

手書き漢字認識における大分類のための特徴抽出†

小川 秀夫** 谷口 慶治**

漢字認識では処理速度及び識別率の向上を計る上で大域的特徴による大分類機能を導入することが不可欠である。本論文では手書き漢字を対象として、大域的特徴の一抽出法を示し、それによる分類実験の結果を示す。また、大域的特徴を安定に抽出するために、前処理として2値化及び大きさの規格化について言及する。漢字パターンでは、分類のための情報が文字周辺部に多いこと、及び文字線の方向を4種類に量子化可能であることを基に、観測点を文字枠上の左右上下4点及び重心を原点として文字を4象限に分割した各領域の中心に置き、大域的特徴として各観測点から見た文字線の方向別分布量を取り上げ、これを20次元ベクトルとして表現する。この20次元大域的特徴空間において、距離、類似度及びK-NN法により分類を行った結果、誤り率0.09で候補カテゴリーを全体(教育漢字881字)の2%まで減少することができ、全カテゴリー数の約20%を候補とした場合には、ほぼ誤りなくその中に正しいカテゴリーが含まれることが示された。

1. ま え が き

漢字認識において大分類は極めて重要な処理過程である。これにより膨大な数のカテゴリーを少数の候補カテゴリーに減少することができ、その結果処理時間及び識別精度の向上を図ることが可能となる。現在、印刷漢字を対象として多くの有効な手法^{1)~4)}が報告されているが、手書き漢字については少ない⁵⁾。本論文では手書き漢字を対象として、大分類のための大域的特徴を提案し、それによる分類結果について述べる。

一般に大分類のための大域的特徴としては、観測が容易であり、文字の歪みあるいは雑音に対して安定であること及び分類効率が良いこと(候補カテゴリーが少数であること)が必要とされる。漢字では大分類のための情報が文字の周辺部に多く存在し²⁾、文字線を4方向に近似できるとされており、手書き漢字においてもこの性質は保存されていると見なすことができる。そこで、本稿ではこれらの性質に着目し、大域的特徴の観測点を文字周辺部におき、文字線の方向別分布量を抽出する手法を提案する。また、手書き漢字では文字の位置、大きさ及び形状に大きな変動を伴うため一般に前処理として、これらの規格化が行われる。この規格化法についても多くの方式^{6),7)}が報告されているが、本稿では特徴抽出方式との関係から文字の重心を基準にした平行移動及び重心を原点とする2次元直交座標系の各象限別に文字を伸縮する手法を用いる。

なお、本稿における“大域的”特徴とは、その観測範囲及び観測手法から局所性がかかなり緩和されていること、及び大分類のための特徴という意味で用いたものであり、漢字パターンの識別あるいは記述に用いる特徴と明確な区別はない。

2. 前 処 理

前処理として、ここでは観測パターンの2値化、細線化、文字線の方向付け及び大きさの規格化を行う。

2.1 2値化、細線化及び文字線の方向付け

2値化法⁸⁾として多くの手法が提案されているが、ここでは比較的安定に2値化できる周囲の平均濃度と比較する次の手法を用いた。今、 $N \times N$ にデジタル化された観測パターンを $f(i, j)$ とし、閾値 $T(i, j)$ を次の式で定める。

$$T(i, j) = \frac{1}{(2m+1)^2} \sum_{k=-m}^m \sum_{l=-m}^m f(i-k, j-l) + \alpha \quad (1)$$

ここで、 α は補正項でありこの値は現在実験的に定めている。この閾値により次のように2値パターン $P(i, j)$ を得る。

$$P(i, j) = \begin{cases} 1 & f(i, j) \geq T(i, j) \\ 0 & \text{その他} \end{cases} \quad (2)$$

次に本稿で大域的特徴とした文字線の方向別分布量は $P(i, j)=1$ である要素の数が重要な意味を持つ。したがって、文字線幅の規格化が必要となり、ここでは細線化⁹⁾により1要素幅とした。また、文字線の方向を定める手法はすでに文献¹⁰⁾で報告したものであり、それを参照されたい。

2.2 大きさの規格化

† Feature Extraction for Global Classification of Hand Written Chinese Characters by HIDEO OGAWA and KEIJI TANIGUCHI (Faculty of Engineering, Fukui University).

** 福井大学工学部電子工学科

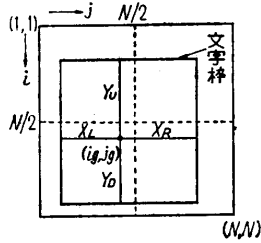


図1 文字の位置及び大きさの規格化
Fig. 1 Diagram for standardization.

ここでは文字の重心が局所的な歪みに対して比較的安定していること、及び本稿における大域的特徴の観測点が画面の周囲にあるため文字枠が各観測点に一致する必要があることから、規格化として文字の平行移動及び各象限別の伸縮を行う。すなわち、図1に示すように文字を囲む枠を定め、重心の座標を (i_c, j_c) としたとき、任意の要素 (i, j) は(3)式に従い (\bar{i}, \bar{j}) に平行移動される。

$$\left. \begin{aligned} \bar{i} &= i - (i_c - N/2) \\ \bar{j} &= j - (j_c - N/2) \end{aligned} \right\} \quad (3)$$

ここで、 $(N/2, N/2)$ は画面の中心座標である。次に重心の位置すなわち $(N/2, N/2)$ を原点とする2次元直交座標系の各象限において文字の伸縮を行い $N \times N$ 画面の大きさに統一する。ここでは簡単のため第1象限の要素 (i, j) に対する伸縮された座標 (\bar{i}, \bar{j}) の関係を(4)式に示す。

$$\left. \begin{aligned} \bar{j} &= \frac{N}{2} + \frac{N/2 - C}{X_R} \left| \frac{N}{2} - j \right| \\ \bar{i} &= \frac{N}{2} - \frac{N/2 - C}{Y_U} \left| \frac{N}{2} - i \right| \end{aligned} \right\} \quad (4)$$

このとき、第1象限水平方向伸縮率 W_{1x} 及び垂直方向伸縮率 W_{1y} をそれぞれ次のように定義する。

$$\left. \begin{aligned} W_{1x} &= \frac{N/2 - C}{X_R} \\ W_{1y} &= \frac{N/2 - C}{Y_U} \end{aligned} \right\} \quad (5)$$

ここで、図1に示すように X_R は重心から右文字枠までの距離、 Y_U は上文字枠までの距離であり、 C は規格化後周囲を枠のために残す任意の要素数である。同様に他の各象限においても、 X_L, Y_D を用い伸縮後の座標 (\bar{i}, \bar{j}) 及び水平方向伸縮率 W_{2x}, W_{3x}, W_{4x} 及び垂直方向伸縮率 W_{2y}, W_{3y}, W_{4y} を求めることができる。この結果、文字は重心を $N \times N$ 画面の中心とし、 $(N-2C) \times (N-2C)$ の様な大きさに規格化される。ここで、 $W_{1x} = W_{4x}, W_{1y} = W_{2y}, W_{2x} = W_{3x},$

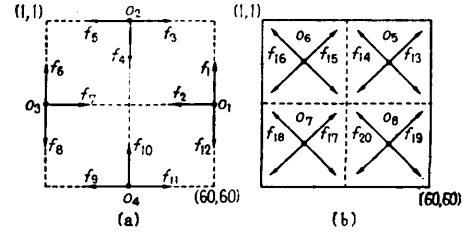


図2 観測点の位置
Fig. 2 Observation points.

$W_{3y} = W_{4y}$ であることに注意したい。

3. 大域的特徴抽出

漢字パターンの属性である。(1) 周辺部に大分類のための情報が多い、(2) 文字線の方向を垂直、水平、右斜め及び左斜めの4方向に量子化できる、ことを基に垂直、水平成分に対しては図2(a)に示す画面周辺の観測点から、また斜め成分については画面の対角線附近に比較的隣接して分布することを基に同図(b)に示す観測点から文字線の方向別分布量を求め、これを大域的特徴とする。したがって、大域的特徴 G は次のように20次元ベクトルとして表現される。

$$G = (f_1, f_2, \dots, f_{12}, f_{13}, \dots, f_{20}) \quad (6)$$

各要素の抽出は基本的に同じであるため、ここでは観測点 O_1 における第1象限の垂直成分分布量 f_1 について具体的に述べる。

$$f_1 = \sum_{(i,j) \in A} P(i, j) \cdot W_1(\theta(i, j)) \cdot M(D) \cdot V(\theta(i, j)) \quad (7)$$

ここで、 A : 観測領域 (この場合第1象限)

$$P(i, j) = \begin{cases} 1 & \text{要素 } (i, j) \text{ が文字線上の場合} \\ 0 & \text{その他} \end{cases}$$

$$W_1(\theta(i, j)) = \sqrt{(W_{1x} \cos \theta(i, j))^2 + (W_{1y} \sin \theta(i, j))^2}$$

W_{1x} : 第1象限水平方向伸縮率

W_{1y} : 第1象限垂直方向伸縮率

$M(D)$ は距離に対する重み関数であり、次のように定める。

$$M(D) = \begin{cases} 1 - 2D/N & D \leq N/2 \\ 0 & \text{その他} \end{cases}$$

$$D = \sqrt{(m-i)^2 + (n-j)^2}$$

(m, n) : 観測点の座標 (この場合 O_1)

$V(\theta(i, j))$ は文字線角度に対する重み関数であり表1のように定めた。

なお (7)式において $W_1(\theta(i, j))$ の項は、文字の

表 1 文字線方向に対する重み関数

Table 1 Weighting function concerning the stroke directions.

	$0 \leq \theta(i, j) < 45$	$45 \leq \theta(i, j) < 90$	$90 \leq \theta(i, j) < 135$	$135 \leq \theta(i, j) < 180$
水平 $H(\theta(i, j))$	$\frac{\cos 4\theta(i, j)+1}{2}$	0	0	$\frac{\cos 4\theta(i, j)+1}{2}$
垂直 $V(\theta(i, j))$	0	$\frac{\cos 4\theta(i, j)+1}{2}$	$\frac{\cos 4\theta(i, j)+1}{2}$	0
右斜め $R(\theta(i, j))$	$1-H(\theta(i, j))$	$1-V(\theta(i, j))$	0	0
左斜め $L(\theta(i, j))$	0	0	$1-V(\theta(i, j))$	$1-H(\theta(i, j))$

規格化処理における伸縮後の実質要素数を考慮したものである。すなわち、本稿では伸縮後の文字構成要素間の補間を行わないため、伸縮後の1要素は実質的に $W_1(\theta(i, j))$ 個の要素がそこに集約されていると近似した。同様に観測点 O_1 における水平成分分布量 f_2 は(7)式において観測領域 A を第1象限及び第4象限とし、それぞれの領域の伸縮率を用い、さらに $V(\theta(i, j))$ に対して表1で与えられる $H(\theta(i, j))$ を用いることにより抽出される。さらに f_3, f_4, \dots, f_{12} についても同様に類推されよう。また、 $f_{13}, f_{14}, \dots, f_{20}$ については、(7)式における $M(D)$ に対して次の(8)式を用い、さらに観測領域 A すなわち f_{13}, f_{14} については第1象限、 f_{15}, f_{16} は第2象限、等及び角度に対する重み関数を適宜選ぶことにより同様に抽出される。

$$M(D) = \begin{cases} 1-4D/N & D \leq N/4 \\ 0 & \text{その他} \end{cases} \quad (8)$$

次にこの大域的特徴 G による大分類について述べる。

4. 大分類

大分類における候補カテゴリーの選出法として、観測された大域的特徴 G に対し、

(1) 定距離(類似度)内にあるカテゴリーを選出する手法、

(2) 特徴空間を重複したクラスに分割し、 G の属するクラスを選出する手法、

(3) 特徴空間を重複しないクラスに分割し、各クラスに適当なメンバーシップ関数 M_k を定め、 G に対する関数 $M_k(G)$ の大きいクラスを選出する手法⁵⁾、などがある。本稿では、提案した大域的特徴の有効性あるいは性質を調べることを目的とし、上記(1)の手法による分類を行った。しかし、カテゴリー数が極めて多い場合にはあらかじめ特徴空間のクラス分割を行いクラス単位で選出することが大分類処理時間を小さ

くするために必要となることに注意したい。また、尺度として次に示す類似度 $S(G, \bar{G}_j)$ 及び距離 $D(G, \bar{G}_j)$ を用いた。

$$S(G, \bar{G}_j) = \frac{\sum_{i=1}^{20} f_i \cdot f_i'}{\sqrt{\sum_{i=1}^{20} f_i^2} \cdot \sqrt{\sum_{i=1}^{20} f_i'^2}} \quad (9)$$

$$D(G, \bar{G}_j) = \sqrt{\sum_{i=1}^{20} (f_i - f_i')^2} \quad (10)$$

ここで、 G は観測された大域的特徴、 \bar{G}_j は j 番目のカテゴリーの基準大域的特徴である。

5. 分類結果及び評価

本稿では対象を教育漢字 881 字の枠内手書き漢字とした。また大域的特徴の基準値として各カテゴリー(漢字1種ごと)につき 15 文字から 20 文字の楷書手書き漢字(総計 15,000 字)から抽出した大域的特徴の平均値を用いた。対象の手書き漢字は約 9 mm × 9 mm の枠内に万年筆、ボールペン及び鉛筆で書かれたものであり、これをテレビカメラで観測し 60 × 60 のデジタル画像とした。分類実験はテスト用パターンとして総計 9,500 字、各カテゴリーにつき 7 文字から 20 文字の枠内手書き漢字を用いて行ったものである。なお、この実験で用いたテストパターンは訓練用(基準値作製用)パターンとは別に抽出したものである。図3にそれらの一例を示す。大分類では小さな誤り率でいかに少ない候補カテゴリーを選出できるかが重要な問題である。したがって、機能は次に定める誤り率 e 及び候補として選出されるカテゴリーの数、ここでは r の両者により評価される。

$$e = \frac{\text{正しいカテゴリーが選出されなかったテストパターン数}}{\text{全テストパターン数}}$$

r = テストパターン全体における r_i の平均値
ここで、

材強院君討
 情横童承結
 通我街胃料
 期雜塩深現
 留形独死成
 君画兎俗資
 牧運歩方信
 帯直夏千産
 経面福后持

図3 手書き漢字の一例 (左第1列は訓練用手書き漢字, 他の文字はテスト用手書き漢字の一例)
 Fig. 3 Handwritten Chinese characters, left most column; examples of training patterns which are written in the square style, the others; examples of test patterns.

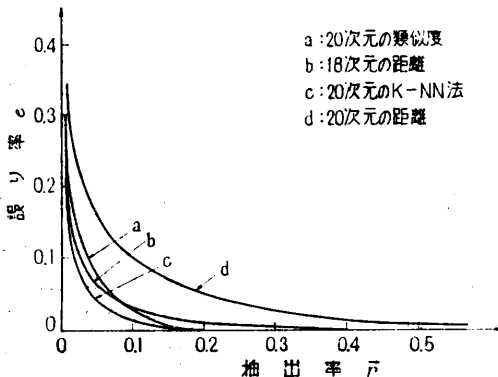


図4 大分類結果 (f と e の関係)
 Fig. 4 Results of global classification showing the relation between the rate of extracted categories and error rate.

$$r_i = \frac{(i \text{ 番目のテストパターンに対して候補カテゴリーとして選出されたカテゴリー数})}{(\text{全カテゴリー数 (881)})}$$

図4に大分類結果を \bar{r} と e の関係で示した。ここで、同図 (a) は 20 次元大域的特徴を用いた結果として (9) 式に定めた類似度を用いた結果であり、(b) は (11) 式により定められる距離を尺度として用いたものである。

$$D(G, \bar{G}_i) = \sqrt{\sum_{n=1}^{20} (f_n - \bar{f}_n)^2 - (M_i^2 + M_i'^2)} \quad (11)$$

広: 広展後氏降感武后兄坂
 官: 管算官管賞尊宜建春管
 問: 明南開開朝試角の郎問
 徳: 体健備歷術視祝風位極
 詩: 魚試説角業許討誤勢有
 協: 械術航研個荷船輪細林
 飛: 界報休球体服林版味板
 焼: 鳥錦林竹術拡細供体結
 引: 訓動別制則列創功利打
 供: 使状板捺件体付横作竹
 宿: 短儘庭鳴増程店招預律
 賃: 重尊尊首音温自告差愛
 旧: 個院臨陸田加博府内曲
 歩: 共接岸拳走方友交幸誤
 帳: 帳眼修練味横博昨限使
 仁: 住在后任延経往社血低
 習: 宜重喜書書雲胃言運專

図5 近接カテゴリーの一例

Fig. 5 Examples of similar categories in the sense of proposed global feature.

ここで、

$$M_i = \max_{k=1 \sim 20} \{(f_k - \bar{f}_n')\}$$

$$M_i' = \max_{k=1 \sim 20, k \neq i} \{(f_k - \bar{f}_n')\}$$

これは特に差が大きな特徴要素、すなわち多くの雑音あるいは歪みの影響を受けていると考えられる要素を除外して距離を評価したものであり、実験的に2つの要素を除外した場合、平均的に最も良い結果が得られた。また、(c) は 20 次元大域的特徴を用いた K-nearest neighbour 法により分類した結果であり、(d) は 20 次元大域的特徴により尺度として (10) 式で定めた距離を用いたものである。ここで、同図 (c) では誤り率 0.09 で全カテゴリー数の 2% (約 18 個) まで候補カテゴリー数を減少することができ、全カテゴリー数の 20% (約 176 個) を候補カテゴリーとして選出した場合にはほぼ誤りなくその中に正しいカテゴリーが含まれていることを示している。これらの結果は、本稿で提案した大域的特徴により大分類が効果的に行われることを示すものであり、これにより手書き漢字認識システムの高い識別能率または処理速度が期待される。

図5に任意のカテゴリーに対して距離的に近いカテゴリーを順に示した。ここで、これらの類似カテゴリーにおいて、その局所的パターンを見た場合かならずしも似ていないことがわかる。このような性質はこれらの候補カテゴリーから局所的な特徴により識別を行う場合、極めて有利なものである。

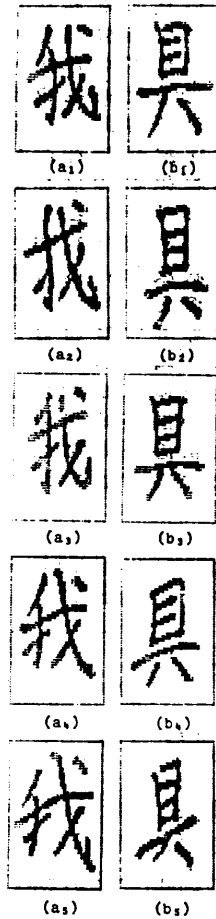


図6 2値化された手書き漢字の一例

Fig. 6 Examples of binarized handwritten Chinese characters (test patterns).

図6に2値化された手書き漢字の一例を示し、その大域的特徴を表2に示した。ここで、文字線の位置あるいは傾きの変動による大域的特徴量の変化の一例と

して例えば、文字 a_3 の大域的特徴 f_{19} 及び a_4 の f_{15} はこの観測点において観測の対象となる文字線が、観測点に近い位置にあり、その結果特徴値が比較的大きくなっている。また、 a_5 の f_{10} では観測の対象となる縦線が観測点から他の同一文字と比較して離れているため特徴値は小さい。 b_5 の f_2 及び b_5 の f_7 は、文字の下部が小さく書かれ、また文字線が回転しているため、これらの影響を受けたものである。また、 b_5 の f_5 については、文字の左上が離れており、このため細線化により f_5 の観測対象となる文字線がほとんどなくなったために特徴値が小さくなったものである。このように、文字線の位置及び回転により対応する観測点の特徴値は少なからず影響を受ける。しかし、表2に示すように他の特徴要素は比較的稳定な特徴値を示しており、特徴値全体として良い分類結果を与えている。

さらに、文字線の位置及び回転の影響を軽減するためには、観測点の位置を文字に適応して変化させること及び特徴値の動的な類似性の評価法を導入することなどが考えられるが、これは今後の課題としたい。

6. むすび

漢字の周辺部に定めた観測点から観測される文字線の方向別分布量を大域的特徴として提案し、その大分類結果を示した。それによると誤り率 0.09 で全カテゴリ数 の 2% まで候補カテゴリ数を減少することができ、全カテゴリ数の約 20% を候補カテゴリとして選出した場合には、ほぼ誤りなくその中に正しいカテゴリが含まれることが知られた。また、選出された候補カテゴリの局所的パターンはかならずしも類似したものではなく、これは局所的特徴による

表2 図6における“我”と“具”に対する大域的特徴量

Table 2 Examples of global features for the Chinese characters 我 and 具 in Fig. 6.

	value of global features																				distance to the correct category	R	
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}			
我	a_1	0	5	2	6	1	1	5	1	1	4	1	2	0	3	5	2	9	0	5	9	6.00	1
	a_2	1	3	3	5	1	1	6	2	1	5	0	2	2	2	4	1	9	1	3	10	6.16	1
	a_3	0	5	3	4	1	1	6	1	0	8	0	5	2	1	6	1	6	1	7	7	7.68	1
	a_4	1	6	1	3	0	1	9	0	1	2	0	1	0	7	7	0	7	1	4	9	8.94	1
	a_5	1	9	2	2	0	6	0	0	1	0	3	0	5	5	0	8	1	5	8	9.89	1	
具	b_1	3	12	7	5	7	1	8	0	2	0	3	0	0	1	0	4	8	0	0	7	6.86	1
	b_2	2	9	7	4	6	1	4	0	5	1	5	0	2	2	1	1	5	0	0	8	9.05	1
	b_3	1	10	6	2	6	1	7	1	5	0	4	0	2	2	1	3	4	0	0	6	9.16	1
	b_4	5	11	6	3	4	1	4	0	5	0	4	0	1	2	3	1	9	0	1	7	10.00	2
	b_5	2	6	6	2	3	0	1	0	4	0	2	0	2	2	0	1	6	0	5	3	13.16	17

R: The ranking of appearance of the correct category, which is based on the distance.

漢字の識別に対して極めて有利な性質であり、提案した大域的特徴の有効性を示すものである。今後は対象を当用漢字にまで拡張し、大域的特徴の雑音構造を考慮した類似性の尺度あるいはクラス分割の手法を開発し、クラス単位で選出した場合の大分類機能を評価していきたい。

参 考 文 献

- 1) 坂井, 森: 漢字パターンの大分類, 信学会パターン認識と学習研資, PRL 73-14 (1973).
- 2) 河田, 平井, 森: 漢字認識のための漢字大分類の一手法, 昭和 48 年度情報処理学会全国大会, 114, p. 227 (1973).
- 3) 山本, 中島, 中田: 階層的パターンマッチング法による漢字認識の実験—印刷漢字認識の研究—信学論 (D), Vol. 56-D, No. 12, p. 714 (1973).
- 4) 梅田, 有野, 増田: 粗いメッシュパターンによる明朝体マルチフォント印刷漢字の分類, 信学会パターン認識と学習研資, PRL 77-65 (1978).
- 5) 小川, 谷口: 手書き漢字の大分類法, 信学論 (D), Vol. 60-D, No. 9, p. 766 (1977).
- 6) Fujimoto, Y., et al.: Recognition of Hand-printed Characters by Nonlinear Elastic Matching, Proc. 3rd Joint Conf. on Pattern Recognition (1976).
- 7) 小高, 荒川, 石井: オンライン入力漢字の特徴記述法の検討, 信学会パターン認識と学習研資, PRL 77-58 (1977).
- 8) 森, 大津: 認識問題としての二値化と各種方式の検討, 情報処理学会イメージプロセッシング研資, 15-1 (1977).
- 9) 田村: 細線化法についての諸考察, 信学会パターン認識と学習研資, PRL 75-66 (1975).
- 10) 小川, 谷口: 手書き漢字認識における部分パターンの観測, 信学論 (D), Vol. J 59-D, No. 8, p. 588 (1976).

(昭和 53 年 4 月 3 日受付)

(昭和 53 年 9 月 1 日採録)