四隅部分の特徴抽出による文字の分類 *

4R－4

○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

文字を分類する方法として，従来は，＂つくり＂，＂へん＂を用いることが多いが，文字認識を行なう場合，これらの部分を抽出することが困難である．そこで，本研究では，文字の分類および認識を効率よく行なうことを目的として，文字の四隅部分の特徴に基づく分類法を提案する．本方式は，四隅部分の特徴をコード化する「四角コード」と称するもので[1]，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，漢字自動認識などへの応用に適する．具体的には楷書体常用漢字1945字を対象に本方式による分類実験を行ない，その有効性を確認した．

2　分類処理

文字を分類する過程は図1に示すように，以下の手順に従う．

1. 計算機内の文字フォントのビットパターン（48×48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[2]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[3][4]．グラフ構造では一つの文字は，いくつかの節点（Node：端点，交差点，角[3]）と線分から構成され，（1）Nodeの座標値　（2）接続しているNodeの数（3）接続している2つNodeの間の線分の種類，長さ，勾配(slope)などを含んでいる．
4. 文字の枠の四隅（左上，左下，右上，右下）にそれぞれもっとも近いNodeを抽出する．隅との距離が同じNodeが二つある場合，両方を抽出する．
5. 抽出されたNode近傍の形状的な特徴を抽出する．特徴を抽出する際，後述する8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＝3261）

なお，四隅の特徴抽出には表1に示すように，8種類のタイプを取り上げ，これに対応してコード化している．これは旧来の「四角コード」の基本的なものを選んでおり，この数によって分類の精度は異なるといえる．

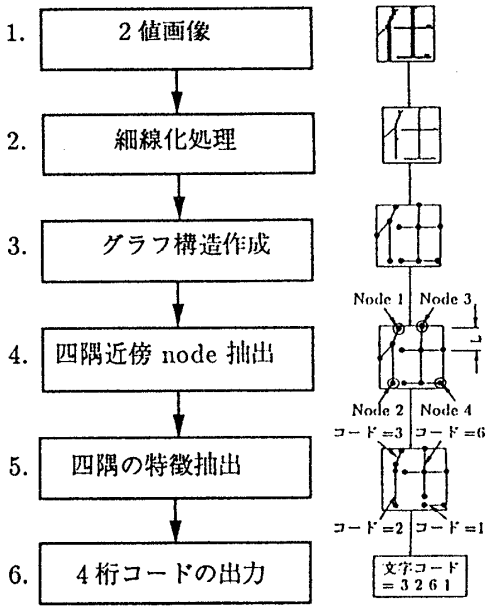

図 1: 文字の分類過程

3　文字頻度分布

本実験で使用したデータは，昭和56年内閣告示第一号に掲げられた1945字の常用漢字である[5]．ビットパターンで表された文字について各位置におけるビットの表れる頻度分布を調べた結果を図2に示す．図より文字が大体平均的に枠内に分布していることがわかる．つぎに，本実験で抽出されたNodeの頻度分布を図3に示すように，Nodeも図2のように全体に分布しているが，下部と四隅にある程度集中している．このことより，四隅にもっとも近いNodeを抽出すれば，文字識別の情報としては十分と考えられる．

---

* 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

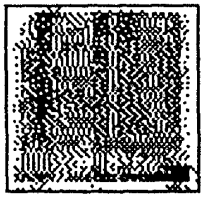

図 2: 文字ビットパターンの頻度分布

図 3: Node の頻度分布

4 分類結果と考察

- 設定した特徴を容易に識別することができるため，分類時間をかなり短縮できる．
- 本方式により，常用漢字 1945 字を 619 グループに分類することができた．4 桁のコードが 1 文字だけに対応する場合が 289 あり，文字認識を行なう際に，効率よく識別できると考えられる．
- 同じ 4 桁コードで表される文字の例を表 2 に示す．Code 7377 は，もっとも多い場合で文字数 49 個につけられる．つぎは，Code 7777 で，文字数 40 個に対応する．このような場合には，もう一段階の分類を行なうことが不可欠といえる．
- 本規則により，文字の形がかなり類似していると同一コードになるが，あまり類似していない場合もある．例えば表 2 に示すように，Code 4547 では，沿と浴で，かなり類似しているが，Code 6374 の賀と想は，一般的には類似していると認めにくい．これらの問題は設定した特徴コードを詳細に検討することで改善できると予想される．
- 細線化により，文字の変形が生じ，これが誤認識の原因になっていることが確認された．例えば，図 4 に示すように，文字" 何 " の左上の所は，かなり変形している．この斜線部分は 角に変形してしまうため，左上の特徴は Code 7 とみなされる．したがって細線化による変形に対する対策も今後の重要な課題であろう．

5 おわりに

四隅の特徴を抽出する方式を考案し，実際の楷書体の分類に適用した結果，十分な精度で分類可能であることを確認した．今後は，一般的な手書き楷書体漢字にも適用していきたいと考えている．

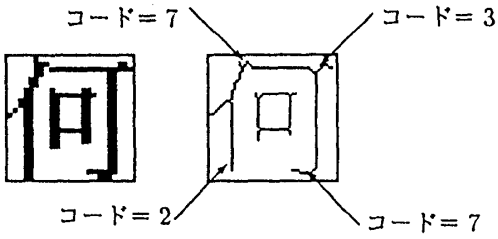

図 4: 細線化処理による変形

表 1: 特徴 Code 表

番号	特徴図	条件
Code 0	片一方空白	隅にもっとも近い Node と隅の距離 L, L >= 23 ドット
Code 1	" 一" 横線	| slope | <= 0.2
Code 2	" 丨" 縦線	| slope| >= 5
Code 3	" ／" 負斜線	−5 < slope < −0.2
Code 4	" ＼" 正斜線	5 > slope > 0.2 線の両側端点（点）
Code 5	" ＼" 正斜線	5 > slope > 0.2 線の両側，端点ではない
Code 6	" ＋，×" 二線以上の交差	隅にもっとも近い Node とつながる Node のつながる Node の数 >= 4
Code 7	" 「,」, 口" 角	曲率 curve >= 100

表 2: 4 桁コードと対応する文字

4 桁コードと対応文字	4 桁コードと対応文字
Code 7377 易喝完冠紀鬼客居胸局屈兄月見鋼号昆腸周冗親成即胎貯釣鳥腸胴銅届尼肌晩肥尾胞凡民眠明銘鳴野用容卵郎朗	Code 7777 囲印因園凹回官観寄器曲圏固口国困四自囚昭唱晶申図団中宙田島凸日白冒品富密目由略留
Code 4547 沿浴	Code 6374 賀想
Code 4514 渋渉漂涼	Code 3247 俗代伐

参考文献

[1] 陸師成. 辞書. 文化図書公司, 台湾台北市, 1992.

[2] 鳥脇純一郎. 画像処理のためのディジタル画像処理.

[3] 長尾 真. パターン情報処理.

[4] 奥村彰二 and 前田正弘. 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.32,No.1:pp.50-61, 1991.

[5] 大辞林. 1992.