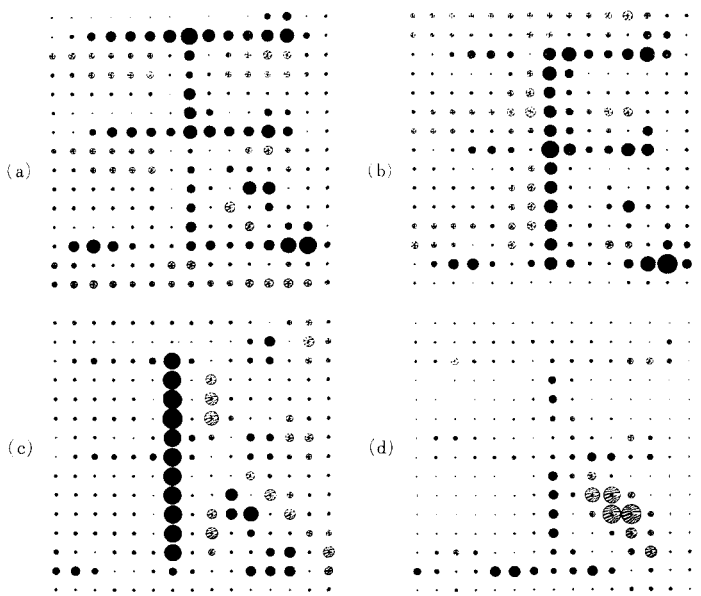


図7　漢字「玉」の複合類似度/混合類似度標準パターン

3. 漢字OCRの性能仕様

以上述べた方法により，漢字パターンを認識し印刷された日本語文書を読み取ることのできるOCRが開発された．この装置の主要な性能仕様は**表1**の通りである．現在市販されているOCRと異なる点は，

① 日常使用される漢字を含めて，2,000種以上の印刷文字を識別す

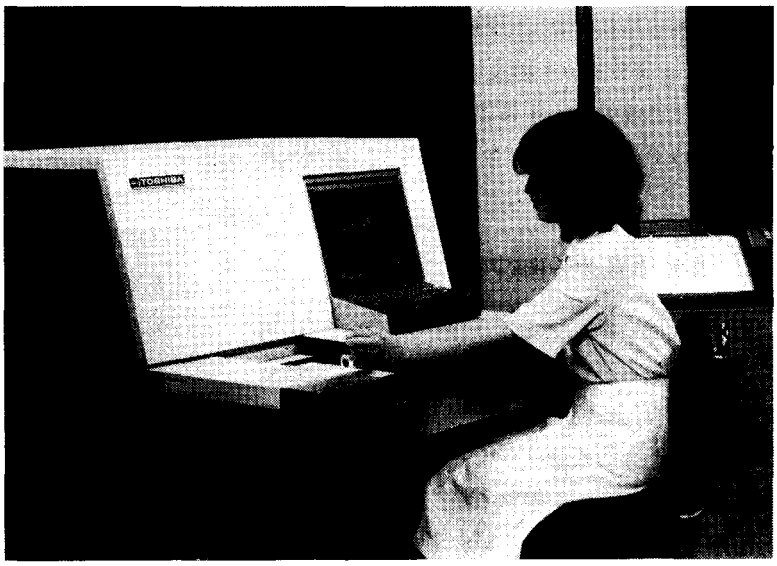


図 8　印刷漢字 OCR の外観

ることができる，

② OCR 用紙に限らず書籍，出版物の印刷に用いられる普通紙を扱うことができる，

③ 図面，写真，表などが混在した文書であっても，その中の文章部分の位置を対話的に指定することによってこれを自由に読みこなすことができる，

などであり，将来，日本語情報の入力装置の一つとして用いられるための条件を備えている．図 8 に装置の外観を示した．

特許公報の文章を入力データとして行なった読取実験では，正読率98.4%，読取拒否率 1.3 %，誤読率 0.3 %を得た．また一般タイプ原稿を入力データとした場合は正読率 99.71 %，読取拒否率 0.25%，誤読率0.04%であった．この性能を漢テレ鍵盤や和文タイプなどを用いた人間のオペレーターの能力と比較すると，入力速度で約 100 倍，入力精度で数倍になる．

表 1　印刷漢字 OCR の性能諸元

型　　式	ページ式
読取文字	漢字，ひらがな，カタカナ，英数字，記号から成る 2,000 字種以上の印刷活字文字
読取方式	候補文字の選択と個別認識を組み合わせた階層構造認識
読取速度	100字/秒
書　　類	普通紙，B 4 サイズまで
フォーマット指定	一頁内の任意の部分を指定可能
リジェクト処理	音訓入力

読取性能をさらに向上させる研究が現在進められているが，それは熟語や文脈情報を用いたソフトウエアにより，OCR の読取結果に修正を施すアルゴリズムの開発である．数値データとは異なり，文章が入力データとなる場合には，同一単語が繰り返し出現することが多い．これらの文字が読取対象外の文字であったり，活字の欠けたものであったり，単独の文字だけからは人間にも区別のつかない類似文字である場合には，読取拒否や誤読が集中して発生することになる．この問題，とくに読取拒否に対して有効な方法は日本語の文法知識を用いて OCR の読取結果の文章の分節分析を行ない，熟語の記憶された辞書を参照して正しい文字を発見・挿入すること（後処理）である．たとえば“ペンシルヴァニア大学における研究の一端を紹介する”云々の文章の読取結果が，“[ペ〈カタカナ)，ペ(ひらがな)]* ンシルヴァニ

*[　] の記号は OCR が読取拒否を起こした箇所を意味する．[　] の中は認識結果の類似度が所定の値を越えた文字を示す．

ア大学における研究の［一(漢数字)，－（マイナス記号)，－（カタカナの長音記号)］端を紹介する”となった場合に，後処理の結果としてはそれぞれカタカナの「ペ」，漢数字の「一」が選択される．

4. 漢字 OCR の将来応用

電子計算機の普及と利用形態の高度化につれて，数値であらわされた情報に限らず日本語情報の入力，記憶，処理，出力に電子計算機を活用することへの社会的要請は今後ともますます増大すると見られている．このための技術上の最大のネックはいかにして日本語情報を高速かつ安価に入力するかということである．このいわゆる入力問題に対する最近の動向は多様な小量データの逐次入力・処理・出力に適した和文ワードプロセッサーの商品化と，蓄積された大量データの高速・高精度の入力に適した漢字 OCR の開発とが目だったものとしてあげられる．

漢字 OCR の開発はそれ自体では日本語情報処理に必要な一つの入力装置を提供することであるが，従来あまりにも高価かつ時間がかかるとして見送られてきたことを実現するという点では大きな意義がある．たとえば特許公報や科学技術文献の読み取りによるデータ・ベースの作成，不動産登記簿，住民登録簿，顧客カード，名簿の読み取りによるファイルの作成と更新，小説・雑誌類の読み取りによる再版工程の省力化などに漢字 OCR を利用することが可能になってきた．これらの応用においては，入力データの印字品質，字体と大きさ，印刷様式，用紙などに特別な制限を設け，これを前提条件とすることはできない．既存のシステムの中で用いられ，また将来の応用範囲を広げるためには，単に漢字の認識ができるというだけでなく，実用ということを充分考えたうえでの周辺技術の開発もまた重要である．

5. あとがき

本研究の一部は通産省の大型プロジェクト，「パターン情報処理システムの研究開発」の一環として行なわれた．昭和53年度には本稿で紹介した漢字 OCR (パイロット・モデル)をさらに改良した実用規模の装置(プロトタイプ・モデル)の開発が予定されている．

参考文献

［1］ 森　健一，坂井邦夫「2,000字種を100字/秒で読む印刷漢字 OCR の開発」，『日経エレクトロニクス』，1977年10月31日号，pp. 102-128.

［2］ 森　健一，坂井邦夫「漢字認識システムへの挑戦」，『情報管理』，Vol. 19，No. 8 (1976).

［3］ 坂井邦夫，平井彰一他「印刷漢字 OCR のためのシミュレーション・システム」，『情報処理』，Vol. 17，No. 7，pp. 587-594 (1976).

［4］ 飯島泰蔵，森　健一「人間の識別能力に迫るOCR，“ASPET/71”」，『日経エレクトロニクス』，1972年5月22日号，pp. 66-80.

［5］ 飯島泰蔵「文字認識装置 ASPET/71」，『テレビジョン』，Vol. 27，No. 3 (1973)，pp. 157-164.

［6］ 飯島泰蔵「混合類似度による識別理論」，『電子通信学会研究会資料』，PRL 74-24 (1974).

さかい・くにお　1944年生
東京大学大学院工学系計数工学修士卒
東芝総合研究所　文字読取装置の研究開発に従事