四隅部分の特徴抽出による手書き漢字の分類 *

4G－5

○沈 英謀 **　津村幸治 **　斎藤 義夫 **　吉田 嘉太郎 **
千葉大学工学部 ***

1　はじめに

従来の文字分類の方法としては，" つくり"，" へん" など文字の部分的な特徴を抽出する方法が多く用いられているが，これらは部分を抽出する過程に，多くの問題を含んでいる．そこで，本研究では，四隅部分の特徴をコード化する「四角コード」[1] と類似の分類方法を提案し，実際への適用を試みた．本方式ＦＣＣ（Four Corner Code ）は，部分的な単純パターンを機械的に認識し，合理的に４桁のコードを生成するものであり，楷書体常用漢字 1945 字に対してその有効性を確認した [2]．今回は，手書き漢字の統計的な特徴量を考慮した四角コードによる文字分類を行ない，その有効性の適用範囲と課題について検討した．

2　分類処理

文字を分類する過程は以下の手順に従う．

1. 文字フォントのビットパターン（64 × 63 ドット）を読み込み，２値化した画像データを得る．
2. 取得した２値画像を前処理として細線化処理 [3], および細線化による変形の修正を行なう．
3. 計算処理を容易にするために，細線化画像をグラフ構造に変換する [4] [5].
4. 文字枠の四隅にそれぞれもっとも近い Node（ＦＣＣ Node ）を抽出する．
5. ＦＣＣ Node 近傍の形状的な特徴を抽出する．特徴抽出は，後述する６種類のタイプと比較し，対応したコードを当てはめる．
6. 最後に，５．で得られた四隅のコードを左上，左下，右上，右下の順序で並べて一文字に対して４桁のコードを対応づける．（例えば" 仕" のコード＝ 3251 ）

表 1: 特徴コード表

Code	線分種類	勾配条件
1	" 一" 横線	$-1/7 <= slope <= 1/6$
2	" 丨" 縦線	$slope <= -8$ or $slope >= 6$
3	" ／" 負斜線	$-8 < slope < -1/7$
4	" ＼" 正斜線	$1/6 < slope < 8$
5	" ＋, ×" 二線以上の交差	ＦＣＣ Node とつながる Node のつながる Node の数 >= 4, 或いは, ＦＣＣ Node のつながる Node 数 >= 4
6	" 「,」, 口" 角	曲率 $curve >= 150$

3　特徴コードの設定

手書き漢字の特徴を考慮して，今回は四隅の特徴抽出を表１に示すように，６種類のタイプを取り上げてコード化した．コードの数としては，前回の楷書体の場合（８種類）より少なくなっている．これは手書きの場合のほうが線の太さや傾きに個人差があり，細分化するとかえって誤認識が増えるためである．そこで，以下に示す点を考慮し，単純で一義的に確定するように特徴コードを選定した．

- 楷書体では，文字の各部の配置が均等で，ずれが少ないため，code 0 を用いたが，手書き漢字では適用しないこととした．
- 手書き漢字では，Ｔ字形の交差線は交差させずに書く場合が多いことから，このコードを削除した．
- 手書き漢字では，横線と縦線は多少斜めに書くため，勾配（*slope*）の範囲は楷書体の場合より，大きく範囲を設定した．

4　ＦＣＣによる漢字の分類結果と考察

電総研データベース ETL8B2 中の 881 カテゴリの文字（ 160 sample/ １文字）を対象に，ＦＣＣによる手書き漢字の分類を行なった．その結果，次のことが確認された．

1. 図１は文字カテゴリ数とＦＣＣ数の関係である．横軸は１つのＦＣＣに含まれる文字数を示

---

* Classification of Chinese Characters by Four Corners Characteristics

** Yin-Mou Shen,Tsumura Kouji,Yoshio Saito, Yoshitaro Yoshida
*** Faculty of Engineering ,Chiba University 1-33 Yayoi-cho Inage-ku, Chiba 263 Japan

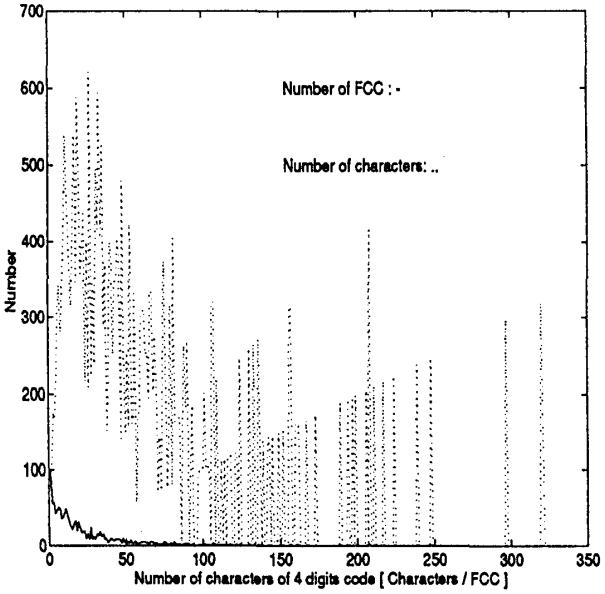

図 1: 文字カテゴリ数とＦＣＣ数の関係

し，縦軸は実線の場合ＦＣＣの総数，破線の場合文字の総数を示している．881 カテゴリの文字は，合わせて 1262 個のＦＣＣをもち，そのＦＣＣに含まれているカテゴリの総数は，35309 個である．したがって，一つのＦＣＣは平均で約 28 個の文字と対応することになる．

2. 同一カテゴリに対し，できるだけ少ないＦＣＣが導かれることが理想であるが，今回の分類結果では，160 サンプル／カテゴリに対し，平均約 40 個のＦＣＣが求められた．一つの文字に対し，ＦＣＣが多くなった理由としてはつぎのことが挙げられる．
   - 抽出特徴では，線分要素を単純に横線，縦線，正斜線あるいは，負斜線に区別することが難しく，許容領域を設定する必要があること．
   - サンプルにより，文字の特徴が大きく変動するため，抽出される隅近傍の特徴も大きく異なる可能性があること．
3. 手書き漢字のＦＣＣはサンプルにより異なるが，約半数近くのカテゴリにおいて，各隅のコードが 1 から 2 種類に確定することが確かめられた．例えば，" 川" に対して各隅におけるコードの頻度を調べた結果を表 2 に示すように，ＦＣＣとして 2222 あるいは 3322 が多くなることがわかる．
4. 各カテゴリにおいて，多くのサンプルと対応するＦＣＣについて整理した結果を表 3 に示す．対応するサンプル数の多いのＦＣＣを二つ選び，その中に含まれる割合い（対応率）を調べ，881 カテゴリの全体に対する平均対応率は 39 % である．" 姉" のように 8 %と対応率の悪い例もあるが，カテゴリによっては二つのＦＣＣで 70 %以上に達することが確認できる．このことは，高い確率で同一カテゴリの文字が数個のＦＣＣコードと対応することを意味している．

表 2: 文字" 川" の特徴コードの頻度

code	1	2	3	4	5	6
左上隅の数	0	105	52	3	0	0
左下隅の数	0	104	54	2	0	0
右上隅の数	1	154	1	1	0	3
右下隅の数	0	157	1	2	0	0

表 3: 各カテゴリにおけるサンプル数の多いＦＣＣ

文字	ＦＣＣと対応するサンプル数	samples/160
人	3334: 88, 2324: 69	98 %
川	2222: 98, 3322: 51	93 %
以	2334: 94, 2324: 46	81 %
演	4364:108, 1364: 17	78 %
信	3216: 45, 3236: 38	52 %
去	5354: 52, 5654: 28	50 %
愛	3334: 18, 3565: 15	20 %
委	1354: 12, 3334: 10	14 %
番	1636: 8, 4336: 8	10 %
姉	2356: 7, 2555: 5	8 %
	881 カテゴリの平均対応率	39 %

5 おわりに

以上のことより，手書き漢字に対して単純なＦＣＣで，ある程度の分類を容易に行なえるといえる．現状では，ＦＣＣのみでは対応する文字の数が多く，漢字認識としては解決すべき問題点も多いが，特徴コードの学習や他の情報と融合することにより改善可能と考えている．

参考文献

[1] 陸師成. 辞書. 文化図書公司, 台湾台北市, 1992.

[2] 沈英謀. 四隅部分の特徴抽出による文字の分類. 情報処理学会第 51 回全国大会, 講演論文集 2 , pp. 171–172, 1995.

[3] 鳥脇純一郎. 画像処理のためのディジタル画像処理. 第 3 章, 3.3,pp.56-59, 1993.

[4] 長尾真. パターン情報処理. 第 3 章, 1991.

[5] 奥村彰二ほか. 漢字画像から文字要素の自動抽出. 情報処理学会論文誌, Vol. vol.32,No.1, pp. 50–61, 1991.