特徴選択による文字認識高速化の一手法

5P-8

大田　裕、　西村　康、　富本　哲雄
(松下電器株式会社　無線研究所)

1、まえがき

漢字ＯＣＲにおいては、多次元の特徴を使用する特徴マッチングが広く使用されている。しかし特徴数を増していくつれて、マッチングの計算量が増し、認識時間が長くなる。

そこで我々は、文字ごとに識別に有効な特徴を選択し、この特徴を多段階に構成した識別に使用して認識の高速化を図る手法について検討を進めた。

ここでは、本手法の内容と特徴選択の有効性を確認する予備実験および多段階構成の認識についての実験結果について報告する。

2、分離率の定義

特徴の識別力を数値化するため、ある特徴ｋによってカテゴリｉ〔平均m_i，標準偏差σ_i〕よりカテゴリｊ〔平均m_j，標準偏差σ_j〕が分離される確率である分離率Ｂ(k)ij を図１の斜線部分の面積になるように定義する。

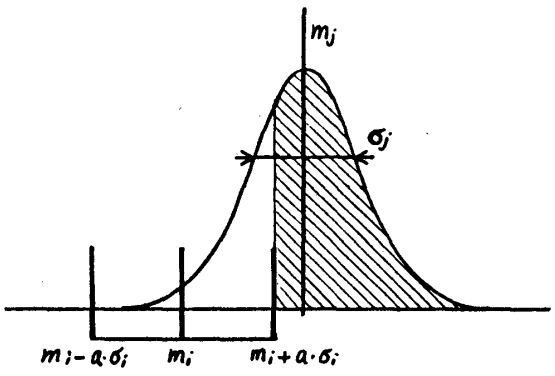

図１

3、逐次特徴選択の方法

各特徴の分離率の総和Ｕを式１で算出し、Ｕが最大の特徴を逐次的に選択する。〔カテゴリ数をＭとする〕

$$U(k)i=\sum_{\substack{j=1\\ j\neq i}}^{M} B(k)ij \qquad \cdots \text{式１}$$

ただし第２番目の特徴選択からは、それまでに選択したＮ個の特徴〔Ｂ(m)ij,m-1..N〕と識別力が重複するのを避けるために分離率Ｂの代りに式２で算出した累積分離率Ｂ'を使用して総和Ｕを求める。

$$B'(k)ij=1-(1-B(k)ij)\cdot\prod_{m=1}^{N}(1-B(m)ij) \qquad \cdots \text{式２}$$

4、特徴選択の有効性を確認する予備実験

前述の方法で各カテゴリごとに特徴を選択して認識する。選択する特徴数の上限を順次変えて、認識率・マッチング時間の変化をみた。実験の方法は次のとおりである。

文字数	３１６カテゴリ
全特徴次元数	３７０次元
使用データ	ＥＴＬ８Ｂ２
学習サンプル	４０セット
未知サンプル	４０セット

〔実験結果〕図２に示す。

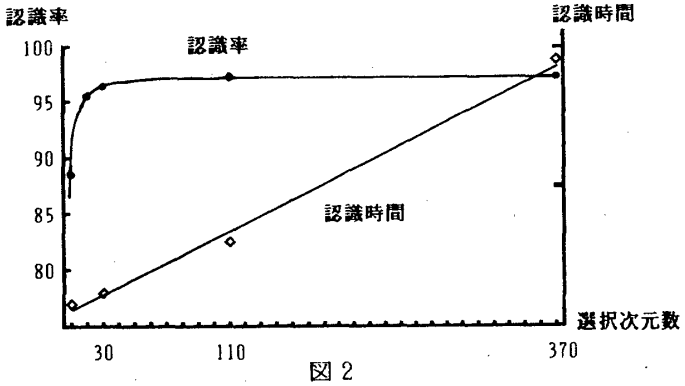

図２

High speed character recognition method using feature selection

Hiroshi Ohta ,　Yasushi Nishimura ,　Tetsuo Tomimoto
Matsushita Electric Industrial Co., Ltd.

・　特徴３０次元で、ほぼ３７０次元の特徴の認識性能を得ており前述の特徴選択の方法が有効であることがわかった。

・　特徴選択が認識時間の短縮につながることがわかった。

5、多段階識別の構成手法

直列に並べた３段階の識別で構成した。第１段と第２段は、前述の方法でカテゴリごとに３７０次元の特徴から選択した３０次元で識別する。第３段は３７０次元の全特徴を使って識別する。各段階の処理は次のとおりで、従来の予め定められた数に候補カテゴリ数をしぼり込む大分類とは異なる構成である。

【　第１段　】

３０次元の特徴〔平均m_k, 標準偏差σ_k, k=1 ～30〕について入力文字の特徴値x_kが、

$$|(x_k - m_k)/\sigma_k| < c_1$$

をみたすカテゴリを候補とし、候補カテゴリ数が１ならばそのカテゴリを出力し、第１段で認識処理を終る。〔c_1 は定数〕

【　第２段　】

３０次元の特徴〔平均m_k, 標準偏差σ_k, k=1 ～30〕について入力文字の特徴値x_kが、

$$\sum_{k=1}^{30} |(x_k - m_k)/\sigma_k| < c_2 \cdot 30$$

をみたすカテゴリを候補とし、候補カテゴリ数が１ならばそのカテゴリを出力し、第２段で認識処理を終る。〔c_2 は定数〕

【　第３段　】

370次元の特徴〔平均m_k, 標準偏差σ_k, k=1 ～370 〕で入力文字〔特徴値x_k〕と各カテゴリの辞書との距離Ｄが最小のカテゴリを出力する。

$$D = \sum_{k=1}^{370} |(x_k - m_k)/\sigma_k|$$

6、多段階識別による認識実験

前述の方法で第１段、第２段の定数c_1、c_2を順次変えて、認識率・マッチング時間の変化をみた。実験の方法は次のとおりである。

文字数		９５２カテゴリ
特徴次元数	第１、２段	３０次元
	第３段	３７０次元
使用データ		ＥＴＬ８Ｂ２
学習サンプル		８０セット
		（奇数データセット）
未知サンプル		８０セット
		（偶数データセット）

［実験結果］図３に示す。

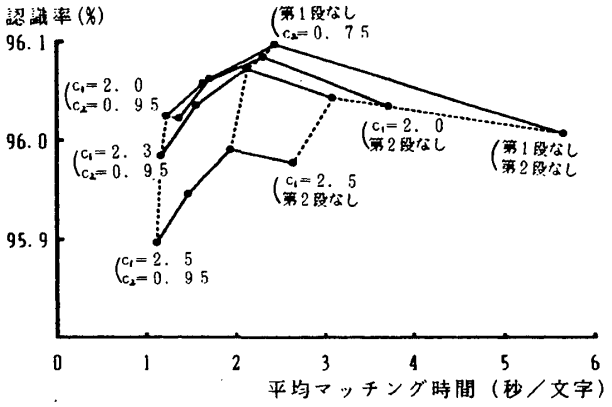

図３

・　c = 2.0, c = 0.95 で、認識率を下げることなく平均マッチング時間を約１／５に減少させることができることがわかった。

6、まとめ

各カテゴリの平均とマッチング度の高い入力文字について計算量を削減することによって、平均マッチング時間を減少させた。そのための特徴選択として分離率による特徴選択を提案した。

今後は、多段階識別の構成方法についてさらに検討を進めていきたい。