多重辞書類似度法による手書き類似漢字の識別実験

大倉　充*・藤原敬己**・塩野　充***

*岡山理科大学大学院理学研究科

博士課程システム科学専攻

**㈱三菱電機東部コンピュータシステム

***岡山理科大学工学部電子工学科

（1989年9月30日受理）

1．まえがき

手書き漢字は，字種（カテゴリー）が非常に多く，しかも形状の類似した文字パターンも数多く存在する。また筆記者の書き癖による文字パターンの変形（手書き歪み）が多様である。このような性質を持つ手書き漢字を認識するために考えられた手法の一つとして，各カテゴリーに用意する標準（辞書）パターンの複数化（以後，多重辞書パターンと呼ぶ）があり，種々の報告が成されている[1)-3)]。筆者らも「多重辞書類似度法」という多重辞書パターンを用いる認識手法を提案し，報告を行ってきた[4)-6)]。

多重辞書パターンを用いる最も大きな目的は，上述の手書き歪みの吸収にある。しかしながら，手書き漢字の認識を困難とする主な要因の一つである，類似カテゴリー（類似カテゴリーの定義はかなり困難な問題であり，本論文では，筆者らの主観で似ていると思われるカテゴリー群を指すこととする）の存在に対して，多重辞書類似度法が効果的に対処できるかどうか明確ではない。なぜならば，類似カテゴリーの識別の困難さと手書き歪みの存在との間には，密接な関係のあることが推測され，多重辞書パターンを用いて手書き歪みに対処するということは，逆に類似カテゴリーの識別を妨げることに結びつく可能性がある。

本論文では，多重辞書類似度法の類似カテゴリーの識別能力を調査するために，電子技術総合研究所（以後，電総研と記す）作成のJIS第1水準手書き漢字データベースETL−9(B2)[7)]より7組（4カテゴリー／組）の類似カテゴリー群を選出し[8)]，各カテゴリーに対する多重辞書パターン数，多重辞書パターンを構成する学習サンプル数及び，文字パターンの画面次数の3種類のパラメータを変化させて識別実験を行った。さらに，カテゴリーを増加させた場合の，上述の類似カテゴリー群の分類実験を行い，最も代表的な類似度法である単純類似度法[9)]により得られた結果との比較を行った。

2. 多重辞書類似度法

2.1 多重辞書類似度法

多重辞書類似度法とは，各カテゴリー$C^{(i)}$（$i=1$，…，K：Kは全カテゴリー数）に対して多重辞書パターン$M_j^{(i)}$（$j=1$，…，h，…，n：nは一つのカテゴリーに対して用意された多重辞書パターン数）を用意し，入力されたサンプルXとの間で式(1)に示す類似度[9]を計算し，

$$r_j^{(i)}=\frac{(X,M_j^{(i)})}{\|X\|\cdot\|M_j^{(i)}\|} \tag{1}$$

式(2)によって，決定カテゴリー$C^{(d)}$を定める手法である。

$$r_h^{(d)}=\max_{i,j} r_j^{(i)} \tag{2}$$

2.2 多重辞書パターン作成法

多重辞書類似度法では，文字パターンより構造情報を担った特徴量の抽出は行われず，原文字パターンそのものが多重辞書パターン作成及び認識に用いられる。そのため，多重辞書パターンの作成法がこの手法の認識能力を左右することになる。

本論文では，一つのカテゴリーに所属するサンプルに対してクラスタリングを施し，類似したサンプル（ここでの”類似”は，黒点分布形状の似かよったという意味で用いている）同士でクラスターを形成した後に，各クラスター内のサンプルをすべて加え合わせることによって，濃度レベルを持った辞書パターンを得るという方法を採用した。

クラスタリング手法としては，Ward法を用いている[5),6)]。Ward法は，各サンプルが類似しているものから順次集められ，クラスタリングの過程において，2つのクラスター間でサンプルの入れ替えが生じない階層的手法[10)]の一つである。Ward法では，クラスター間の距離は次のように定められる。クラスター$L^{(a)}$に属する全てのサンプルについての，クラスター重心（クラスター$L^{(a)}$に所属するサンプルの平均サンプル）からの偏差2乗和$I^{(a)}$を情報損失量と定義する。クラスター$L^{(a)}$と$L^{(b)}$が併合されて，新しいクラスター$L^{(ab)}$となるときの情報損失量の増加量$\Delta I^{(ab)}$をクラスター$L^{(a)}$,$L^{(b)}$間の距離と定める。

$$\Delta I^{(ab)}=I^{(ab)}-I^{(a)}-I^{(b)} \tag{3}$$

従ってWard法では，情報損失量がその一つのクラスター内のサンプルのまとまりの度合いを示すため，情報損失量の増加量を最小化するようなクラスターを選択し，それらを結合させていけばよい。

3. 認識実験

3.1 実験に使用したデータ

実験に使用したデータは，電総研のETL－9(B2)[7]である。このデータベースは，2値化及び，位置と大きさの正規化済みの文字パターンから構成されているため，そのまま重ね合せ的手法に適用できる。全200データセットから成り，1データセットにひらがなを含めJIS第1水準漢字3036カテゴリーが収納されている。実験に際し，1カテゴリー200サンプルがデータセット番号の順番に並ぶように編集し直して用いた。なお，画面次数は63×64である。

類似カテゴリーとして，表Ⅰに示す7組（4カテゴリー／組）を選出した[8]。組番号①～③は，遍と旁のそれぞれ2種類の組み合せで各カテゴリーが表現できるもの[11]を，④は，形状自体が似ていると筆者らが判断したものを，⑤～⑦は，構え，旁，遍がそれぞれ同じものを選んでいる。また，この類似カテゴリー群の大分類率の調査のために，多重辞書パターンを作成した100カテゴリーを表Ⅱに示す。これらは，漢字JISコードの先頭からの100カテゴリーで，表Ⅰに示したカテゴリーと重複したものはない。

（表Ⅰ）　類似カテゴリーの組

組番号	類似カテゴリー			
①	諭	輸	輪	論
②	詰	結	紹	詔
③	渇	掲	湯	揚
④	丑	五	互	瓦
⑤	閏	開	閑	関
⑥	狙	祖	租	粗
⑦	鏡	鎖	鐘	鎮

（表Ⅱ）　大分類率調査のために追加したカテゴリー

亜	啞	娃	阿	哀	愛	挨	姶	逢	葵
茜	穐	悪	握	渥	旭	葦	芦	鯵	梓
圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎
或	粟	袷	安	庵	按	暗	案	闇	鞍
杏	以	伊	位	依	偉	囲	夷	委	威
尉	惟	意	慰	易	椅	為	畏	異	移
維	緯	胃	萎	衣	謂	違	遺	医	井
亥	域	育	郁	磯	一	壱	溢	逸	稲
茨	芋	鰯	允	印	咽	員	因	姻	引
飲	淫	胤	蔭	院	陰	隠	韻	吋	右

3.2 実験概要

まず，各組における類似カテゴリーの識別実験を行った。実験条件として，次に示す3種類のパラメータを設定した。

（Ⅰ）各カテゴリーあたりの多重辞書パターン数**n**：本論文では，$1 \leqq n \leqq 20$を用いた。各カテゴリーに対する**n**は同数である。なお，$n=1$の場合は，通常の単純類似度法[9]を意味する。

（Ⅱ）多重辞書パターンを構成する学習サンプル数**m**：本論文では，$m=50,100,150$を用いた。それぞれ，データセット番号1～50，1～100，1～150のサンプルを学習サンプルとしている。従ってその各々の場合に対して，未知サンプルのデータセット番号は，51～200，101～200，151～200である。

（Ⅲ）サンプルの画面次数**p**：オリジナルのサンプルは，63×64の**p**を持つ。本論文では，オリジナルの画面に1行（64個の0）付け加え，64×64の画面にした後に，32×32，16×16，8×8の3種類の**p**を持つ縮小画面を新たに作成した。次数変換の方法は，多重辞書パターン，入力サンプル共に同じで，2×2の4画素を次の縮小画面の1画素に対応させ，4画素値の合計値を縮小画面の1画素に持たせた。従って縮小画面を持つ入力サンプルは，多重辞書パターンと同じく濃度レベルを持つ。なお，縮小画面を持つ多重辞書パターンは，63×64の画面において得られたものを基にして，上述の方法で単純に作成したものであるため，縮小画面を持たせた学習サンプルに対して，クラスタリングを施して得られるものとでは異なる可能性がある。

次に，各組ごとに類似カテゴリーの大分類率の調査を行った。用いた（多重辞書パターンを作成した）カテゴリーは，表Ⅰ，Ⅱに示した128カテゴリーである（K=128)。本論文では，$n=10$，$m=100$，$p=63\times64$の条件下で行った認識実験結果を示す。比較のために，単純類似度法による結果も併せて示す。単純類似度法での辞書パターンは，学習サンプル（$m=100$）を全て加え合わせて作成した。

なお本論文では，上述の全ての認識実験において，認識不能と判断するリジェクト設定は行わず，正読と誤読のみとしている。

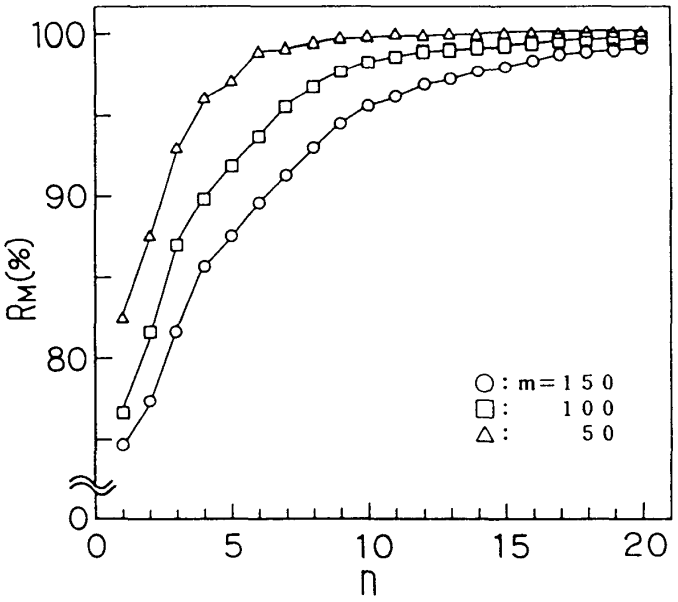
実験に使用した計算機は，本学情報処理センターのFACOM／M380で，使用した言語は，FORTRAN77である。

3.3 認識実験結果

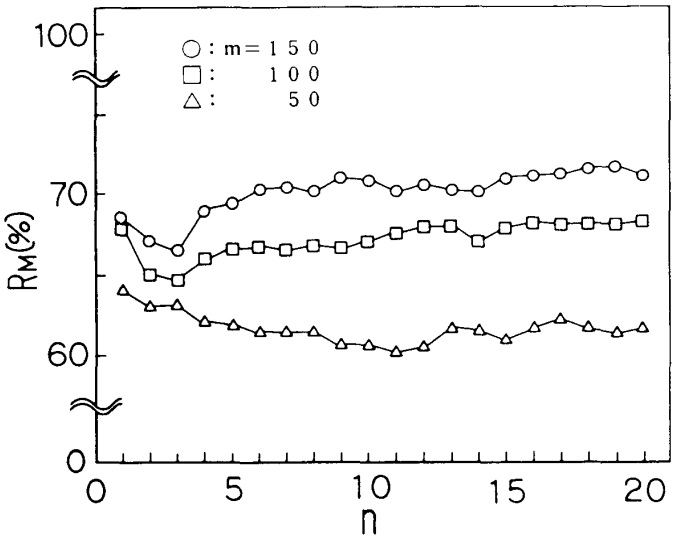
3.3.1 画面次数固定（$p=63\times64$）

まず，$p=63\times64$に画面次数を固定した場合の結果について述べる。図1(a),(b)にそれぞれ，学習及び未知サンプルに対する識別実験結果を示す。図中，横軸は各カテゴリーあたりの多重辞書パターン数**n**を，縦軸は各組の平均識別率の平均を取った値R_Mを表す。ここでの未知サンプルに対する結果は，学習サンプル数**m**による識別率の違いを明確にするため，データセット番号151～200のサンプル（以後，固定未知サンプルと呼ぶ）に対し

てのものである。



(a)　学習サンプル



(b)　固定未知サンプル

図1　識別実験結果（p＝63×64）

学習サンプルに対しての結果は，mに関係なく定性的にはほぼ等しい。nの増加と共にR_Mの上昇が見られ，グラフはほとんど不連続な変化を見せず，R_M＝100（%）に漸近していく。ただしmが大きくなれば，グラフの立ち上がりが緩やかになり，R_M＝100（%）に達するnも大きくなる。n＝mのときR_M＝100（%）となるのは当然であるが，ここでの結果は，そのように極端にnを大きくする必要のないことを示唆しており，学習サンプルに関しては，mによらずある程度のnで，類似カテゴリーの識別が可能と考えられる。ただしmに対するnの決定は，更に調査が必要な課題として残されている。

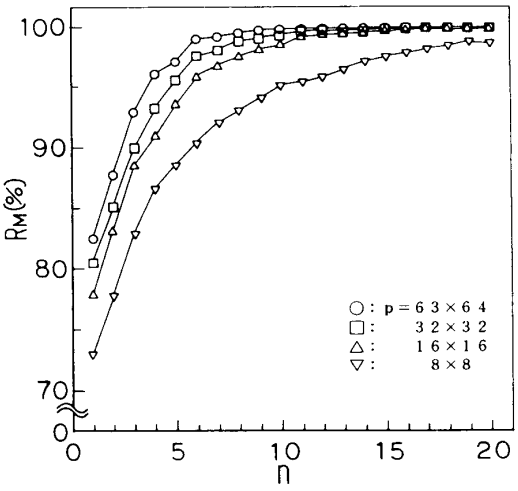
固定未知サンプルに対しての結果では，nを増加させてもm＝50,100の場合には効果が見られず，単純類似度法によるR_Mよりも低下したR_Mが得られている。m＝150の場合には，若干であるがR_Mの上昇の傾向が見られる。またnを固定した場合には，mの増加につれて常にR_Mの上昇が生じているが，類似カテゴリーの識別という面では，不十分な値である。従ってこの結果より，本実験で用いたmでは多重辞書パターンの効果が見られず，多重辞書類似度法の類似カテゴリーの認識能力は高くないと考えられる。ただしグラフの傾向から，mの増加（150＜m）によって，ここで得られたR_Mよりも若干高いR_Mの得られる可能性はあると思われる。

3.3.2 画面次数の変化

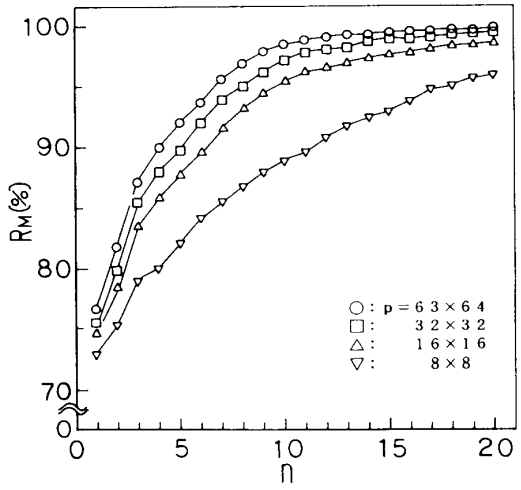
図2(a),(b),(c)に学習サンプルに対しての結果を，図3(a),(b),(c)に未知サンプルに対しての結果を示す。ここでの未知サンプルは，学習サンプル以外のサンプルを指す。

学習サンプルに対する結果では，3.3.1で述べた傾向と同様の傾向が見られ，nの増加と共にR_Mの上昇が見られ，mが大きい程グラフの立ち上がりが遅い。そしてpが小さい（縮小画面）程R_Mは低い値を示している。ただし3.2で述べたように，縮小画面を持つ多重辞書パターンは，p＝63×64において得られたものを単純に縮小したものであるため，そのことが影響を及ぼした可能性がある。従ってpを小さくした場合に，カテゴリー内のクラスタリングを行って多重辞書パターンを作成し認識実験を行う必要があると思われ，これは今後の課題の一つである。

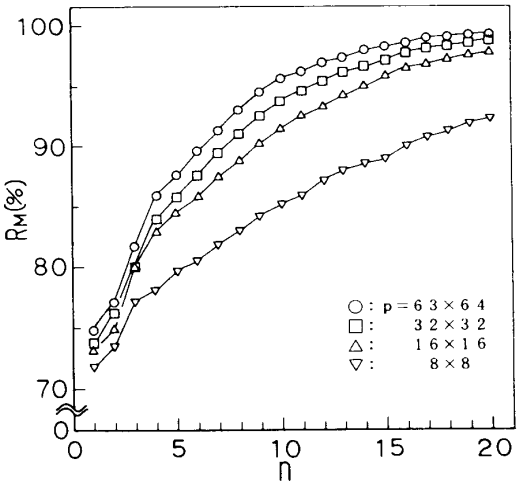
未知サンプルに対する結果では，学習サンプルに対する結果とは逆の傾向が見られ，pが小さい程，高いR_Mを示している。多重辞書類似度法は重ね合せ的手法の応用手法であることより，学習サンプルの質及び量に対する依存性が極めて高く，多重辞書パターンの作成が良好な程，入力サンプルにおける文字パターンの局所的な位置ずれに敏感になると思われる。そのため，p＝63×64では，用いた学習サンプルとは黒点分布形状が若干異なった未知サンプルの識別が困難になったと考えられる。そしてpを小さくするということは，一種のボケ処理を施すことに相当しており，上述の局所的な位置ずれの吸収が成されたと考えられる。しかしpを小さくした場合においても，R_Mの値自体は低いと言わざるをえない。未知サンプルにたいしても，上述の学習サンプルに対する結果で述べた今後の課題が残されている。



(a)　m＝50

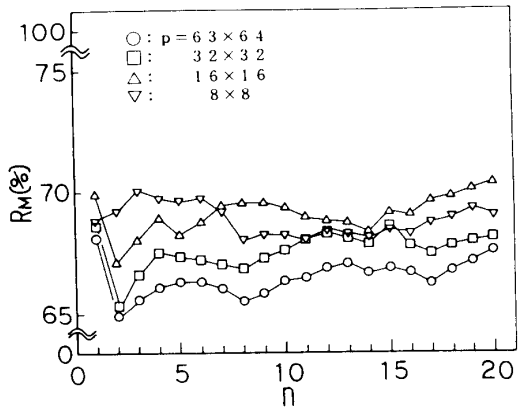


(b)　m＝100

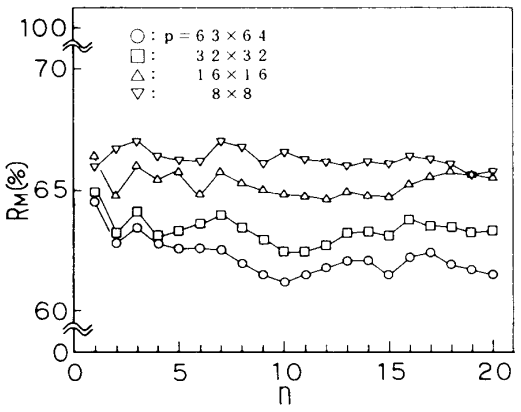


(c)　m＝150

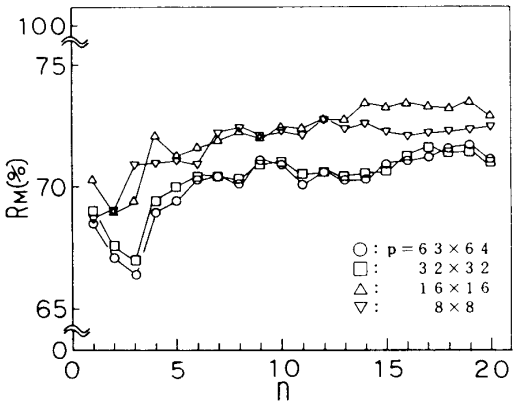
図2　学習サンプルに対する識別実験結果



(a)　m＝50



(b)　m＝100



(c)　m＝150

図3　未知サンプルに対する識別実験結果

3.3.3 類似カテゴリーの大分類率

ここでは，本論文で用いた類似カテゴリー群の大分類率の調査結果について述べる。多重辞書パターンを作成したカテゴリーは，表Ⅰ，Ⅱに示す128カテゴリーである。

図4に$m=100$，$n=10$，$p=63\times64$の条件下での多重辞書類似度法による，大分類10位までの結果を示す。ここでの未知サンプルのデータセット番号は，101～200である。横軸は順位を，縦軸は類似カテゴリー各組の平均累積分類率を平均した値R_Aを表している。なお比較のために，単純類似度法による結果も併せて示している。図中の数字は分類順位10位におけるR_Aを表し，多重辞書類似度法による学習サンプルに対する結果では，7～10位まで同じ値を得た。また表Ⅲ，Ⅳに，それぞれ上述の実験条件と同じ条件下での類似カテゴリー識別実験結果と分類順位1位でのR_Aを示す。

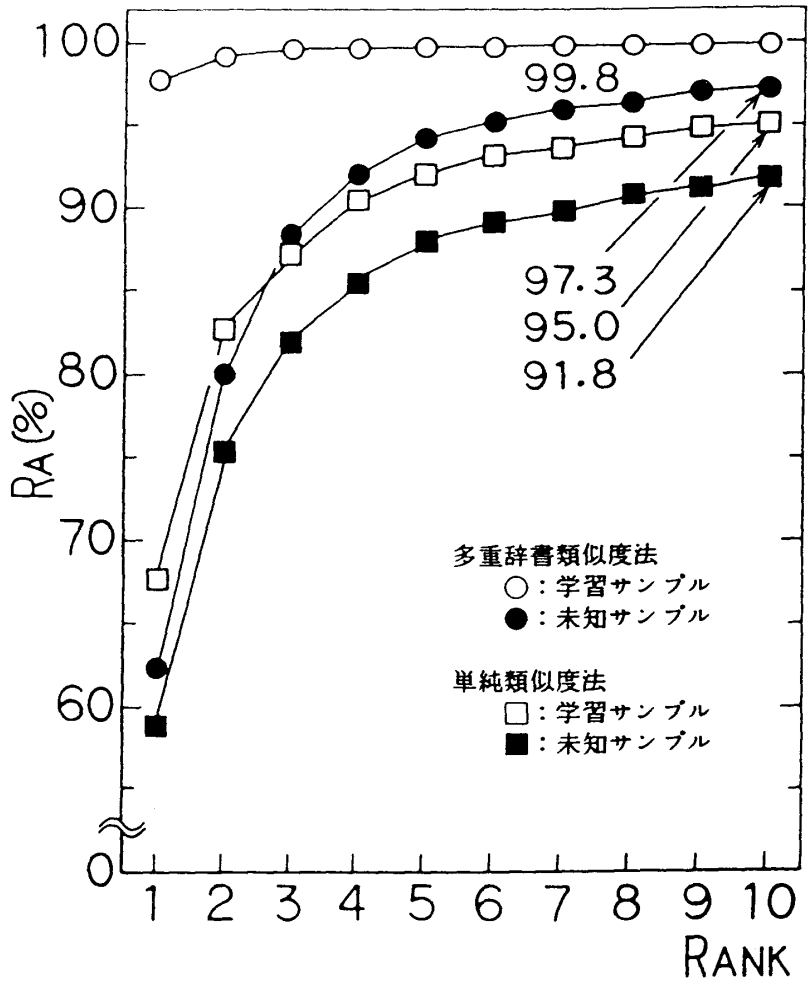


図4 類似カテゴリーの平均累積分類率

($K=128$，$m=100$，$n=10$，$p=63\times64$)

（表Ⅲ）　類似カテゴリー識別実験結果（%）の一例

（m＝100，n＝10，p＝63×64）

組番号	多重辞書類似度法		単純類似度法	
	学習	未知	学習	未知
①	99.5	61.8	79.3	67.0
②	99.3	73.3	80.5	75.3
③	100	74.8	85.8	77.8
④	96.8	71.0	70.0	66.5
⑤	96.3	54.3	69.3	58.5
⑥	97.0	55.3	73.8	59.0
⑦	99.8	73.8	77.5	72.8
平均	98.4	66.3	76.6	68.1

（表Ⅳ）　正読率（大分類1位）（%）

（K＝128，m＝100，n＝10，p＝63×64）

組番号	多重辞書類似度法		単純類似度法	
	学習	未知	学習	未知
①	99.0	58.8	69.5	54.8
②	98.8	67.8	72.8	66.0
③	100	70.3	76.0	68.3
④	96.5	68.5	60.3	59.8
⑤	95.8	51.3	62.5	50.8
⑥	94.8	48.0	60.0	45.8
⑦	99.5	72.3	72.5	66.3
平均	97.8	62.4	67.7	58.8

これらの結果より学習サンプルに対しては，設定した条件下での多重辞書類似度法の類似カテゴリーの識別能力は，非常に高いことがわかる。

未知サンプルに対しては，表Ⅲ，Ⅳよりカテゴリーを増加させた場合の正読率の低下が，単純類似度法によるものよりも小さく，正読率自体も単純類似度法によるものより良くなっており，単純類似度法に比べて認識結果に対する信頼度が高いことがわかる。また図4に示されているように，多重辞書類似度法では分類順位10位で97.3（%）のR_Aが得られており，この値は，単純類似度法によって学習サンプルに対して得られた値よりも高い。

3.3.4 多重辞書類似度法の類似カテゴリー識別能力についての検討

3.3.1～3.3.3で示した結果に基づいて，多重辞書類似度法の類似カテゴリー識別能力についての検討を行う。

まず学習サンプルに関しては，3.3.1で述べた課題，すなわち**m**に対する妥当な**n**の決定という問題を解決するならば，多重辞書類似度法の類似カテゴリー識別能力は高いと言えるであろう。しかしこの問題は，各カテゴリーに対する**n**の決定という別の問題とも関連しかなり困難な問題と考えられ，更に研究の必要がある。

未知サンプルに関しては，種々の条件を変えて認識実験を行ったにもかかわらず，良好と思われる正読率は得られなかった。また本論文で用いた以外の条件（例えば，150＜**m**）を採用したとしても，大幅な正読率の向上はあまり期待できないであろう。従って未知サンプルに対しては，多重辞書類似度法の類似カテゴリー識別能力は高くなく，詳細認識に用いることは難しいという結論が導き出される。しかし大分類実験の結果を考慮すると，上述した多重辞書パターンの構成方法（**m**に対する**n**の決定及び，各カテゴリーに対する**n**の決定）が明確にされるならば，多重辞書類似度法は大分類用の手法としての可能性があると思われる。

4. むすび

本論文では，多重辞書類似度法の類似カテゴリーの識別能力を調査するために，種々の実験条件を変えて認識実験を行い，本実験の範囲で次に示す知見を得た。

(1) 多重辞書パターンを構成する際の方法を明確にすることが可能ならば，学習サンプルに対する識別能力は極めて高い。ここでの構成方法とは，学習サンプルの数に対する妥当な多重辞書パターン数の決定及び，各カテゴリーについての多重辞書パターン数の決定方法を指す。

(2) 未知サンプルに対する識別能力は低いが，(1)と同じく，多重辞書パターンの構成方法を明確にできるならば，本手法は，大分類用の手法としての可能性があると考えられる。

今後の課題としては，本文中で述べた次の事柄が挙げられよう。

(1) 最適な多重辞書パターンの構成方法の明確化。

(2) 画面次数を小さくした後に作成された多重辞書パターンを用いた場合の，多重辞書類

似度法の識別能力の調査。

謝　辞

本研究で使用させていただいた電総研手書き漢字データベースETL-9(B2)を作成された関係各位に感謝します。また，本研究に際し，有益な御助言をいただいた，本学大学院システム科学専攻藤田志郎教授，宮垣嘉也教授に深謝します。

参考文献

1）江島俊郎，市村　洋，木村正行：”重ね合わせ的手法による手書き文字の識別に関する基礎的検討”，情処学論，**28**,11,pp.1207-1210(昭62-11).
2）襄　東善，森下哲次，蕪山幸和，伊崎保直，山本栄一郎：”手書き漢字認識におけるテンプレート複数化の検討”，信学技報，**PRL81**-42,pp.49-56(1981).
3）大田　裕，西村　康，富本哲雄：”特徴マッチングによる手書き漢字認識”，信学技報，**PRU86**-41,pp.9-16(1986).
4）塩野　充：”多重辞書類似度法による手書き漢字識別の基礎実験”，情処学論，**27**,9,pp.853-859(昭61-09).
5）Mitsuru OHKURA and Mitsuru SHIONO :" On the intra−category clustering to make multidictionary patterns for multidictionary templet matching method", in Proc. 9th Int. Conf. on Pattern Recognition, Ⅱ, pp.1029−1031 (14−17 Nov. 1988).
6）大倉　充，塩野　充：”カテゴリー内クラスタリングによる多重辞書類似度法の辞書パターン作成の一検討”，信学論(D-Ⅱ),**J72−D−Ⅱ**,4,pp.499-506(平元-04).
7）斎藤泰一，山田博三，山本和彦：”手書文字データベースの解析（Ⅷ）”，電総研彙報，**49**,7,pp.487-525(1985).
8）大倉　充，塩野　充：”多重辞書類似度法における類似文字識別能力について”，情処学第37回（昭63後期）全大，5W-2,p.1633.
9）橋本新一郎（編著）：”文字認識概論”，pp.34-35,オーム社（昭57).
10）柳井晴男，高木廣文（編）：”多変量解析ハンドブック”，現代数学社（昭61).
11）森　吉弘，横澤一彦，梅田三千雄：”ＰＤＰモデルによる手書き漢字認識”，信学技報，**MBE87**-156,pp.407-414(1988).

A Discrimination Experiment of Similar Handprinted KANJI Characters by Multidictionary Templet Matching Method

Mitsuru OHKURA* ·Takaki FUJIWARA**and Mitsuru SHIONO

*Graduate School of Science, Okayama University of Science,

1−1 Ridaicho, Okayama 700 Japan

**Mitsubishi Electric Computer Systems (Tokyo) Corporation

***Faculty of Engineering, Okayama University of Science

(Received September 30, 1989)

The results of discrimination experiment of similar handprinted KANJI characters by multidictionary templet matching method is presented. This recognition method is one of the applications of templet matching method and uses multiple temlets for each category. To examine the discrimination ability of this method, three experimental parameters are used. The first is the number of multiple templets for each category, the second is the number of training samples which make the templets and the last is the picture size of character patterns. The handprinted KANJI character data base ETL−9(B2) made at Electrotechnical Laboratory is used as the test data.