

部首の抽出による類似文字の詳細識別 Fine Classification of Resembling Characters by Radical Extraction

鈴木 道孝[†]
Michitaka Suzuki

伊藤 彰義[‡]
Akiyoshi Itoh

1. はじめに

統計処理による手書き文字認識ではごく少数の類似文字[1][2]に誤読が集中するが、それらの詳細識別に構造解析的な手法が有用であることが報告されている[3]。本報告では、類似文字の「推」と「推」におけるような木偏と手偏の違いを構造的に識別する方法を提案する。この問題は、単に類似文字の識別のためだけでなく、部首を単位とした認識の可能性をも示すものである。本報告で部首とは、必ずしも漢和辞典で規定されるものではなく、単に文字を構成する部品という意味である。

2. 識別の流れ

類似文字の片方が最終候補に残ったとき、以下の図1の流れに従って、類似文字の対のどちらを答えとするかを決定する。

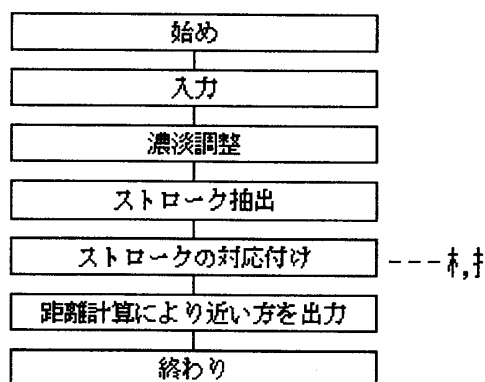


図1 詳細識別の流れ

入力された多値の文字画像データに対して、まず、濃淡値の調整を行う。一般に、一文字を筆記する間にも筆圧は変動し、濃度値はストロークの中心線上においてもばらつきがある。ばらつきによって一箇所での高い濃淡値の最大が起こると、その値で全ての領域の濃淡値が規格化されるので、かすれ領域が出やすくなる。これを回避するために、ストローク抽出に使われるモデルの濃淡値ヒストグラムと入力の濃淡値ヒストグラムが一致するように入力の濃淡値を調整する。

ストロークの抽出は、図2のような巾を持った線分の連なり（堤防モデルと呼ぶ）によって多値の画像データを最小二乗法でフィッティングすることによって求められる

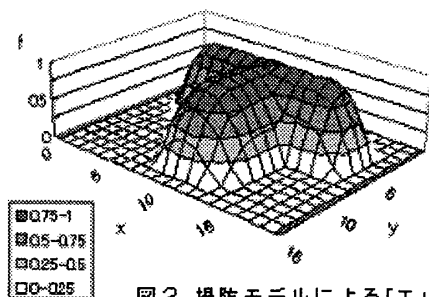


図2 堤防モデルによる「T」

[3][4]。図2では、巾が3の線分(5,5)-(15,5)と巾が4の線分(10,5)-(10,10)が重なり、「T」字形を形成している。

堤防モデルによるストロークの抽出は、従来の細線化の方法に比べ、いくつか特長を持っている。それらは、(1)ひげ状のノイズが出ない、(2)ストロークの交差部が正確に求まる（従来の細線化では三叉路の組合わせが求まる）、(3)多値データを扱える、(4)ベクトルデータ（端点や角点の座標）としてのストロークが直接求まる、などである。

この方法で求められたストロークの例を図3に示す。図3では、濃淡調整済みの画像データの上に、求まったストロークの骨格線分を巾の情報をぬいて白色で重ねて表示している。例として選んだ文字は、ETL9Gの中で明らかに誤記された文字である。

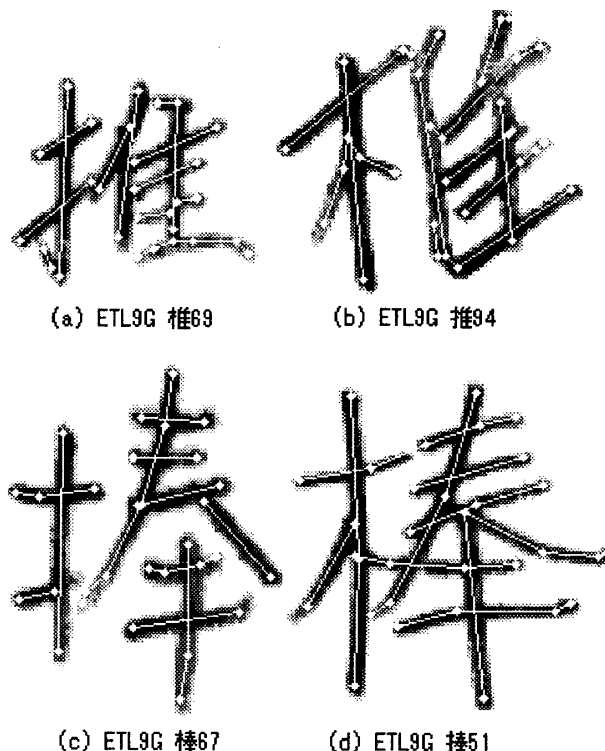


図3 ETL9の誤記文字と抽出された骨格線分

[†] 日本大学大学院理工学研究科

[‡] 日本大学理工学部

抽出されたストロークの点座標データは、外郭長方形を正方形にする正規化をし、角度変化の小さい内点を削除したのち、標準パターンのストロークと対応付けをする。対応付けは、距離

$$d_k = \sum_{i=1}^{n_k} \left(\frac{x_i - \mu_{ki}}{\sigma_{ki}} \right)^2 \quad (1)$$

を最小にするような配置をキューブサーチ[5]により求めることで得られる。ここで、 k は部首の種類、 i はそれを構成する成分、 x_i はその入力値を表し、 μ_{ki} 、 σ_{ki} はそれぞれ学習によって得られる標準パターンの平均値、標準偏差である。類似文字対の両方について対応付けを行い、そのとき求められた距離を使い、次式の量を用いて、近い文字を判断する。

$$a_{kk'} = r_{kk'} \frac{d_k/n_k}{d_{k'}/n_{k'}} \quad (2)$$

ここで、 k' は対抗する部首を表す。 $a_{kk'} < 1$ のとき k を正解とし、 $a_{kk'} > 1$ のとき k' を正解とする。次元数の違う距離を比較しなければならないので、次元数あたりの距離の比に $r_{kk'}$ による補正因子が必要になる。 $r_{kk'}$ の値は認識率が最大になるように学習によって求める。

3. 識別実験

ETL9G の「推」と「推」および「棒」と「棒」の2組の字種 800 パターンを対象に詳細識別の実験を行う。類似文字対の二者択一問題の正解率を算出し、それを認識率とする。対象データには、図3の例以外にも誤記がある。「棒」の 68 番、96 番、105 番、135 番は、いずれも手偏を書くべきところに木偏が書かれている。これらのすべての誤記データは、正しいカテゴリに移して実験を行った。

ストロークの対応付けを行うために事前分布が必要であるが、その平均の配置を図4のようにとる。木偏の第2画と第3画のストロークは、重心の x, y 座標、角度、長さを独立変数としてとる。木偏の第1画、第4画については、図3(d)で例示されているように、傍の部分と接触し、誤認識を引き起こすことがあるので、ストローク全体ではなく、左側の端点と角度のみを扱う(半ストローク)。したがって、木偏の自由度は、14 となる。

事前分布には標準偏差も必要であるが、角度を除く標準偏差はすべて 0.05 として、角度に関する標準偏差は 20 度とした。事前分布を多少変えても学習結果は同じ値に収束する。「推」を用いた木偏の学習結果は表1のようになった。

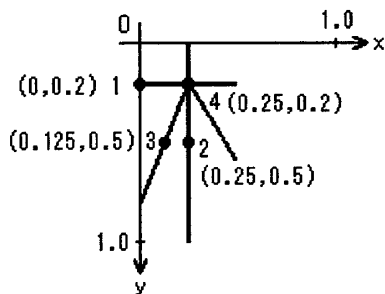


図4 木偏の事前配置

表1 「推」から学習した木偏のデータ

画	x 座標	y 座標	角度(度)	長さ
1 b	0.08±0.05	0.82±0.08	-16±10	-
2 a	0.23±0.05	0.52±0.04	-89±3	0.94±0.07
3 a	0.12±0.04	0.61±0.08	-61±8	0.17±0.11
4 b	0.23±0.06	0.51±0.09	33±16	-

a: ストローク. x, y は重心の座標.

b: 半ストローク. x, y は左端点の座標

「推」の手偏についても、同様の学習を行う。この場合の自由度は 10 になる。認識率最大の条件から $r_{kk'} = r(\text{推}, \text{推}) = 1.4$ を得た。同様に「棒」「棒」から、木偏と手偏の学習をし、 $r_{kk'} = r(\text{棒}, \text{棒}) = 1.4$ を得た。学習から得られた両者の部首の形状はほとんど等しくなっている。

認識実験は、同じ字種からの学習データは使わずに行う。すなわち、「推推」の識別には、「棒棒」のデータを使い、「棒棒」の識別には、「推推」のデータを使う。結果を表2に掲げる。

表2 類似文字の認識率

項番	対象文字対	認識率(%)
(1)	「推」と「推」	98.25
(2)	「棒」と「棒」	96.25

表2の認識率が 100%に至らない原因としては、ストローク間の相関を考慮せず、個々のストロークを独立に標準パターンのストロークと比較していることがあげられる。また、「棒棒」の認識率が「推推」より悪くなっている理由は、次のように考えられる。「推推」は字足をそろえて書かれるのに対し、「棒棒」は図3(c)の例のように傍の最終画が偏より下まで延びて書かれる傾向がある。すると、それによって規格化された偏の垂直方向の長さが小さくなり得、変動が大きくなり認識率が下がると考えられる。

4. むすび

構造解析が類似文字の詳細識別に有効であることを確認した。部首を単位としたオフライン手書き文字認識の可能性を確認した。

参考文献

- [1] 孫, 安倍, 根元: "部分整合領域の自動学習による手書き文字の詳細識別に関する一手法", 信学論 D-II Vol.J78-D-II No.3 pp.492-500 (1995).
- [2] 鈴木, 加藤, 根元, 市村: "2次混合関数を用いた類似文字識別手法", 信学論 D-II Vol.J84-D-II No.8 pp. 1557-1565 (2001).
- [3] 鈴木 道孝: "骨格ベクトルによるオフライン手書き文字の構造解析", 情報科学技術フォーラム(FIT) I-016 (2004).
- [4] 鈴木 道孝: "手書き文字画像データから骨格ベクトルを抽出する", 情報科学技術フォーラム(FIT) 第3分冊 pp.37-38 (2003).
- [5] 慎, 迫江: "筆順・画数自由オンライン文字認識のための画対応決定法", 信学論 D-II Vol. J82-DII No. 2 pp. 230-239 (1999).