5W-2

多重辞書類似度法における類似文字識別能力について

大倉 充　　塩野 充
岡山理科大学 工学部 電子工学科

[1] まえがき

手書き文字の認識を困難とする要因は，主なものとして次の2つのことが挙げられる．それは，著しい手書きによる変形（手書き歪み）と，類似した漢字が多数存在することである．近年，このような特徴を持つ手書き文字に対して，最も基本的な認識手法である重ね合せ的手法を応用しようという試みが活発化している．その応用の1つとして，手書き歪みを吸収するために，各文字（カテゴリと呼ぶ）に対する辞書パターンを複数化（多重辞書パターンと呼ぶ）する多重辞書類似度法という手法が提案されている[1]．しかし，後者の問題に対しては，この手法が有効なのかどうか明確ではない．そこで，本報告では，多重辞書類似度法を用いて類似カテゴリの識別に関する基礎的な実験を行い，この手法の類似カテゴリの識別能力の評価を行う．

[2] 多重辞書類似度法

多重辞書類似度法とは，各カテゴリに対して多重辞書パターンを用意し，入力されたサンプルとの間で類似度を計算し，最大の類似度を与える辞書パターンの所属するカテゴリを入力サンプルの決定カテゴリとする方法である．

多重辞書パターンの作成方法は，1つのカテゴリ内に所属するサンプルに対してクラスタリングを行い，類似したサンプル同士でクラスタを形成した後に，各クラスタ内のサンプルを加え合わせて各クラスタごとに濃度値辞書パターンを得るというものである．本実験で用いたクラスタリング手法は，Ward法である[2]．Ward法とは，各サンプルが似ているものから順次集められ，2つのクラスタ間でサンプルの入れ替えが生じない階層的手法の1つであり，クラスタ間の距離を次のように定めるものである．クラスタ$C^{(i)}$に属する全てのサンプルについてのクラスタ平均値（重心）からの偏差2乗和$I^{(i)}$を情報損失量と定義し，$C^{(i)}$と$C^{(j)}$が併合して$C^{(ij)}$となる時の情報損失量の増加量$\Delta I^{(ij)}$

$$\Delta I^{(ij)} = I^{(ij)} - I^{(i)} - I^{(j)} \qquad (1)$$

を$C^{(i)}$，$C^{(j)}$間の距離とする．

[3] 認識実験

3－1 実験データ　実験に用いたデータは，電総研JIS第1水準手書き漢字データベースETL－9（B2）である[3]．このデータベースは，200データセット（3036カテゴリ／データセット）より編成されているが，本実験では，1カテゴリ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，表1に示す類似カテゴリ（4個）の7組を選択した．①～③は，偏と旁のそれぞれ2種類の組み合せで各カテゴリが表現できるもの[4]を，④は，形状自体が似ているものを，⑤～⑦は，構え，旁，偏がそれぞれ同じものを選んでいる．また，クラスタリングは，前半の100サンプルに対して行い，認識実験における学習サンプルとして，後半の100サンプルを未知サンプルとして用いた．なお，本実験では1カテゴリの辞書パターン数は10とした．

表1．実験に用いた類似カテゴリの組

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

3－2 認識実験結果　表2に得られた認識実験結果を示す．比較のために，前半100サンプル全てを加え合わせて得られる，濃度値辞書パターンを用いた単純類似度法の結果も合わせて示している．この結果より，多重辞書類似度法は類似カテゴリに対して，学習サンプルでの識別能力は非常に高いが，未知サンプルではかなり低く，未知サンプルに関していえば，単純類似度法より低い認識率を示す場合があることがわかる．これは，学習サンプル数が十分でないことを表しているものと考えられる．

表2．認識実験結果（%）

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	99.5	45.8	75.5	49.8
⑥	97.0	55.3	73.8	59.0
⑦	99.8	73.8	77.5	72.8
平均	98.8	65.1	77.5	66.9

[4] むすび

多重辞書類似度法の類似カテゴリの識別能力を評価するために基礎的な認識実験を行った．現在，学習サンプル数と辞書パターン数／カテゴリを変化させた場合の認識率の変化の調査を行っている．また，カテゴリ数を増加させた場合の，類似カテゴリの大分類率の調査も行いたいと考えている．

参考文献

（1）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”，情報処理学会論文誌，27，9，pp. 853－859，1986－09．
（2）大倉，塩野：”手書き漢字データベースETL－9を用いたカテゴリ内クラスタリングの実験”，信学技報 PRU87－91，1988－01．
（3）斉藤，山田，山本：”手書文字データベースの解析（Ⅷ）”，電総研彙報，49，7（1985）．
（4）森，横澤，梅田：”PDPモデルの類似漢字識別への応用”，昭和63年度春季信学総全大，D－436．

On Discrimination Ability of Resembled Characters in the Multidictionary Templet Matching Method.

Mitsuru OHKURA, Mitsuru SHIONO
Okayama University of Science.