手書きカナ文字の切り出し信頼度の定量化とその活用法

4R－2

堀田 悦伸　直井 聡

（株）富士通研究所

1　はじめに

フリーピッチで書かれた文字列から1文字ずつ切り出していく場合、複数の切り出し候補を求め、それらに対して文字認識処理を施し、切り出し結果を一意に確定するという方法がある。しかし処理時間の増大、誤切り出し候補を正解と誤る誤認識を防ぐためには、切り出し候補を限定したうえで認識処理を用いる必要がある。一方、切り出しパラメータの評価に際しては、基準となる文字列データベースがないために、切り出しの正解／失敗の判定を人間が目視で行なう必要があり、評価データ数の絶対数を多くしづらい、という問題がある。 本報告では手書きカナ文字を対象に、少ない評価データ数を考慮したうえで切り出しの信頼度を定量化する手法について述べる。

2　切り出し処理の概要

カナ文字には、濁点・半濁点や左右にパターンが分離した「ル」「ハ」などの文字（以下、分離文字）が存在する。前者は隣接する文字とオーバーハングしやすく、後者は分離文字であるため、外接矩形の並びをみていくだけではこれらを正確に処理することができない。そこで、本方式では連結黒画素の外接矩形座標に基づく処理に加え、連結黒画素自体の形状特徴に基づく処理を併用する。以下に両処理の内容とそこで用いるパラメータについて示す。なお実際には”文字の切り出し”とは、連結黒画素の”統合”処理となる。

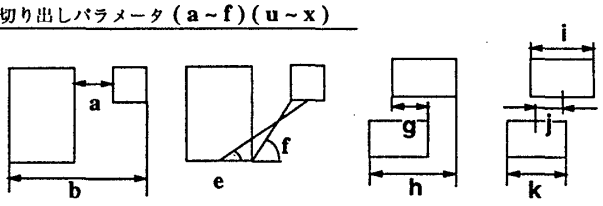
3　切り出し信頼度の定量化

3.1　切り出しパラメータ評価

まず学習データに対し、連結黒画素同士を統合すべきかどうかについて目視で判定し、判定結果とそのときの各パラメータ値を記録する。そして統合すべき群とそうでない群の2群に対し、N次元空間上（パラメータ数 N の場合）で判別分析処理により判別面を求める。次に2群の各データに対し、判別面からの距離に関するヒストグラムをとる。求めたヒストグラム分布に対して分布の平均値と分散値を求め、正規分布近似する。そして2群の正規分布曲線が交わる領域に入力パラメータセットが含まれる場合に、切り出し処理と認識処理を融合する。分布曲線の両端を決定する際には、データ数が少ない場合を考慮して、ヒストグラム分布の正規分布への適合度を重みとして用いる。

		処理内容
外接矩形処理	配置処理	パラメータ a～f の値に基づきパターンを統合
	上下重なり処理	パラメータ u～x の値に基づきパターンを統合
黒連結成分形状処理	濁点処理（ツ，シ など）	濁点候補パターンを抽出後、パラメータ p～r の値に基づきパターンを統合
	分離文字処理（ハ，ル など）	分離文字候補パターンを抽出後、パラメータ p～r の値に基づきパターンを統合

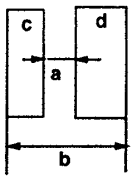
切り出しパラメータ（a～f）（u～x）



$c = a/b$
$d = b/MX$
（MX: 外接矩形平均幅）

$u = g/h$
$v = g/k$
$w = g/i$
$x = j/MX$

切り出しパラメータ（p～r）



c, d: 外接矩形面積
MX: 外接矩形平均幅
MY: 外接矩形平均高さ

$p = a/b$
$q = b/MX$
$r = (c \times d)/(MX \times MY)$

¶A Reliability of Handwritten Kana Character Segmentation and Its Usage
Yoshinobu Hotta, Satoshi Naoi
Fujitsu Laboratories Ltd.

3.2 信頼度算出

連結黒画素同士について、各切り出しパラメータを算出し、そのパラメータセットの判別面からの距離を求める。それがヒストグラム分布上で2群の交わり領域に属する場合に、交わり領域内での位置に基づき切り出し信頼度を算出する。

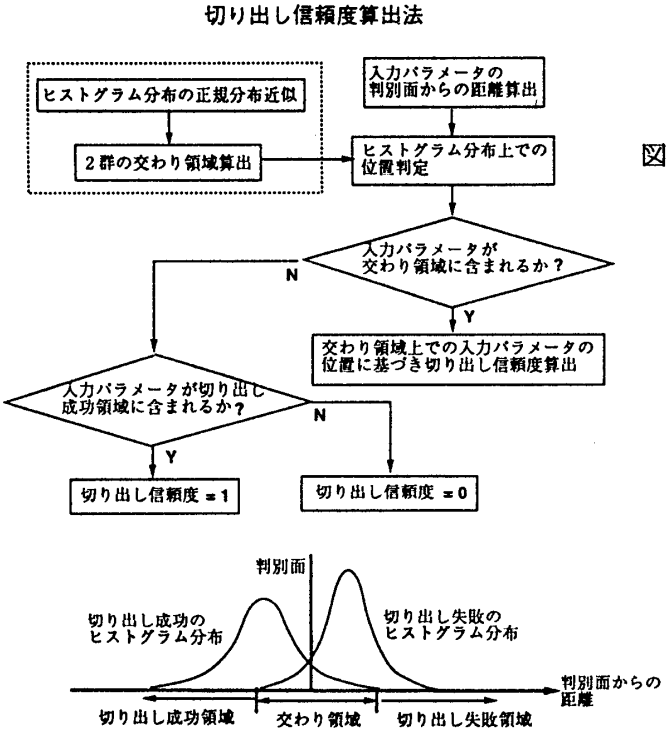


図 1: 切り出し信頼度算出

が困難であることがわかる。

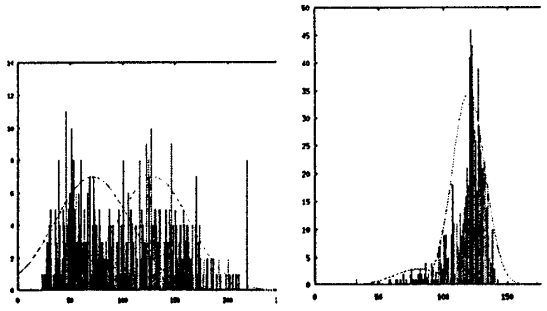


図 2: 外接矩形処理（配置処理、上下重なり処理）

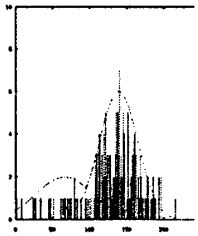


図 3: 濁点処理

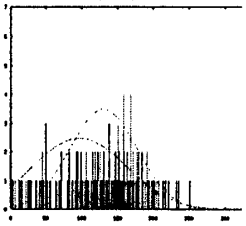


図 4: 分離文字処理

4 実験結果

手書きカナ文字データに対して行なった切り出しパラメータの評価結果を示す。横軸は判別面に垂直にとった座標軸で、100 の値のところが判別面に相当する。縦軸は頻度である。実際のヒストグラム分布データとそれを正規分布近似したものを表示してあり、右側の分布が切り出し正解となるパラメータ分布、左側が切り出し失敗となる分布である。これより外接矩形処理、とくに外接矩形同士が重なった場合の処理については、正解／失敗パラメータ分布の分離度が高く、切り出しパラメータだけで切り出し処理が可能であることがわかる。濁点処理についても、正解／失敗パラメータ分布の分離度が比較的高い。一方、分離文字処理については正解／失敗パラメータ分布の分離度が低い。これは、「ル」という文字を「ノ」「レ」と間違う場合やその逆の場合があり、切り出しパラメータだけで判断すること

5 おわりに

手書きカナ文字の切り出しに関して、切り出し信頼度の定量化とそれを用いた認識処理との融合方式について述べた。今後は実験による評価を進めるとともに、ヒストグラム分布の正規分布近似の妥当性評価、2群の判別に関する判別分析以外の手法も検討していく予定である。

参考文献

[1] 米倉:" フリーピッチ帳票の文字切り出しにおける 切り出し確度 ", 信学会, 秋季全国大会, (1995).

[2] 石寺, 西脇, 山田:" 手書き住所読み取りのための文字切り出し方法 ", 信学会, 秋季全国大会, (1995).

[3] 堀田, 直井:" 文字列特徴の定量化に基づく手書き数字分離法 ", 信学会, 秋季全国大会, (1993).