論　文

帯パターンの時刻変化を用いた印刷漢字認識方式*

藤田孝弥** 中西道明** 宮田清徳**

Abstract

Information processing system using Japanese sentence has been developed in these days, and demand for simple imput device of Japanese sentences including Chinese characters is very strong and urgent. But automatic character recogition systems are not yet in practical use, because Chinese character consists of a large number of categories and their structures are highly complex. So, the most important problem in the recognition of Chinese characters is how to efficiently reduce the candidate categories.

In this paper, we propose a new recognition method, which is based on a new clustering method using a time variation of peripheral belt pattern and binary pattern matching method. In computer simulation test, a recognition rate of 99.1% is obtained. As a result, it has been proved that we can realize a Chinese character recognition system using our recognition method.

1. まえがき

日本語情報を計算機で直接扱うために必要な漢字情報処理システムの開発が近年さかんに行われている．しかし，漢字情報処理を行うにあたって最大の問題点は，漢字データをいかに入力するかであり，現状では漢字けん盤装置がほとんど唯一のものであるために，専門のオペレータを必要とし，なおかつ処理速度が遅いという欠点がある．そこで漢字情報処理を行うためには，漢字入力の自動化が必須であり，各所で漢字認識の研究が進められている[1)~4)]．

しかし，漢字認識は単一字体の印刷文字に限っても対象文字数（カテゴリー数）が多く，通常の用途でも2,000 文字の認識が必要であり，また，漢字が英数字に比べて複雑なために．一文字を表現するためのパターンのメッシュが数倍必要となる．そこで，漢字認識においては，英数字の数百倍の情報量を扱わなければならず，標準パターンの記憶容量，認識ハードウェアにもこの違いが効いてくるので，漢字認識システムの構成を考える上では，この点について十分考慮する必要がある[1),2)]．

ところで，単一字体の印刷文字の認識手法としては，パターン整合法が安定な認識法として確立されており，英数字，カタカナ等の認識においては十分にその性能が実証されている．しかしながら，漢字認識においてこの方法を直接用いようとすると，漢字のもつ情報量が英数字の数百倍にものぼるために，ハードウェアの増大ないしは，認識処理速度の低下はさけられず，また，標準パターンメモリも高価なものとなる．そこで，漢字認識においては，情報量を圧縮して特徴量での標準パターンを記憶しておき，メモリー容量を減らすと同時に処理速度の低下をさけるという方法がひとつには考えられる[3)]．しかし，情報量を少しでも失った状態である特徴量でのマッチングで最終的な識別を行うことは，対象文字が少ない場合や，外字の出現を想定しない場合には良いが，対象文字数が 2,000 文字程度に増えてくると，情報量の圧縮をいかに有効に行ったとはいえ，元の情報を完全に保つことはできないために識別が困難になったり，あるいは，外字を誤読せずに，確実にリジェクトとすることができなくなる[1)]．そこで，識別部でのパターン整合は，特徴量でのマッチングではなしに，文字のもつ全情報を保っ

* Recognition of Printed Chinese characters using time variation of peripheral belt pattern by Takaya FUJITA. Michiaki NAKANISHI and Kiyonori MIYATA (Fujitsu Laboratories Ltd.)

** (株)富士通研究所

た状態である全メッシュの2値パターンとの間で行うことにより，最終的な識別部での判定を確実なものとすることが望ましい．また，位置ずれ等の補正を考えた場合にも，特徴量に変換された標準パターンを持っておくよりは，元のパターンのまま記憶しておいたほうが扱いやすい面が多い．

以上の観点から漢字認識においても全メッシュの2値パターンマッチングを用いる事を考えると，処理速度の低下をいかにさけるかということが実用的な漢字認識のシステムを構成する上で重要となり，その点を考慮した認識方式を用いなければならない．

そこで，漢字認識においては，いかに有効に漢字文字を分類し候補文字を絞るか，ということが最も重要となり，そのために安定な分類方式が必要である．ところが分類方式として複雑な演算処理を行ったり，多くのメモリを必要とする方法や，文字のかすれ，ぼけ等によって分類が不安定となりオーバーラップが多くなることはできるだけさける必要がある．そのため，我々は，文字のかすれ，ぼけ等の変形に対しても安定であり，また装置化の容易な分類方式を新たに開発した．本方式は漢字文字が持っている有効で安定な情報量の多く含まれている輪郭部分に注目し，文字パターンを段階的に太らせながら作る**帯パターンの時刻変化**を用いることによって，文字の持つ平面的な特徴を文字枠上での時刻的な変化としてとらえており，芯線すら確保されないような欠けをのぞいては，太め処理によってかすれた文字パターンが再現されるために安定でオーバーラップの少ない分類方式となっている．

本方式を用いたシミュレーションとして，12 ポイント明朝体活字約 2,000 種，1万文字についての認識の結果，正読率 99.1%，読取不能率 0.77%，誤読率 0.13% の結果を得た．

2. 漢字認識システムの構成

本システムの特徴は，前章でのべたように識別能力を十分に上げ，かつ安定でオーバーラップの少ない分類方式を採用したことにあり，**Fig. 1** に示すように，認識回路はすべて専用のハードウェアで実現し，マッチング処理の高速化をはかっている．また，標準パターンを記憶しておくメモリは低価格のミニディスク装置（PF 6036 A，1M バイト）を用いることにより，メモリ容量の増大に対処できるようにしている．

構成は大きくわけて，観測部，認識回路部，及び外部記憶装置部とからなり，観測部から得られた黒白2

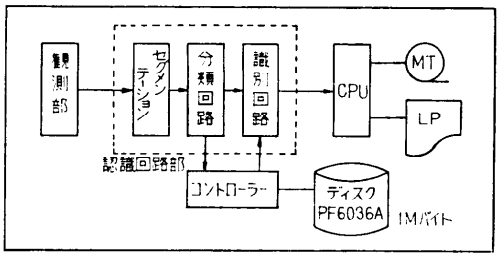


Fig. 1 Block Diagram of Chinese Character Recognition System

値の映像信号は**セグメンテーション回路**で1文字ずつに区分された後，**分類回路**に入り本認識方式の特徴である**帯パターンの時刻変化**を用いた分類によって，候補文字グループの番号が決定される．次のディスクコントロール回路では分類回路からのグループ番号にしたがってディスクのトラックを選択し，標準パターン及び文字コード等を識別回路へ転送している．**識別回路**では，ディスクからの2値パターンとのマッチングを行い，判定基準にしたがって識別結果の文字コードを CPU 側に転送している．

本論文で主にのべる認識回路部は，セグメンテーション回路，分類回路，及び識別回路からなり，セグメンテーション回路では，固定ピッチで印刷されていない場合にも適用できる方式を考案している．**セグメンテーション方式**としては，横方向の連続性のある線分のあつまりを1セグメントとして切り出し，その文字幅が標準文字寸法に満たない場合には，横幅の狭い文字（日，目，あるいは数字，記号等）なのか，あるいは分離文字の片側（例えば「化」の偏(人)）なのかの区別をつけるために，横幅の狭い文字グループの標準パターンとのマッチングをとり，対象文字がみつからなかった場合は分離文字の1部とみなして，さらに次のセグメントまでを1文字として切り出す方法を用いている．次に候補文字を絞るための**分類方式**としては，**帯パターンの時刻変化**による分類方式を考案しており，複雑な特徴抽出や演算処理は行わずハードウェア実現が容易であって，文字のかすれ，ぼけ等に対しても安定な特徴をとらえるために，文字パターンを段階的に太らせながら文字枠部分で形成する帯パターンの時刻変化によって安定な輪郭情報をとり出して分類を行っている．**識別方式**は，標準パターンと入力パターンとの単純な排他的論理和を不一致数とする方法でハードウェアの簡単化をはかっており，外字等をリジェクトにするための**判定基準**としては標準パターンの黒

白境界点数を用いて識別を行っている．

3. 認識方式

3.1 セグメンテーション回路

文字認識においては，個々の文字を認識する前に1文字ずつに区別するためのセグメンテーションが重要な問題となる．特に印刷文字の場合には，手書文字のようにドロップアウトカラーで印刷された文字枠や，リファレンスマークのようなものを想定することができない．また，固定ピッチで文字が印刷されているとはかぎらず，特に隙間なく詰めて印刷されている場合には，英数字のように非分離文字だけでなく，上下左右に複数のセグメントに分かれた文字もあるために，活字文字とはいえ横幅の狭い文字や記号等と分離文字との区別をつけることは非常に困難である．

そこで，我々が用いているセグメンテーション方式は，あらかじめ横幅の狭い文字や記号等の標準パターンだけはP-ROMのような内部メモリにもち，横方向への文字の連続性を検出することによって，1セグメントごとに横幅の狭い文字グループの標準パターンとのマッチングをとり，分離文字か否かを判別している．回路構成は **Fig. 2** に示すようになっており，観測部で縦方向に順次走査して得られる黒白2値の文字映像信号をバッファ1に書き込みながら，連続点検出回路において入力文字パターンを縦方向に投影し，横方向への文字の連続性を検出している．そして，連続点の途切れた所で文字映像信号の転送及びバッファ1への書き込みを停止する．このようにして得られた1セグメントの入力文字信号をバッファ1から読み出しながら，位置決め回路で文字パターンの上下左右の位置を検出し，中心合わせを行った状態でバッファ2への書き込みを行う．このバッファ2に書かれた入力文字パターンは，文字の横幅寸法が標準寸法以上であれば，バッファ1への書き込みアドレスカウンタをリセットした上で，分類回路へ送られるが，分離文字の片側もしくは横幅の狭い文字あるいは記号の場合には，あらかじめこのセグメンテーション回路内に設けられている横幅のせまい文字グループ（漢字14文字及び数字，記号）の標準パターンとのマッチングを行い，対象文字がみつからない場合には分離文字の一部とみなして，さらに観測部から文字映像信号を受け取り，バッファ1に追加書き込みを行う．

このようなセグメンテーション方式を用いることにより，分離文字についても標準文字と同一に処理でき，

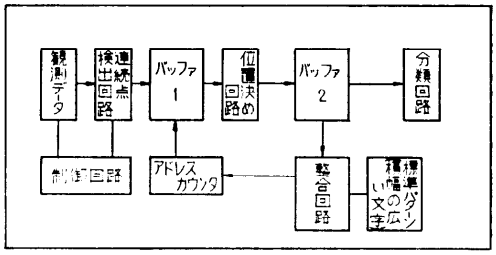


Fig. 2 Segmentation block

また横幅の狭い文字や記号等の少しの文字パターンだけをP-ROMのような内部メモリに用意しておけば，どのようなピッチで印刷されていても，また，数字，記号等が混在していてもセグメンテーションを正しく行うことができる．

3.2 分類回路

3.2.1 帯パターンの作成

漢字認識方式において最も重要な要素は，文字の分類をいかに効率よく行い，候補文字を絞るかということにある．そこで，あらかじめ文字カテゴリーをいくつかの候補文字グループに分類しておき，そのグループごとに標準パターンを記憶装置内に格納しておく．こうすることで，ディスクのようにアクセスに時間のかかるメモリであっても一度のアクセスでグループ内のすべての標準パターンを呼び出すことができる．しかし，P-ROMのような半導体メモリやコアメモリのようにランダムアクセスすることができないので，同一のカテゴリーが複数のグループにわたって存在すること，つまり，オーバーラップをなるべく少なくすることが必要である．そのため，文字のかすれ，ぼけ等の変形に対しても安定な特徴によって分類する必要がある．

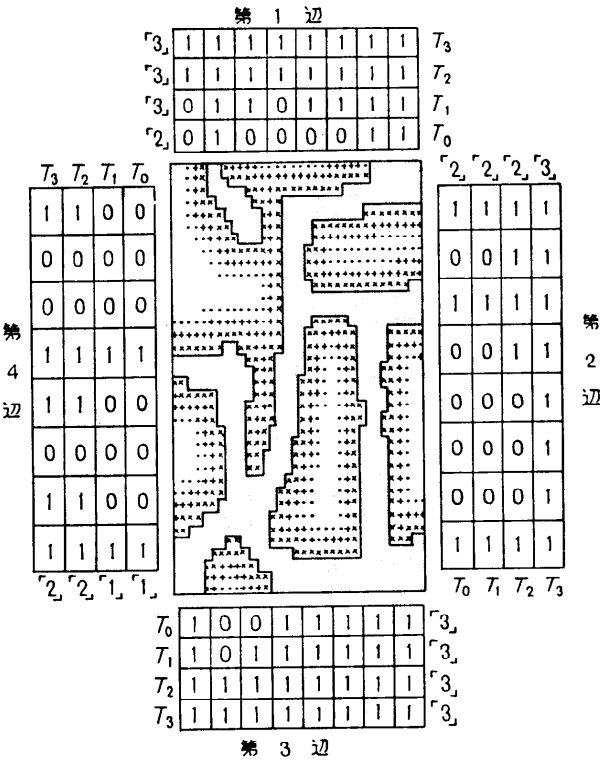
以上の観点から，我々は新しい分類方式として，漢字文字のつぶれに対しても安定であり，有効な情報を多く含んでいる文字の輪郭部分に注目し，**帯パターン**と称するものを作成している．この帯パターンは文字パターンを段階的に太らせながら文字枠上で作成する圧縮パターンであり，文字枠という一次元的な所から文字のもつ平面的な輪郭情報を時刻変化として取り出すことができ，ハードウェアの実現を容易なものとしている．

Fig. 3(次頁参照)が帯パターンの一例である．図中黒の部分が観測された「近」という字の原パターンであり，*印，＋印，●印が，原パターンに太め処理を施して段階的に文字が太められた時の状態である．こ

の入力パターンでの文字の外接枠位置から4メッシュ中までを含む帯状の部分パターンを，入力および3段階の太めパターンのそれぞれから切り取り，それぞれの枠の長さ方向に4メッシュごとに区切り，この 4×4 メッシュ単位ごとに黒が多いか白が多いかによって，“1”，“0”に変換する．これが図で文字の外側に示されたパターンで本方式で用いる帯パターンであり，$T_0-T_1-T_2-T_3$と太め処理を行うごとに上下左右の文字枠上で作られる**帯パターンの時刻変化**の様子を示している．

このようにして，1辺8ビットずつ4時刻，計128ビットの帯パターンが作られる．

帯パターンの**特徴**をながめてみると，Fig. 3の例からもわかるように，原パターンから作った帯パターン（T_0）からだけでは第1辺に表れている白の部分も，第4辺の白も同等な意味しかないが，時刻変化している方向に帯パターンをながめてみると，第1辺の“0”はT_3時刻には“1”に変わっているが，第4辺では最終のT_3時刻まで“0”でのこっている．つまり，原パターンT_0から作られた帯パターンだけでは文字の輪郭部分のもつ平面的な情報が含まれておらず，同じ“0”でも安定で深さ情報を持ったものなのか，不安定で文字のぼけ等によって“1”に変わりうるようなものかの区別がつかない．しかし，太め処理による時刻変化としての帯パターンを調べることにより，原パターンから見ただけでは表れない平面的な**輪郭情報**を，文字枠という一次元的な所で取り出すことができ，かなりのかすれやぼけがあっても，また多少の欠けがあっても，芯線すら確保されないような欠けでなければ，太め処理によってパターンが再現されてくるために安定で有効な特徴をもったパターンとなっている．**Fig. 4**（次頁参照）では最上段の2値化入力パターンから得られた帯パターンとその模式図およびグループ番号と，このグループ番号を指定した全体コードの時刻変化を示してある．全体コードは各時刻ごとの4つの枠の帯コードの組み合わせで作ったもの（**Table 1** 次頁参照）であり，詳細は後述する．



全体コードの時刻変化

$$T_0-T_1-T_2-T_3 = \boxed{7}-\boxed{14}-\boxed{14}-\boxed{19}$$

Fig. 3 Input pattern and Belt pattern

3.2.2 全体コードの時刻変化による分類

帯パターンの時刻変化から文字を分類するためには128ビットの帯パターンからさらに分類用のコードを付ける必要がある．ここでは，まず1つの帯ごとに帯コードを付けている．**帯コード**としては帯パターン上の黒又は白の連続量をとらえたのでは，少しの欠け等に対しても不安定になるので，各帯ごとに黒の絶対量が多いのか，白が多いのか，半々かによって，それぞれ「3」，「1」，「2」の3段階に帯コードを付けている．Fig. 3 の例では，T_0 時刻では第1辺，第2辺は「2」，第3辺は「3」，第4辺は「1」と帯コードが付けられている．さらに，この帯コードを4辺あわせて1から 20 までの全体コードを付けている．**全体コード**は Table 1 に示すように，帯コード「3」の表れる位置によって分けられ，4辺とも黒が多く帯コードが「3」の場合には，全体コードとしては⑳となり，図の T_0 の例では，第3辺のみ帯コード「3」であるので，全体コードは⑦となる．同様にして，全体コードを各時刻ごとに付けると，⑦—⑭—⑭—⑳と変化する**全体コードの時刻変化**が形成される．

全体コードは表のように 20 種類あるが，全体コードの時刻変化の組み合わせは，625 通りとあまり多く

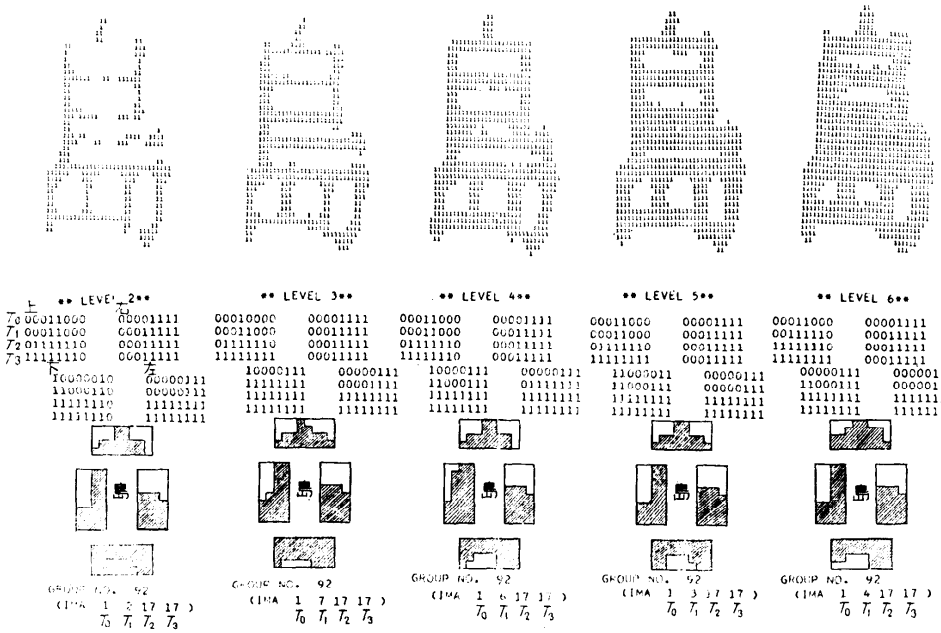


Fig. 4 Example of digitized input patterns and belt patterns

Table 1 Code Table

全体コード	帯コード	帯コード「3」の位置	(例)
1	「1」が4辺		
2	「2」が1辺		
3	「2」が2辺		
4	「2」が3辺		
5	「2」が4辺		
6	「3」が1辺		院
7			土
8			刹
9			天
10	「3」が2辺		
11			向
12			
13			
14			工
15			司
16	「3」が3辺		凶
17			区
18			岡
19			
20	「3」が4辺		国

はない．これは，各時刻の帯パターンが入力パターンの太め処理ごとに作られるために，全体コードが前の時刻より小さな値になることはなく（$T_i \leq T_{i+1}$），また表からもわかるように⑭—⑯，⑯—⑱というような変化も存在せず，全体コードの①から⑤までを1まとめにしているために，T_0 から T_1 への変化は 81 通りであり，T_2 までの変化は 256 通り，T_3 までの変化は625通りの組み合わせが存在する．今回の実験での，標準パターンおよびその想定される変形の全体コードは，**Fig. 5**（次頁参照）に示すように最大グループ数の時でもオーバーラップが非常に少なくなっている．これは，帯パターンを図形的（連続量的）に取り扱わずに，白黒の絶対量として扱い，コードを作成しているからである．またコードは規則性をもっているので，取り扱う文字数や標準パターン格納のデバイス等によって，分割数を決め，対象文字のコードの分布状態を合わせて，625 通りを幾つかずつ組み合わせ，少ないグループ数にしてオーバーラップを減少させることができる．

本実験のシステムでは，8×16＝128 トラックのミニディスクを標準パターン格納デバイスとしているので，グループ数としては，128，64，ないしは 32 に分割することが望ましい．コードの分布と規則性から，

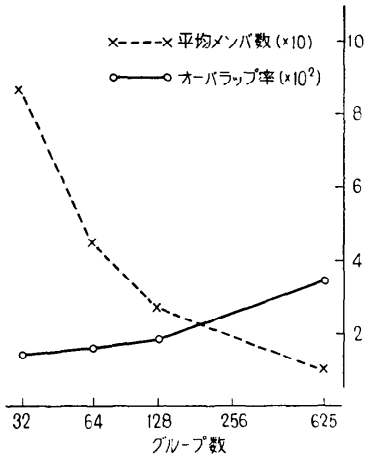


Fig. 5 Number of group vs. overlap

各グループ数に 625 通りを組み合わせて作った場合のそれぞれの平均メンバ数とオーバーラップ率を示したのが Fig. 5 である.

3.3 識別回路

文字識別のための最終決定は識別部でのパターンマッチングにより行われる. そこで, この識別回路の性能が認識率を確保する上で重要な要素となる. そこで本方式では識別部での能力を十分発揮できるように, 漢字のもつ情報量を失ったような特徴量でのマッチングではなしに, その文字のもつすべての情報を保っている全メッシュの 2 値パターンマッチングを行うことで, 2,000 文字の識別を可能なものとしている. またマッチングはセンターマッチングを基本としているが, 位置ずれ補正等のために, 従来から数字, アルファベット等で用いてきた 4 方向にシフトした信号とのマッチングも同時に行っている.

類似度の計算は, ハードウェアの容易なことから, 標準パターンと入力パターンとの排他的論理和を計数することで不一致数を求めている. この不一致数により認識あるいはリジェクトの判定を行っているが, 不一致数の一番少ないものを単純に取り出すことや, 次大類似度との差だけを判定基準とすることだけでは外字を正しくリジェクトにすることができない. またリジェクトを作るための基準を一定の閾値にとると, 同一文字間においても文字の複雑さによって不一致数が異なるために適当ではない. **Table 2** がその一例であり線幅変動等をもった同一カテゴリーの 5 種類の文字について「一」及び「園」の場合の不一致数を示した

Table 2 Example of "exclusive-oR" point

	2	3	4	5	6
一	58	29	15	38	41
園	385	268	196	273	356

ものである.「一」のように簡単な文字と「園」のように複雑な文字とを比べると, それぞれの文字の線幅変動等の変形が同程度であっても, 不一致数の絶対値がかなり違っており, リジェクトを作るための閾値を一定にすることはできない. そこで, 判定基準を一様にするために何らかの正規化操作が必要であり, ここでは不一致が出やすい部分が文字の境界部分で起こっていることから, **Fig. 6** のように標準パターンそれぞれの境界点数をあらかじめカウントし, 標準のメッシュパターンとともにディスクに書き込んでおき, この判定基準以上の不一致のある文字はリジェクトすることで外字の処理を行っている. また, この基準以内に入るカテゴリーが 2 文字以上になる場合には, 別に設けてある補助特徴回路からの信号で, リジェクトないし判別を行っている.

Fig. 7 は認識回路部のフローであり, 入力観測文字は太め処理回路及び帯パターン作成回路を通って全体コードが形成され, この全体コードの変化によりグループ番号が作成される. このグループ番号に基づいてトラック位置を指定し, ディスク制御回路では選択されたトラックから標準パターンを識別回路に送り入

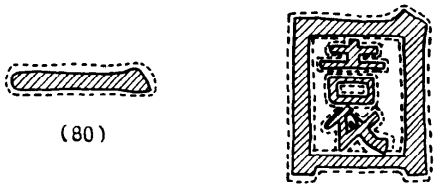


Fig. 6 Number of border point

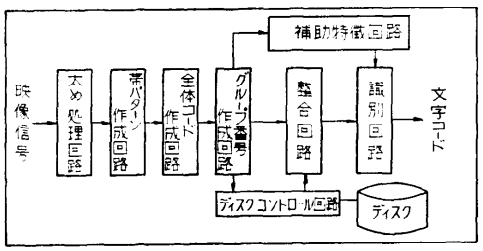


Fig. 7 Recognition flow

力パターンと標準パターンとのマッチングを，中心及び4方向それぞれ独立に行うと同時に，判定基準及び文字コードを抽出し，補助特徴回路からの信号をも参照して判別を行っている．ここで用いている補助特徴とは，あらかじめ類似文字が1つ以上存在することが解っている文字について，それらの文字間での相違点で，黒(文字部分)，白（空白）でそれぞれあるべき点と，白でも黒でも文字間に差がなくて，どちらでもよい(Don't Care の）点の3種類に分けて，各文字ごとにこの3種類の点におきかえた差分マスクを作り，その文字が候補文字となった時に，この差分マスクの Don't Care 以外の点で入力パターンと不一致があった場合に，その文字を候補文字から除外するもので，類似文字の存在しない文字の場合は補助特徴回路からの信号は無視される．

4. 識別結果

実験で用いた文字サンプルは 12 ポイントの邦文タイプライタの明朝体当用漢字等約2,000文字である．標準パターンとしては，あらかじめ1文字あたり5回の印字を行ったものをそれぞれ2値化し，5つのサンプルを各点ごとに重ね合わせ，その平均をとったものを用いた．標本点数は標準の文字で平均 38×38 メッシュ程度であり，文字の横幅が 29 メッシュ以下のものは横幅の狭い文字グループとした．分類情報としては標準パターンの帯パターンからの全体コードおよびコード作成時の閾値の前後を変形として全体コードを追加し，その他の文字図形上の想定変形に対する全体コードを若干含めたものを基礎データとして用いた．本実験でのデータサンプルは比較的新しい活字で印字したものと，別の古い活字で印字したものとを用いてさらに光電変換後の2値化レベルを変えることにより1カテゴリーあたり5データとし約1万データについて識別実験を行った．結果は **Table 3** のようであり，総合の認識率は 99.1%，リジェクト率 0.77%，エラー 0.13% であった．データの2はかなりつぶれた古い活字のものであり，レベルを下げた状態(レベル3)の時にリジェクトが多くなっているが，2値化レベルを下げることと，実際の印刷物上での印字の欠けとが，かなり違ったものになる場合のあることを示している．これらは分類情報作成時点での想定変形に若干の追加をすることで減少させられる．リジェクトはこの他，類似文字対のための補助特徴である差分マスクが今回の実験段階では，まだ不充分であったために，候補カテゴリーをひとつに絞れなかった場合であり，差分マスクの追加で容易に減少させ得る性質のものである．またエラーとなったものは，差分マスクの用意されていなかった文字対の相手の文字だけが判定基準内に残り，補助特徴回路からの信号がなくリジェクトにできなかったもので，前述の差分マスクの追加によるリジェクトの原因除去と平行して減少する性質のものである．

Table3 Recognition Rate

Data No	Level	Recognition rate	Rejection rate	Errer rate
1	3	1,880 (99.05)	12 (0.63)	6 (0.32)
	4	1,890 (99.6)	8 (0.42)	0 (0.0)
	5	1,886 (99.4)	11 (0.58)	1 (0.05)
2	3	1,872 (98.6)	24 (1.26)	2 (0.1)
	4	1,877 (98.9)	18 (0.95)	3 (0.16)
Total		9,405 (99.1)	73 (0.77)	12 (0.13)

( ) %
Vocabulary; 1898 Chinese characters.

5. むすび

漢字認識システムとして，文字パターンのもつ平面的な輪郭情報を安定に取り出すために，文字パターンを段階的に太らせながら作る帯パターンの時刻変化としてとらえる分類方式を考案し，文字のかすれ，ぼけ等に対しても安定でオーバーラップの少ない分類が可能となった．また，識別部でのパターンマッチングは2値の全メッシュのマッチングを用いることで，認識率の向上がはかられ，実用的な漢字認識システムが可能であることがわかった．今後は，さらにデータを蓄積し，補助特徴回路をさらに改良することで，読取不能文字あるいは誤読文字を減らすことができるものと考えられる．

最後に，日頃御指導いただく，電子研究部 宮川部長，徳永部長代理 ならびに社内関係各位に深謝いたします．

参考文献

1) 藤田，中西，宮田：印刷漢字の認識方法，電子通信学会，パターン認識と学習研究会資料，PRL 75-38 (1975).
2) 藤田，中西，宮田：印刷漢字の識別実験，電子通信学会，パターン認識と学習研究会資料，PRL 75-63 (1975).
3) 中野，中田：周辺分布とそのスペクトルによる漢字の認識，電子通信学会論文誌，Vol. 56-D, No. 3, pp. 146～153 (1973).
4) 坂井，森：漢字パターンの大分類，電子通信学会，パターン認識と学習研究会資料，PRL 73-14 (1973).

（昭和51年2月20日受付）
（昭和51年4月14日再受付）