



印刷漢字 OCR のためのシミュレーション・システム*

坂井 邦夫** 平井 彰一** 河田 勉**
天野 真家** 森 健一**

Abstract

Chinese character machine recognition is a difficult problem because of a large character repertoire and structural complexity of each Chinese character pattern. In order to overcome these difficulties, an efficient "candidate selection" method and a powerful recognition principle should be developed.

In this paper, an experimental study on a practical Chinese character reader is described. Newly developed two stage recognition process is applied to realize an efficient and precise recognition. Two types of indexing parameters are proposed. It is indicated that about 100 candidate characters are obtained using these parameters in the first stage. In the second stage, a recognition principle, called multiple similarity method, is applied. Recognition rates of the order of 97.3 percent are achieved on the data of Japanese Official Patent Gazettes.

Experiments were carried out using a simulation system which was developed in advance of the design of a practical Chinese character reader.

1. ま え が き

我が国における電子計算機利用の高度化につれて、入出力を漢字で行うことの要請は近年とみに増大する傾向にある。漢字の出力機器に比べると入力機器の方はオペレータの手作業が介在するために、大量のデータの高速な入力には適していない。この問題を解決するためには漢字を直接読み取ることのできる装置、つまり「漢字 OCR」の開発が必要である。

漢字 OCR を実現するためには多種類の複雑な形状をした漢字パターンを正確に読み取ることのできる文字認識技術が必要である。そしてまたこの OCR が漢字入力機器の1つとしての利用価値を有するためには、普通紙に様々な様式で印刷された文書を扱おうことのできるシステムを構成する必要がある。これらはいずれも従来の OCR の限界をはるかに越えた要求である。鮮明に印刷された文字データに基づく読取実験

は漢字 OCR の具体的問題の解決には役立たない。実際の文書の読取実験を行うことによってはじめて必要な諸技術を開発することができる。

本論文ではまず印刷漢字の機械認識という困難な問題がどのような方法で解決されたかを述べる。次に特許公報を入力データとし、文字パターンの走査から読取結果の表示と記録までを実時間で行うことのできるシミュレーション・システムを紹介する。そしてこのシステムを用いた読取実験の結果を分析し、印刷漢字 OCR を開発するために解決されなければならない種類の問題点を明確にし検討を加える。

2. 印刷漢字の認識方法

漢字 OCR を日本語文章の入力機器として用いるためには、日常使用される漢字、ひらがな、カタカナ、英数字記号のうち少なくとも2,000種の文字を読取対象とする必要がある。これらの大部分を占める漢字は英数字記号など従来の OCR で扱われてきた文字に比べて文字線の構造が格段に複雑である。従って読取速度と読取率を維持するためにはまず第一に効率の良い大分類方式を開発して辞書パターンと入力文字パター

* A Simulation SYSTEM for Optical Chinese Character Reader by Kunio SAKAI, Shoichi HIRAI, Tsutomu KAWADA, Shinya AMANO and Ken-ichi MORI (Toshiba Research and Development Center)

** 東京芝浦電気(株)総合研究所情報システム研究所

る。全体として平坦で広い範囲の分布が得られている。そこでこの平面上に多くの基準点*とその点を中心とした領域を設定し、個々の領域内に存在する文字を1つの類としてまとめ上げれば、これらの文字を複雑指数の意味で類似した文字群に類別することができる。この場合個々の文字の複雑指数が入力文字パターンに混入する雑音に対してこの平面上でどのように変化するか、つまり変化の量とその方向 (Fig. 1 矢印) を調べておき、これを基にして領域の大きさと形 (ほとんどの場合楕円形で近似できる。) を定める。隣接する領域は互いに重なり合うのでほとんどの文字は複数の類に登録されることになる。Fig. 2 はこの方法によって得られた類似文字群の例である。

次に四辺コードは文字パターンの周囲部に存在する文字線の量を抽出するものである。コード化の方法は Fig. 3 に示すように文字の上下左右に設けられた矩形内に存在する文字線の量を 0, 1, 2 の3段階に分け、1文字を4桁のコードで表わすものである。たとえばこの文字においては上側と下側の矩形内には文字線がほとんど存在しないのでコードは0, 左側の矩形内には長い文字線が含まれているのでコードは2, これらの中間の状態にある右側の矩形のコードは1となる。

- ($C_x=8, C_y=19$).....義 建 置 選 語 費 集 壇 層.....
- ($C_x=12, C_y=10$).....被 滅 催 燃 鋼 筋 滅 粧 緒.....
- ($C_x=10, C_y=12$).....務 森 権 僅 鉄 歴 鹿 惑 明.....
- ($C_x=8, C_y=6$).....外 印 門 伯 材 比 ば ぬ あ.....

Fig. 2 Examples of Mutually Similar Character Patterns viewed from a Complexity Index Standpoint.

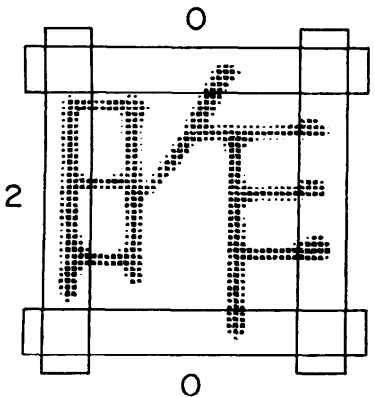


Fig. 3 Surrounding Area Code Extraction Method

* 基準点の設定法としては種々考えられるが、最も簡単な方法はこの平面を格子状に分割し、格子の交点を基準点とするものである。基準点の周囲に文字がほとんど存在しない場合には、適宜、統合を行う必要がある。

- 0021.....仕 任 住 佐 借 俣 枚 舍 含.....
- 1111.....絡 給 勃 統 維 縦 級 緒 鋭.....
- 0121.....逮 道 達 遂 遠 途 連 速 走.....
- 2222.....囚 囚 囚 囚 囚 囚 囚 囚.....

Fig. 4 Examples of Mutually Similar Character Patterns viewed from a Surrounding Area Code Standpoint.

る。従ってこの文字の四辺コードは 0201 となる。矩形の幅は文字線幅の約2倍程度にすることが位置ずれや線幅変動の影響を軽減するために必要である。個々の文字の四辺コードが入力文字パターンに混入する雑音に対してどのように変化するか、つまり取り得る四辺コードの種類 (平均 4~5 種類) を調べ、同一のコードを有する文字を1つの類としてまとめ上げれば、読取対象文字すべてを四辺コードの意味で類似した文字群に類別することができる。この場合もまた個々の文字は複数の類に登録されることになる。Fig. 4 はこの方法によって得られた類似文字群の例である。

以上の方法を用いて読取対象文字の大分類をあらかじめ行っておき、この結果を大分類辞書として記憶しておけば、未知の入力文字パターンの候補となり得る文字を読取対象文字の中から選定することができる。すなわち入力文字パターンの複雑指数と四辺コードを計算し、これらの値を Key として大分類辞書の検索を行い、それぞれの類似文字群を求めて両者に共通する文字を残せばよい。

複雑指数と四辺コードを併用した大分類法⁴⁾の特長はまず第一に効率の高い候補文字の選定を行えることである。複雑指数は文字全体の構造を表わす特徴であり、一方四辺コードは文字の周囲のみに関係した量である。複雑指数によって得られた類似文字群は四辺コードによって得られた類似文字群と互いに独立であると考えられる。特許公報の実データを用いた実験では、複雑指数を用いて得られた類似文字群の平均文字数は190、四辺コードのそれは207に対し、両者を併用した場合は平均文字数を98、つまり読取対象文字数の1/20以下にすることができた。両者の独立性は複雑な漢字の場合に特に顕著であり、候補文字数はおおむね50字内外に納まる。一方ひらがな、カタカナ、英数字など文章中の6~7割を占める文字は漢字に比べて大きさ(文字枠)の小さい文字が多く、四辺コードは0000となることが多い。この場合候補文字数が120字を越えることもある。

第二の特長は雑音に対して安定な候補文字の選定を行えることである。複雑指数はその定義から明らかのように分母、分子の次元は共に長さである。従ってこ

の量は無次元の量であり文字の位置と大きさには無関係である。しかし複雑な文字で、もともと文字線と文字線の間隔が狭い場合には、印字の濃度が濃くなると文字線の間隔が埋まり（文字線のつぶれ）見かけ上文字線が消失し、複雑指数値はつぶれの程度に応じて減少する。文字線の間隔は主として文字の内部に生じることが確かめられている¹⁾。文字の周囲のみに関係した量である四辺コードは文字線の間隔に対しては安定である。しかしコード化を行う位置が大幅にずれるとコードの変化を生じることが明らかである。つまり両者は雑音に対してのふるまいが相補的な関係にあるといえる。

そこで入力文字パターン（平均線幅及び隣接する文字との接続の有無を前処理段階で検出し、それぞれ文字線の間隔の程度、大幅な位置ずれ可能性の判定を行い、許容範囲を超えた場合は2つの類似文字群のいずれか1つを選択的に用いる方法が考えられる。すなわち入力文字パターンの平均線幅が基準値から大きくずれる場合は四辺コードによって得られた類似文字群を用いる。一方入力文字パターンが左右の文字と同時につながっていると判定された場合は、認識対象文字を正しく分離することは非常に困難であるので、この場合は複雑指数によって得られた類似文字群を用いる。

第三の特長は簡単な方法を使って前処理段階で候補文字の選定を行えることである。複雑指数の計算は入力文字パターンの位置正規化の際に、四辺コードの計算は標準化⁵⁾の際に行うことができ、候補文字の選定のために特別な処理時間は必要でない。このことは実際にハードウェアを作成する際に極めて有利な特長であるといえる。以上述べた如く、入力文字パターンの属する類をまず決定し（候補文字の選定）、続いて類内の文字の辞書パターンとのみ照合を行う（個別認識）2段階認識法を用いれば、読取対象文字数が多いことに起因する困難を克服することができる。

2.2 漢字パターンの個別認識

複合類似度法が印刷漢字の認識にとっても有効であることは先に報告した⁶⁾。個々の印刷漢字を類似度を尺度として分類すると、

- 1) 他のいずれの文字とも際立って相違している文字。（兼、射…等全文字の大部分を占める。）
- 2) 複数個の文字と類似している文字。（質、賀、賢、質、買、賛…等）
- 3) 1個または2個の文字と極めて類似している文

字。（微一微、減一減、哀一哀、情一情一慣…等）の3つの型に大別することができる。これらの文字のうち1)の型に属するものは、鮮明に印刷されたものでかつ辞書パターンとの照合が正しい位置で行われれば単純類似度法もしくはそれを変形した方法で充分認識が可能である^{9,10)}。しかし実際の文書に印刷された文字パターンには様々な雑音が混入しており、辞書パターンとの照合を正しい位置で行うこと自体が困難であり、高精度の認識を期待することはできない。

複合類似度法は文字の位置ずれ雑音、文字線の間隔やかすれの雑音に対して安定な認識を行える点が大きな特長である。いま、任意の文字パターンの標準パターン g_0 が与えられたとき、複合類似度法では入力文字パターンに混入する雑音に関する理論モデルを設定し、3つの互いに正規直交した辞書パターン ϕ_0, ϕ_1, ϕ_2 を次のようにして設計する。

$$\phi_0 = \frac{g_0}{\|g_0\|}$$

$$\phi_1 = \frac{\frac{g_1}{\|g_1\|} + \frac{g_2}{\|g_2\|}}{\sqrt{2(1+g)}} \quad g = \frac{g_1 \cdot g_2}{\|g_1\| \cdot \|g_2\|}$$

$$\phi_2 = \frac{\frac{g_1}{\|g_1\|} - \frac{g_2}{\|g_2\|}}{\sqrt{2(1-g)}}$$

但し $g_1 = \frac{\partial}{\partial x} g_0, g_2 = \frac{\partial}{\partial y} g_0, g = \frac{(g_1, g_2)}{\|g_1\| \cdot \|g_2\|}$

次に各文字カテゴリごとに得られた辞書パターンと入力文字パターンとの照合を行い、その出力値から類似度を計算し、最大類似度を示す文字カテゴリ名を答とする。Fig. 5は「絶」の文字について計算され

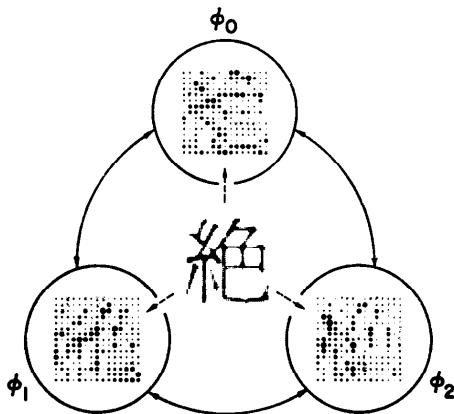


Fig. 5 Triple Reference Pattern based on Multiple Similarity Method

た辞書パターン、 ϕ_0, ϕ_1, ϕ_2 を各点の濃度に対応した大きさの円で示したものである。入力パターン・ベクトルを p 、そのノルムを $\|p\|$ で表わせば、複合類似度 ζ^* は次の式で与えられる。

$$\zeta^* = \sqrt{(\phi_0, p)^2 + (\phi_1, p)^2 + (\phi_2, p)^2} / \|p\|$$

上式の値はある範囲内の位置ずれ雑音に対しては値が変化しないという性質がある。同時に文字線のつぶれやかすれの雑音に対しても値の変化がゆるやかであるという性質も兼ね備えている。しかしこの方法は辞書の設計に際して他の類似した文字の存在を考慮していないため上記 2) または 3) の型の文字に対してはなお一層の改良が望まれていた。混合類似度法⁷⁾は複合類似度法を一般化したものであり、相互に類似した文字であっても精密に識別することができる。この方法の詳細は参考文献に譲ることとして、実験結果のみを報告する。

実際の特許公報を走査して得られたデータに対して、まず複合類似度のみで識別を行い、読取拒否もしくはその恐れのある文字を選択した。次にこれらの文字の辞書パターンを混合類似度法の辞書パターンに置き換え、再度実験を行った。その結果、正読率 99.76%、読取拒否率 0.24%、誤読率 0.00% の認識率が得られた。この実験結果が示すように、複合類似度法と混合類似度法を合わせ用いれば**すべての漢字パターンの高精度な認識が可能である。

3. シミュレーション・システムの 機器構成及びソフトウェア

3.1 機器構成

本シミュレーション・システムはミニコンピュータ TOSBAC-40 C を中心として構成され、実際の文書の読み取りを実時間で行うことができる。これは候補文字の選定を個別認識の前段で行うことにより辞書パターンとの照合回数を大幅に削減したために可能となったものである。本システムの機器構成を Fig. 6 に示す。各機器の主要な性能は次の通りである。

● ITV カメラ図形入力装置

走査領域	前後の文字を一部含めた縦横 70 点
濃淡レベル	6 ビット/点
解像度	56 μm
画素数/文字	縦横 48 点 (9 ポイント活字の場合)

** 本論文のシミュレーション・システムではディスクの容量の制約から混合類似度法は現在まだ組み込まれていない。

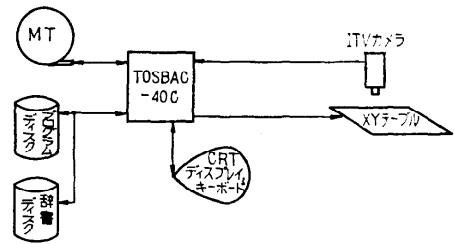


Fig. 6 Simulation System Configuration

● XY テーブル

移動範囲	縦 25 cm, 横 30 cm
送り精度	50 μm /パルス
改行復帰	計算機制御により自動的に行う。

● 辞書ディスク

容量	2.4 M バイト
字体	明朝体及びゴシック体
字種	漢字, ひらがな, カタカナ, 英数字, 記号
辞書パターン数	1,600 字

3.2 ソフトウェア

シミュレーション・システムのプログラムは 20 k バイトのプログラムと 14 k バイトのバッファ領域とからなっている。各部のつながりを Fig. 7 (次頁参照) に示す。

- 前処理部 入力された文字パターンの二値量子化、検出切出し、位置の正規化を行う。二値量子化の方法は各点の濃度とその点の周囲の点の濃度との差を併用したものである⁸⁾。各点を固定の閾値を使って二値化する方法に比べて個々の文字の印字濃度の変動に対して安定な二値化を行える特長がある。検出切出しの方法はあらかじめ与えられた文字配列のピッチを用いて隣接する文字の境界位置を予測、修正しつつ個々の文字の分離を行うものである。文字間のつながりやピッチの部分的変動に対して強い検出切出しを行える特長がある。この段階ではまた入力文字パターンの平均線幅⁹⁾、隣接する文字とのつながりの有無を検出する。
- 標本化部 位置正規化後の 2 値量子化文字パターンを 3 ビットおきに取られた縦横 15

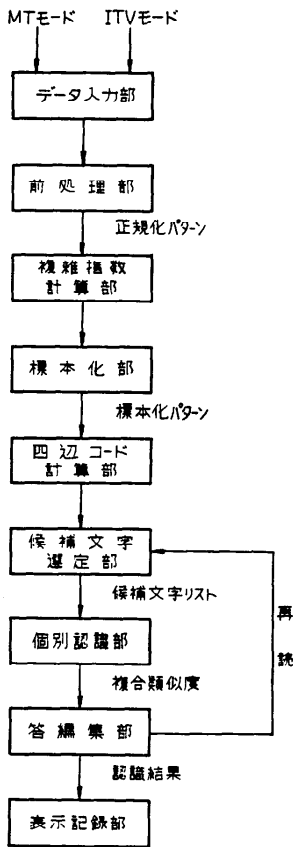


Fig. 7 Simulation Program Flow Chart

点の位置で標本化⁵⁾し、1点6ビットの濃淡パターンに変換する。この操作は一種のフィルタリングであり、次元数の削減(48×48点→15×15点)を行うと共に高周波雑音を除去する効果がある。

- 候補文字選定部 複雑指数、四辺コードそれぞれの類似文字群を記憶した大分類辞書の読み出しを行い、共通する文字を残して入力文字パターンの属する類の決定、すなわち候補文字リストの作成を行う。
- 個別認識部 辞書ディスクから候補文字の辞書パターンを順次読み出し、標本化パターンとの照合を行い、複合類似度を計算する。
- 答編集部 複合類似度のソートを行い、最大類

似度値、最大類似度値と次最大類似度値との差を計算する。これらの値がそれぞれの閾値を超えている場合は、最大類似度に対応した文字名を認識結果として出力する。そうでない場合は候補文字選定部に再読を指令する。再読の場合は入力文字パターンの平均線幅と隣接する文字とのつながりの有無とから、複雑指数または四辺コードの類似文字群のいずれか一方を選択し、個別認識を再度実行する。

4. 読取実験

4.1 実験データ

特許公報の請求範囲の文章を入力データとして読取実験を行った。この理由は本実験が将来漢字OCRを入力機器に用いて日本語のデータベースを作成する際の1つの典型例としての意味を有すると考えたからである。特許公報の文章は、日常使用されることの少ない文字、たとえば又、或、於、附、亘、而、勿、斯、如、且などが頻繁に表われること、使用される文字が技術に関連したものに偏ること、数式・記号が文章中に含まれることなど特有の性質がある。そこで読取対象文字の選定に際しては文字の出現頻度の調査を行い、上位から1,600字を選定した。

4.2 実験結果

特許公報請求範囲の一連の文章の読取実験を行い、認識結果を公報ごとに集計した。読取率の内訳はTable 1に示す通りである。

Fig. 8 (次頁参照)はCRTディスプレイに表示された文字「範」についての実験結果である。読取実験が始まるとまず二値に量子化された入力文字パターン「範」が前後の文字「の」、「罍」の一部と共に左上の枠内に表示される。続いて検出切出し、位置正規化後のパターンが右上の枠内に表示される。次にこのパターンについて計算された複雑指数、標本化パターンから抽出された四辺コード、そして候補文字リストが順次表示される。最後に類似度値の大きい順に上位5位までの文字が表示される。この図の候補文字リストの2行目の左から3番目に入力文字パターンの答、「範」が

Table 1 Recognition Results

正読	読取拒否	誤読	条件
96.4%	2.7%	0.9%	読取対象外の文字を含む。
97.3%	1.9%	0.8%	読取対象外の文字を除く。

存在する。類似度の第一位の文字は「範」、その類似度値は 10、第 2 位の文字は「難」、その類似度値は 2 である。実験結果は正読である。なお、再読の場合はこの画面が消去され変更後の候補文字リストと認識結果のみが新しい画面に表示される。読取拒否となった場合はその旨のメッセージが表示される。

入力文字パターンの走査から最終結果の出力までに要する時間は、候補文字リストの長さに依存するが、約 10~15 秒である。MT モード入力の場合でかつ CRT への表示を一切行わない場合、つまり認識のための計算だけに要する時間は約 4 秒である。

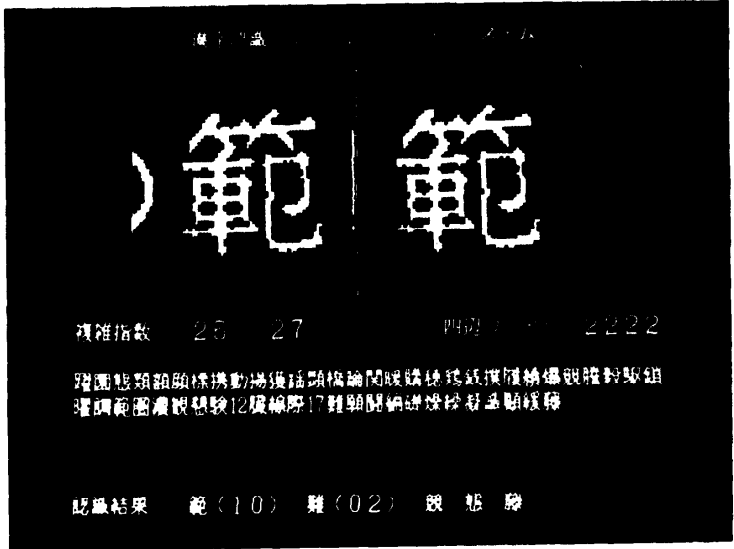


Fig. 8 Experimental Result obtained from the Simulation System

Fig. 9 は実際の特許公報の個々の文字がどのように分離されるかを示すもので、検出切出しの行われた位置に縦棒が挿入されている。相互につながった文字の場合もほぼ正しい位置に分離線が挿入されていることがわかる。

4.3 結果の分析と考察

Table 1 の実験結果の分析と考察を行った。読取拒否の原因の第一は類似文字対の場合に最大類似度値と次最大類似度値が接近することである。(たとえば、すーず、ばーば、べーべ、プーブなど。)この問題に対しては混合類似度を導入することによって解決されることが分かっている。第二の原因は字形の違いである。特許公報の印刷に用いられているタイプ活字の種類は印刷所、発行の日時によってまちまちである。

そして字形の違いは頻繁に取り替えが行われるひらがな、カタカナについて著しい。変形の大きな文字を辞書パターンに追加登録する必要がある。第三の原因は読取対象外の文字が出現した場合である。1,600 字の辞書パターンでは特許公報に現われる文字の 99% を扱えるに過ぎない。実際の漢字 OCR の場合は第二、第三の原因に基づく読取拒否をなくすために最低 3,000 字の辞書パターンを用意する必要がある。

誤読の原因は入力データの低印字品質に基づく字形のくずれ、欠け、つぶれの場合である。実際の文書を扱う限り、この問題を選けて通ることはできない。最

⑤特許請求の範囲

1 隣り合つた間隔が線素の標準幅の少なくとも倍以上の距離を有する田字状に配置された少なくとも 9 本の基本構成線素を予め印刷しておき、この田字状の基本構成線素上に任意の二つの文字が互いに少なくとも二つの線素の位置において線素の有無が異なるように形式化された文字を前記基本構成線素と異なる色により表示し、これを前記基本構成線素に対し所定の位置関係で配置された複数の光電変換素子からなる線素検出マスクにより検出することにより判別するようにした文字読取方式。

Fig. 9 Binary image taken from an Official Patent Gazette

大類似度が低いものについては認識結果の上位数文字を記録しておき、後に文脈情報を用いて正しい文字を選択する方法が考えられている。

特許公報の文体は同一単語の繰り返しが多く、繰り返しの多い文字が上記いずれかの原因で読取拒否もしくは誤読となる場合は、読取率が大幅に低下することになる。現在試作中の漢字 OCR には上記各原因に対する解決方法が組み込まれている。以上述べた如く、本読取実験を通じて鮮明に印刷された個々の文字を単独で扱っている限りでは得ることのできない多くの問題点を明確にし、漢字 OCR の実現に不可欠な基礎資

料を得ることができた。

5. むすび

漢字 OCR の研究開発に役立つミニコン・ベースのシミュレーション・システムを紹介した。漢字 OCR は図画、写真の混在した文書を扱う必要があり、読み取りのフォーマットは頁ごとに異なるのが普通である。このため読取開始に先立ちオペレータがフォーマット指定を行う必要がある。読取速度よりもむしろ文書の種類に応じて辞書の入れ替えを可能とする柔軟なシステムを構成することが重要である。

現在試作が進められている漢字 OCR は Fig. 10 の構成を有し、毎秒 100 字の速度で普通紙に印刷された日本語の文章を読み取ることができる。文脈情報に基づき後処理を用いた場合、読取拒否率は 10^{-3} 以下、誤読率は 10^{-4} 以下の達成が見込まれている。

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」の一環として行われたものである。

参考文献

- 1) 坂井邦夫, 森 健一: 漢字パターンの大分類, 電子通信学会研究会資料, PRL 73-14 (1973).
- 2) 河田 勉, 平井彰一, 森 健一: 漢字認識のための漢字大分類の一手法, 第 14 回情報処理学会全国大会, 114 (1973).
- 3) 坂井邦夫, 森 健一: 漢字パターンの性質, 電子通信学会全国大会, 1557 (1974).
- 4) 河田 勉, 坂井邦夫, 森 健一: 漢字パターン

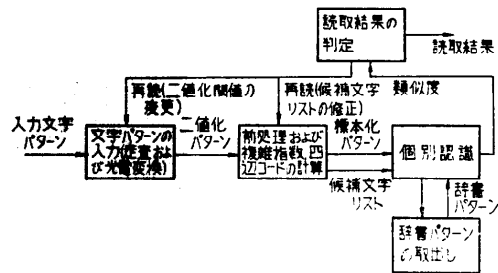


Fig. 10 Block Diagram of Chinese Character Reader under Development

- の大分類, 電子通信学会全国大会, 1556 (1974).
- 5) 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, No. 30 (1972).
- 6) 平井彰一, 森 健一: 漢字認識における単純類似度と複合類似度の比較評価, 電子通信学会全国大会, 1554 (1974).
- 7) 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).
- 8) 河田 勉, 森 健一, 他: 濃淡図形量子化の一手法, 電子通信学会研究会資料, PRL 74-41 (1974).
- 9) 中野康明, 安田道夫, 中田和男: 周波数領域での階層的パターン整合法による漢字認識, 電子通信学会論文誌, Vol. 58-D, No. 2, pp. 94~101 (1975).
- 10) 中野康明, 安田道夫, 中島 晃: 印刷文字認識における位置ずれ補正の効果, 情報処理, Vol. 16, No. 12, pp. 1064~1070 (1975).

(昭和 51 年 3 月 2 日受付)

(昭和 51 年 4 月 5 日再受付)