I-77　クラスタリング手法による文字認識辞書圧縮の検討

A Study on Character Recognition Dictionary Compression Based on Clustering Method

川又　武典†
Takenori Kawamata

1．まえがき

我々は、携帯電話、携帯端末などの小型機器向けのオンライン文字認識方式の検討を行っており、一筆書きした入力パターンから P 形フーリエ特徴を抽出することにより、指筆記文字を読み取り可能な省メモリ英数カナ認識方式[1]、ストローク間情報を用いることにより続け字にロバストで、かつ漢字まで読み取り可能な文字認識方式[2]を開発してきた。しかし、漢字まで読み取り可能な文字認識方式においては、128 次元の特徴を用いるため、例えば日本語（4717 カテゴリ）の場合には、特徴のみで約 590KB の認識辞書容量が必要となり、そのままでは、小型機器への適用が困難であった。また、文字認識では認識処理時に文字単位の特徴ベクトルの復元（解凍）を行う必要があり、必ずしも汎用のデータ圧縮方式が有効とは限らないという問題点があった。そこで、今回は、予め正準判別分析法により認識精度の低下を抑えた特徴の圧縮を行うと共に、圧縮特徴ベクトルをクラスタリングし、クラスタ内の最小値ベクトルからの差分値を保持することにより、高効率かつ低演算量で特徴の復元が可能な圧縮方式の検討を行った。

2．正準判別分析法による特徴圧縮

2.1　原特徴

オフストローク（実際に筆記された実ストローク間の情報）を反映させ、かつ2方向のぼかしを行った 8 方向コード分布特徴である。なお、領域分割を 4×4 としているので、最終的に 128 次元の特徴となる。

2.2　正準判別分析法による特徴圧縮

上記 128 次元の原特徴に対して、正準判別分析法を適用することにより次元圧縮を行う。表 1 に示す、学習・評価データを用いて、辞書設計、及び次元数を変化させた場合の認識率・第 10 位分類率の評価実験を行った結果を図 1 に示す。なお、辞書設計に用いた認識対象文字は、非漢字 348 文字、漢字 3693 文字（第 1 水準漢字 2965 文字＋第 2 水準漢字 728 文字）の合計 4041 文字で、非漢字についてはサブカテゴリ化を行っているため、総カテゴリ数は 4717 である。また、認識実験における認識モードは、全字種混在認識で、評価データ に含まれる認識対象外の非漢字 32 文字は除いた。

表 1．学習・評価データ

	ﾃﾞｰﾀﾍﾞｰｽ名称	ﾃﾞｰﾀ数
学習	当社ｵﾝﾗｲﾝ設計ﾃﾞｰﾀ	750ﾊﾟﾀｰﾝ/文字
評価	HANDS_kuchibue_d-97-06[3]	120人分

†三菱電機（株）情報技術総合研究所, Information Technology R&D Center, Mitsubishi Electric Corp.

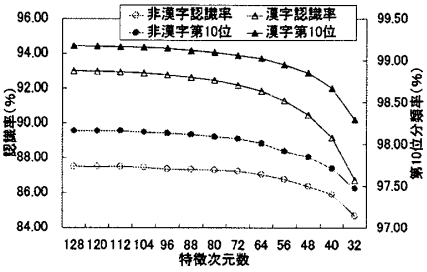

図 1．圧縮特徴の次元数と分類率

図 1 に示す結果より、漢字に比べて非漢字の場合は、特徴次元数削減の影響を受け難いことが分かる。ここで、圧縮特徴の次元数は、当社の従来方式（構造解析的手法）[4]の認識率と同等の性能を実現可能な 56 次元を選択した。これにより、圧縮特徴をバイト型で量子化した場合は、特徴ベクトルの総容量は 258KB となる。

3．クラスタリング手法による特徴圧縮

3.1　圧縮特徴の差分値

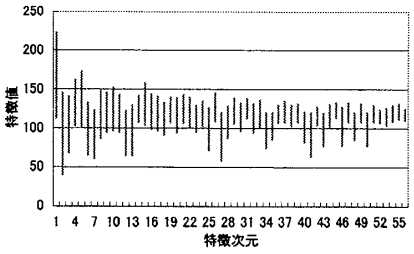

図 2．圧縮特徴ﾍﾞｸﾄﾙの分布

図 2 は、日本語の圧縮特徴ベクトル（4717 カテゴリ）について、次元毎に特徴値のダイナミックレンジをプロットした図である。図に示すように圧縮特徴においては、低次特徴ほどダイナミックレンジが大きく、高次特徴ではダイナミックレンジが小さくなる傾向がある。これより、特徴次元毎に最小値からの差分値を保持することにより、特に高次成分の量子化ビット数を削減することが可能となる。

3.2　適応的ベクトル量子化の応用

適応的ベクトル量子化は、ベクトル空間を等分に賽の目切りにするのではなく、データの分布に応じたクラスタリングを行うことにより、代表ベクトルに置き換えたときの誤差が小さくなるように量子化を行う。文字認識辞書における各文字の特徴ベクトルも、文字の類似性により分布には偏りがあると考えられる。そこで、クラスタリング後のクラスタ毎に、ク

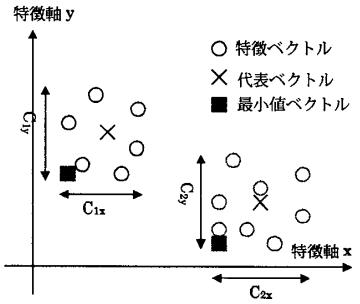

図 3．ｸﾗｽﾀﾘﾝｸﾞによる分類

ラスタ内の最小値からの差分値を各カテゴリの特徴ベクトルとして 3.1 と同様に保持することにより、量子化ビット数の削減を図る（図３）。図４は、実際に図２の例について、２つのクラスタに分類した場合の特徴ベクトルについてプロットしたものである。図に示すように、図２に比べて各特徴次元におけるダイナミックレンジが抑えられていることが判る。

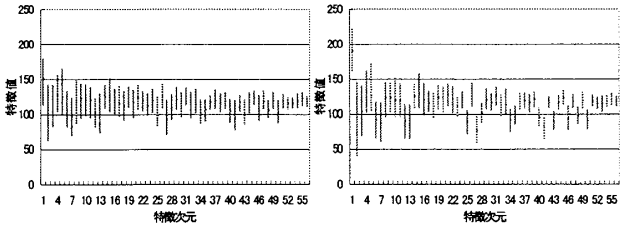

図４　クラスタリング後の圧縮特徴ﾍﾞｸﾄﾙの分布

3.3　クラスタリング手法

3.2 でクラスタリング手法を用いた特徴ベクトルの容量削減手法について説明した。本削減手法では、特徴ベクトルのクラスタリング性能がポイントとなる。そこで、今回の検討では、階層的クラスタリング手法として、群平均法、重心法、可変法（β=0.25）、ウォード法の４種類、非階層的クラスタリング手法として、K-means 法を用い、性能比較を行った。

4. 認識辞書圧縮の評価結果

（１）クラスタリング手法による比較

各クラスタリング手法を日本語の特徴ベクトルに適用した場合の認識辞書容量の比較結果を図５に示す。なお、辞書容量の算出には以下の式を用いた。ここで、各クラスタでは、最小値ベクトル、平均値ベクトル、量子化ビット数ベクトルを保持するので、（３×Nd）としている。

$$S = (3 \times Nd) \times Nc + \sum_{i=1}^{Nc} (Nci \times Sci)$$

Ｎｄ：特徴ベクトルの次元数（56 次元）
Ｎｃ：クラスタ数
Ｎｃｉ：クラスタｉの特徴ベクトル数
Ｓｃｉ：クラスタｉの圧縮後の特徴ベクトルバイト数
Ｎ：特徴ベクトル数（日本語：4717、簡体字 6763
繁体字 13063）

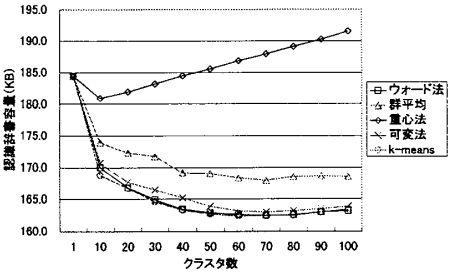

図５．クラスタリング手法の比較

図５に示すように重心法を除いて、クラスタ数を増加させるほど認識辞書容量は減少するが、クラスタ数が 70 前後で最も小さくなり、それ以降は、逆に増加する。重心法では、１つの大きなクラスタ以外は細かいクラスタが形成されるので、逆に辞書容量が増加している。本評価では、ウォード法、k-means 法の性能が良い。クラスタリングを行わない場合の認識辞書容量が 184.4KB、k-means 法の最小辞書容量（クラスタ数 60）が 162.4KB であるので、約 12%の辞書容量削減効果がある。

（２）中国語認識辞書における適用効果

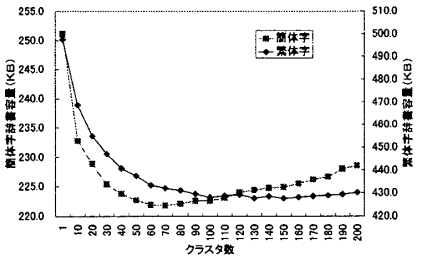

図６．クラスタ数と辞書容量

日本語において最も圧縮性能の高かった K-means クラスタリングを用いた圧縮方式について、中国語（簡体字、繁体字）に適用した場合の辞書圧縮性能を図に示す。図に示すように、簡体字では日本語と同様に最大 12%、繁体字では最大 14%の削減効果があることが判った。

5. クラスタ情報を用いた絞込み

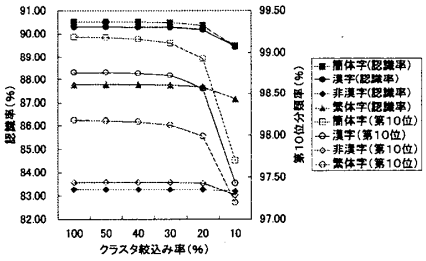

図７．クラスタ絞込み率と認識性能

ここでは、クラスタリング結果を用いた認識処理時の候補クラスタの絞込みによる、低演算量化の効果について評価を行う。具体的には、入力パターンから抽出された特徴ベクトルと、クラスタ毎の平均ベクトルとの距離を用いて、候補クラスタの絞込みを行った場合の認識率、第 10 位分類率の性能を評価した。結果を図７に示す。図に示すように、30%まで絞り込んだ場合でも、分類性能の劣化は僅かであり、本圧縮手法が認識処理時の演算量削減にも有効であることが判った。

6. まとめと今後の課題

正準判別分析法及びクラスタリング手法を用いた辞書圧縮により、低演算量かつ省メモリな文字認識方式の実現が可能となった。今後は、本手法を携帯電話、携帯端末などの小型機器に適用した場合の処理時間評価、認識率評価などを行う予定である。

[参考文献]
[1] 岡野他：“携帯電話向け文字入力システムの試作”　,第 64 回情処学全大　デ-20,2002
[2] 川又他：” ストローク間情報を用いたオンライン文字認識の改良”,信学総大 D-12-20(2001)
[3] 中川他：” 文章形式字体制限なしオンライン手書き文字パターンの収集と利用”,PRU95-110,pp.43-48(1995.9)
[4] 亀代他：” 方向コード特徴とストローク特徴を用いたオンライン文字認識方式”,信学総大 D-12-90(1997)