人文科学とコンピュータ　25－3
（１９９５．　１．　２７）

高次局所自己相関特徴による古文書かな文字認識

山田奨治
筑波技術短期大学

古文書かな文字(変体かな)を自動認識する際に特有な諸課題(文字のくずし、散らし書き、連綿体など)を整理した。そしてこれらの課題に有効な特徴抽出方式である、高次局所自己相関特徴について述べ、それを利用した文字認識実験を行なった。高次局所自己相関特徴は、文字の位相的な特徴を反映し、位置に関して独立で、画像内での加法性を持っており、これらの特性を利用することで、変体かなを厳密に文字を切り出すことなく認識することができる。本方式は文書を限定した上で、システムが文字を学習しながら専門家の読解を支援してゆくシステムへの応用が期待できる。

Japanese Old Kana OCR Based on Higher Order Local Autocorrelation Features

Shoji Yamada
Tsukuba College of Technology
e-mail:shyamada@cs.k.tsukuba-tech.ac.jp

4-12, Kasuga, Tsukuba 305 JAPAN

We arranged the problems of machine recognition of Japanese old Kana characters. The higher order local autocorrelation features, which is efficient to solve these problems, was described. We conducted some recognition experiments using this feature. The higher order local autocorrelation features reflects the topological features of characters. It is independent for position, and satisfies the low of addition. Using these features, old Kana could be recognized without strict character segmentation. Our method can be applied for specialists support system with auto character feature learning for single document.

1 はじめに

歴史学研究における基本的な作業のひとつに、古文書の読解がある。古文書の読解は、それだけでひとつの分野が成立するほど、高度に専門的な作業で、人手によるところの多い作業でもある。一方で文字認識技術の進歩により、活字の漢字かなまじり文をある程度の精度で自動的に読み取るシステムが安価に流通し始めている。もし文字認識技術を利用して、古文書の読解作業を支援するようなシステムが開発されれば、歴史学研究者にとって有用な道具になると思われる。この研究では、古文書かな文字（変体かな）を自動認識する際に特有な諸問題を整理し、変体かなの認識に有効な特徴抽出法を実際のデータに適用して検討する。

2 古文書文字認識の課題点

古文書文字認識は、基本的に手書き文字認識となる。したがって手書き文字認識特有の、筆跡の個人差という問題がある。またすでに書かれた文字であるため、リアルタイムの手書き文字認識で重要な情報として利用されている筆順情報がない。古文書文字認識特有の課題点として、さらに次のようなものがある。

1. 様々な「書風」がある。また、くずしの程度が同一文書上の同一文字においても異なる。
2. 「散らし書」のように、文字配置の自由度が大きい場合がある。
3. 毛筆の場合、筆圧や墨の残量によって線幅が非線形に変化する。また、線のにじみ・かすれがある。

1.の「書風」とは、文字の書き振りのことである。これは使う材料や書く人物によって異なるのはもちろんのこと、習字の師弟関係、流派、階級、時代によっても変化があらわれる [1]。このような無限ともいえるバリエーションのすべてに対応することは、極めて困難である。したがって現時点でとりうる現実的な選択枝は、文字の特徴をシステムに学習させながら、同一文書内での文字判別を支援させるようなシステムであろう。2.の「散らし書」は、通常のように紙面の右から左へ行を進めながら書いてあるのではなく、紙面の各部に文書を散らして書いてあるものをいう [2]。この場合は、活字の文字認識で使われている、最初に行間を切り出すという方法が使えないことになる。むしろ文字群をブロックとして取扱い、ブロックの中から文字を抽出できるような方法が望ましいと考える。3.に対しては、線幅に対して独立性を持った特徴量を採用する必要がある。

本研究は可能性の検討が主目的であるので、漢字は今回の検討範囲外とし、議論を変体かなに絞って進めることにする。変体かなは、万葉がなを草書で書いているうちにまとまったもので、一音に対して幾通りもの字がある。始めは現行のひらがなと同じ範疇に属するものであったが、明治半ば過ぎに現行のひらがな47字が国字として選抜され、除外されたひらがなが変体かなとして区別されることとなった [3]。

文字認識をする上で変体かな特有の課題点としては、以下のようなものがある。

4. 一音に対して幾つもの字源（字母）が存在する。
5. 同一文書上においても、同一音に対して異なる字母のかなが混在する。
6. 2字から4字程度を続け書きにする習わしがある（連綿体）。

4.、5.の課題は、現行のかな文字よりもはるかに多い字種を認識しなくてはならないことを意味している。そして6.の連綿体の存在が、変体かなの認識をとりわけ困難なものにしている。文字が続け書きになっているので、文字の切れ目がはっきりしない。また前文字の最終画が、次文字の第一画と共有される場合もある。

したがって、これらのような特徴を持つ変体かなの文字認識のための特徴量としては、次のような特性を持つことが要求される。

- 文字の位相的な特徴を反映すること。
- 位置に関して独立であること。

- 線幅、大きさに対して独立であること。

文字の位相的な特徴を反映する特徴量を用いることで、1. の書風、くずしに対する頑健性を得ることができると思われる。

位置に関して独立であることは、2. の文字配置の自由度の大きさに備えるために重要である。また線幅に対して独立であることは、3. の毛筆によるストローク幅の変化や、にじみ・かすれに関連して必要となる。

さらに、6. の連綿体の問題を解決する一つの方法として、

- 特徴量が加法性を持つこと

が有用であると考える。特徴量の加法性とは、文字「あ」の特徴量と「い」の特徴量の和が、文字列「あい」の特徴量に等しくなることである。特徴量に加法性があることで、連綿体から文字を厳密に切り出すことなしに、画像に含まれる文字を推定することが可能となる。

このような特徴を持つ特徴量として、高次局所自己相関特徴がある。次にこの高次局所自己相関特徴について検討する。

3 高次局所自己相関特徴について

3.1 定義と性質

時系列処理でよく用いられる自己相関関数の高次への拡張が、高次自己相関関数である。N次自己相関関数は、画面内の対象画像を $f(r)$ とすると、N個の変位 $(a_1, a_2, ..., a_N)$ に対して、

$$x^N(a_1, a_2, ..., a_N) = \int f(r)f(r+a_1)\cdots f(r+a_N)dr$$

で定義される [4]。通常の自己相関関数は、$N = 1$ の場合に相当する。この関数は自己相関関数であるので、平行移動に関しては不変である。また加法性を満たすためには、変位は参照点 r の回りの局所領域に限定されなければならない。

実際の計算にあたっては、参照点 r 近傍にN次までの局所パターンのマスクをかけ、該当するピクセル値の積和を求めるという手順をとる。変位の組み合わせの数は、Nに関して指数関数的に増大するので、応用にあたってはNを限られた次数に限定する必要がある。3×3 近傍の2次までの平行移動に関して独立な局所マスクパターンは25種類で、大津によって導かれている [5](図1)。

局所マスクパターンは、任意の近傍領域と最大次数について生成可能である。本研究では大津の局所マスクパターンに加えて、よりよい判別力を得るために 5×5 近傍の2次までの平行移動に関して独立な局所マスクパターンを独自に生成し使用した。そのパターン数は、0次が1、1次が12、2次が180で、合計193の特徴数となっている。

高次自己相関関数の性質を、先に検討した変体かな文字認識に求められる特徴量の性質と比較すると次のようになる。まず高次局所自己相関関数は、参照点 r 近傍の位相的な特徴を反映したものである。またそれは平行移動に関して独立で、画像内での加法性を持っている。残った条件は、位置に関して回転に対する独立性と、線幅・大きさに対する独立性である。線幅についてはエッジ抽出処理を行うことで、正規化を行った。他の条件については固定的に考えることで、以後の検討を進めることにする。

3.2 加法性を利用した文字の推定方法

高次局所自己相関特徴のもつ、画像内での加法性の特徴は、次のような式で表現することができる。認識対象画像の特徴ベクトルを y、既知文字の特徴ベクトルを F、既知文字の計数値ベクトルを x とすると、

$$y = F \cdot x$$

となる [4]。F はほとんどの場合、非正則行列になる。したがって、$F^\dagger$ を F の一般逆行列とすると、

$$\hat{x} = F^\dagger \cdot y \qquad (1)$$

を解くことで、認識対象画像の特徴ベクトルから既知文字の計数値を推定できる。

一般逆行列 $F^\dagger$ は、次のように求められる [6]。$N \times M$ 行列 F を特異値分解すると次のようになる。

$$F = U \cdot W \cdot V^T$$

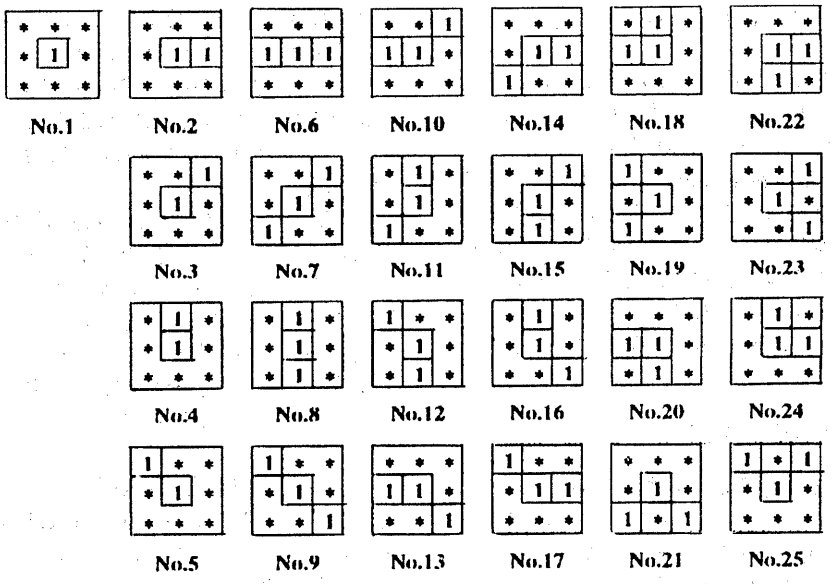

図 1: 高次局所自己相関特徴局所パターン (栗田 (1993) より引用)

ここでUは$N \times M$の直交行列、Wは$M \times M$の非負対角行列、V^Tは$M \times M$の直交行列Vの転置である。

したがって、Wの対角要素をw_jとすると、$F^\dagger$は次のようになる。

$$F^\dagger = V \cdot [diag(1/w_j)] \cdot U^T$$

連立一次方程式$y = F \cdot x$が優決定（特徴数が認識対象文字数よりも多い）であっても、劣決定（特徴数が認識対象文字数よりも少ない）であっても、解$x = F^\dagger \cdot y$は最小自乗の意味で最尤推定解になるという性質がある。

3.3 特徴の自動学習方法

画面内での加法性の特徴を利用して、複数の画像を学習データとし、その中に含まれる字種とその計数値のみを教師信号として与えることで、文字の特徴を自動的に学習させることも可能である。そのためには、誤差

$$\varepsilon^2 = \| y - F \cdot x \|^2$$

を最小とするようなFを求めればよい [4]。これは重回帰分析に他ならず、Fの推定値$\hat{F}$は、

$$\hat{F} = c_{xy} \cdot C_{XX}^{-1}$$

で求まる [7]。ここでc_{xy}はxとyの共分散行列、C_{XX}はxの分散共分散行列で、字種数をN、特徴数をM、学習データ数をLとすると、次のように表される。

$$c_{xy} = \begin{bmatrix} s_{x_1y_1} & \cdots & s_{x_1y_M} \\ & \cdots & \\ s_{x_Ny_1} & \cdots & s_{x_Ny_M} \end{bmatrix}$$

$$C_{XX} = \frac{1}{L} X'X$$

$$s_{x_Ny_M} = \frac{1}{L} \sum_{k=1}^{L} (x_{kN} - \bar{x}_N)(y_{kM} - \bar{y}_M)$$

$$X = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1N} - \bar{x}_N \\ & \cdots & \\ x_{L1} - \bar{x}_1 & \cdots & x_{LN} - \bar{x}_N \end{bmatrix}$$

したがって、認識対象画像の特徴ベクトル y' が与えられたとき、学習データから得た c_{xy} と C_{XX} とから、

$$\hat{x'} = C_{XX} \cdot c_{xy}^{-1} \cdot y' \qquad (2)$$

により、文字の計数値ベクトルの推定解 $\hat{x'}$ を得る。この方式は、文字領域を指定することなしに学習が可能となるという優れた特長があるため、変体かなにおける連綿体の認識に有効である。

4 変体かな文字認識実験

次に高次局所自己相関特徴を、実際の古文書のデータに対して適用し、文字の判別力を試験してみた。

4.1 学習済み文字の認識

文字を手作業で切り出しておき、3.2節の方法により、その高次局所自己相関特徴量と字種を教師信号としてシステムに与えた上で、少数の学習済みの文字をどの程度認識できるかを実験した。

実験に用いた古文書は、嘉永元年に筆記された弓術書「日置流秘歌」[8] の一部である。マイクロフィルム複写で入手した文書を、200dpi でスキャナー入力し二値化処理とエッジ抽出処理を行った。したがって実験画像は、複写による劣化や量子化ノイズを含んだ比較的条件の悪いものである。エッジ抽出処理はすべて Roberts のオペレータによって行った [10]。エッジ抽出後の実験文字列 1「しめてゆるすな」を図 2に示した。

まず実験文字列 1 から、七つのかな文字を手作業で切り出し、それぞれの特徴ベクトルを求めておく。特徴ベクトルを M 次元とすると、$7 \times M$の特徴行列 Fが得られる。Fの一般逆行列 $F^{\dagger}$を求めて、元の画像全体の特徴ベクトル y から式 (1) を計算すると、各文字の計数値の推定解$\hat{x}$が得られる。

3×3 近傍と 5×5 近傍の 2 次までの高次局所自己相関特徴について、判別を試みた結果を表 1に示す。元画像全体についての推定解のほか、その部分画像「しめて」「しめてゆる」についての推定解も求めてみた。表 1から、3×3 近傍マスクでは判別力が不足しているが、5×5 近傍マスクでは、最も近い整数値をとればよく推定できていることがわかる。

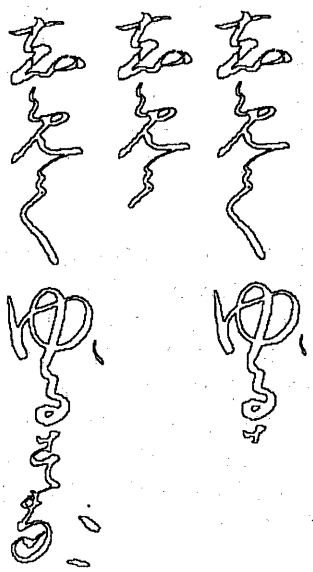

図 2: エッジ抽出後の実験文字列 1「しめてゆるすな」(左) とその部分文字列 1(中央) 部分文字列 2(右)

4.2 特徴の自動学習

次に個々の文字の切り出しを人間が行わず、3.3節の方法で、学習画像の高次局所自己相関特徴と、画像に含まれる字種とその計数値を情報として与えるだけで、システムが各文字の特徴量を自動的に推定する方法について実験を行った。学習画像の特徴ベクトル yと各字種の計数値ベクトル x から C_{XX}, c_{xy}を求め、式 (2) により認識対象画像に含まれる文字の計数値ベクトル$\hat{x'}$を推定した。実験データは同じ「日置流秘歌」の中から、一字のみ簡単な漢字を含む図 3「みな一やうとおもふもそうき」の部分を使用した。この例の中では、「う」と「も」がそれぞれ 2 文字づつ出現するが、字母が同じであればく

文字	3×3マスク			5×5マスク		
	全体文字列	部分文字列1	部分文字列2	全体文字列	部分文字列1	部分文字列2
し	0.99	0.82	1.04	1.19	0.99	1.23
め	0.38	1.33	0.59	0.66	1.30	0.77
て	1.18	-0.07	1.00	1.14	0.02	0.99
ゆ	0.98	0.28	1.06	0.95	0.19	1.05
る	0.06	-1.27	0.14	0.75	-0.54	0.79
す	1.96	-0.47	0.88	1.25	-0.17	0.06
な	1.48	0.61	0.33	1.17	0.07	0.12

表 1: 学習済み文字の認識の場合の文字の計数値の推定解 (実験文字列 1)

ずしの程度にかかわらず同じ文字とした。

原画像から複数の文字を含む部分画像を、重複領域を許しながら 12 枚取り出して、それぞれの 3×3 近傍の 2 次までの高次局所自己相関特徴量と字種の計数値を教師信号として $\hat{F}$ を求めた。そして原画像全体の特徴ベクトル y' から $\hat{x}'$ を推定した結果を表 2に示す。このように限定された字種であるならば、文字のくずしがあっても良好な推定解が得られることがわかる。

図 3: エッジ抽出後の実験文字列 2「みなーやうとおもふもそうき」

かな	計数値	推定解
み	1	1.02
な	1	1.02
ー	1	1.02
や	1	1.01
う	2	2.01
と	1	1.01
お	1	0.95
も	2	2.06
ふ	1	0.95
そ	1	1.11
き	1	1.00

表 2: 実験文字列 2・特徴自動学習の場合の文字の計数値の推定解

もうひとつの例として、文字配置の自由度が高く、30 字程度を含む画像に対して適用してみた。実験データは、小野鷺堂著「小久羅帖」に含まれる中納言行平の和歌である。画像は印刷書籍に収録されている文字 [9] を、300dpi でスキャナー入力したものである。自由度高く配置された文字群から、15 の矩形領域を手作業で切り出し (図 4)、エッジ抽出後のそれぞれの 3×3 近傍 2 次までの高次局所自己相関特徴と、画像にあらわれている字種とその計数値を学習データとした。ここで個々の文字の位置情報は、一切あたえていない。同一音であっても字母が異なるかな文字は、別の文字として取り扱った。

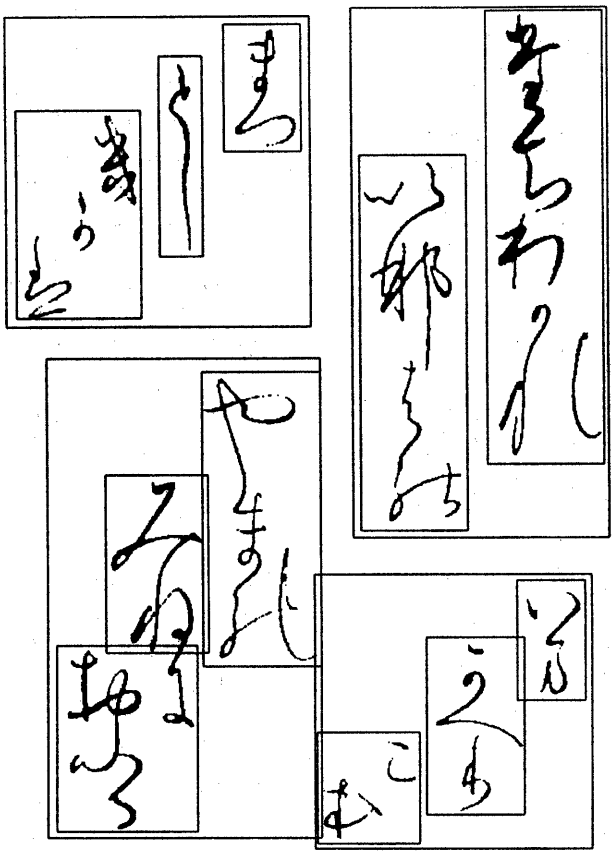

図 4: 実験文字列 3「たちわかれ いなはのやまの みねにおふる まつとしきかは いまかへりこむ」の矩形分割 (笹目 (1981) より引用)

特徴の学習後、元の画像全体の特徴ベクトルに対して、字種ごとの文字の計数値の推定解を、式 (2) で求めた。その結果を表3に示した。くずしの程度が異なる「い」「か」「の」「ま (末)」についても、よく推定できていることがわかる。

かな (字母)	計数値	推定解
い	2	2.03
お	1	1.00
か	3	2.96
き	1	1.00
こ	1	0.99
し	1	0.99
た	1	0.98
ち	1	0.98
つ	1	1.04
と	1	0.99
な	1	1.04
に	1	1.00
ね	1	1.00
の	2	2.03
は (者)	1	1.04
は (盤)	1	1.00
ふ	1	1.00
へ	1	0.98
ま (末)	2	2.03
ま (万)	1	0.99
み	1	1.00
む	1	0.99
や	1	0.99
り	1	0.98
る	1	1.00
れ	1	0.98
わ	1	0.98

表 3: 実験文字列 3・特徴自動学習の場合の文字の計数値の推定解

4.3 未学習文字の認識実験

次に高次局所自己相関特徴が、未学習文字に対してどの程度の判別力を持っているかを検討してみた。「あ」「い」「う」「え」「お」の音に対応する変体かなについて、代表的なくずしのパターンを 5 種の古文書辞典 [11, 12, 13, 14, 15] から採録した。画像は印刷書籍から 200dpi でスキャナー入力し、二値化とエッジ抽出処理をほどこした。くずしのパターン例を図 5に示した。5 音に対応する字種は 10 種類で、学習に用いた文字数は表 4の通りである。

これらの学習文字について 5 × 5 近傍 2 次までの高次局所自己相関特徴を算出し、特徴空間を求めておく。続いて各非学習文字について、その高次局所自己相関特徴が学習データの特徴空間内の最小自乗距離での最近隣点となる文字に判別した。ここでは判別分析などによる線形変換は行なわず、文字から得られた特徴量をそのまま距離算出に使用した。

認識結果は表 5の通りである。本方式での正認識率は 59 %であった。このように正認識率が低い水準に留まった理由としては、次のことが考えられる。第一に、辞典によって採録文字の拡大率に微妙な相違があり、現在の拡大に対する独立性のない特徴抽出方式では問題があるのは明らかである。第二に、学習データの特徴空間上での最近隣点となる文字に分類する方法では、十分な判別力が得られない。そこで線形判別分析や、マハラノビス汎距離を利用する方法が考えられる。しかし、そのためには文字をくずしの程度に応じてクラス分類しておく必要がある。そのようなクラス分類の有効な手法については、まだ検討されていない。クラスター分析による方法を若干試みてみたが、それが判別力向上に寄与し、かつ人間がみて不自然な分類であるという規準からみれば、満足のゆく結果は得られなかった。最後に、変体かなは古来書き手の勝手に種々の漢字を草書体に省略して書き、ほとんど一定する所を知らざる有様であるものである [16]。辞書には代表的なパターンが採録されているとはいえ、無限ともいえるパターンのすべてを限られた学習データから推定することは困難である。したがって、どのようなデータに対しても一定の認識率が得られるようなアプローチは、あまり現実味がないといえる。

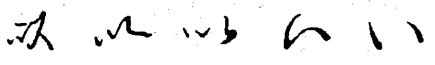

図 5: 「い」のくずしのパターン (浅井他 (1987) より引用)

かな (字母)	学習サンプル数
あ (阿)	14
あ (安)	16
い (以)	15
い (伊)	10
う (宇)	14
え (衣)	13
え (盈)	10
え (江)	14
え (得)	12
お (於)	18

表 4: 学習に用いたかな (字母) とサンプル数

かな (字母)	テストサンプル数	正判別数
あ (阿)	3	2
あ (安)	3	0
い (以)	3	2
い (伊)	1	0
う (宇)	4	4
え (衣)	3	2
え (江)	4	2
え (得)	4	3
お (於)	4	2
計	29	17

表 5: 未学習文字の認識実験

5 まとめと考察

本研究ではまず、古文書かな文字 (変体かな) を自動認識する際に特有な諸課題を整理した。それらの代表的なものは、文字のくずし、散らし書、連綿体などであった。そこでこれらの課題に有効な特徴抽出方式である、高次局所自己相関特徴について述べ、それを利用した文字認識実験を行なった。高次局所自己相関特徴は、文字の位相的な特徴を反映し、位置に関して独立で、画像内での加法性を持っている。これらの特性を利用することで活字に比べて記述の自由度が極めて高い変体かなを、厳密に文字を切り出すことなく認識することができる。学習済みの文字の認識の場合と、特徴の自動学習の場合の 30 字程度までについて良好な認識結果を得た。代表的なくずしパターンを学習した上での、未学習文字の認識実験では、10 字種について 59 %の認識率を得た。認識率が低い水準に留まった理由については、特徴抽出方式、判別方式が現時点では単純なものであることがあげられる。しかしながら、くずしのパターンは事実上無限に存在するため、いかなるくずしにも対応できるようなアプローチは、現実味に欠けると思われる。

本研究から考察される今後の方向としては、次の四点があげられる。第一にシステム全般に関し、文書を限定した上でその文書内の文字をシステムが学習しながら、専門家の読解を支援してゆくようなものを、まず目指すべきであろう。本研究の結果から、辞典を学習していかなる文字にも対応するようなシステムは、その次の段階に位置づけられるべきと思われる。第二に特徴抽出法にまだ改良の余地がある。現在の局所マスクパターンでは、回転に関して独立でない。また文字の大きさによる影響を無視している。今後は回転独立性を考慮した局所マスクを検討するとともに、判別に有効なマスクパターンの選択方法についても検討しなければならない。文字の大きさの影響の除去については、画像ピラミッドの構成などが考えられる。第三に特徴ベクトルの推定値 $\hat{F}$ の推定法について、何らかの拘束条件を設けることが可能であると考える。現在は最小自乗の意味での最尤推定解を使用している。しかし $\hat{F}$ に関しては、非負拘束を加えるのがむしろ自然である。第四にくずしの程度による文字の分類が必要である。古文書では同一文書中であっても、同じ文字のくずしの程度がまったく異なることがある。いかに位相的な特徴を反映する特徴量を用いても、それらを同一クラスに分類しようとするのは、高認識率を得るために不利となる。くずしの程度の分類は、それが判別力に対してプラスになるような手法で、しかも人間の目からみても自然な分類とならなくてはならない。クラスター分類による方法を試みてみたが、満足のゆく結果はまだ得られていない。これについても今後の課題である。

参考文献

[1] 伊木壽一: 日本古文書学 第三版, 雄山閣, p.80 (1990).

[2] 飯倉晴武: 古文書入門ハンドブック, 吉川弘文館, p.78 (1993).

[3] 笹目蔵之助: 続古文書解読入門, 新人物往来社, p.281-282 (1981).

[4] 栗田多喜夫: 柔らかな情報処理のための統計的手法の応用に関する研究, 電子技術総合研究所研究報告, 957, p.132-133 (1993).

[5] 大津展之: パターン認識における特徴抽出に関する数理的研究, 電子技術総合研究所研究報告, 818, p.178-179 (1981).

[6] W. H. Press et al.: Numerical Recipes in C 2nd. ed., Cambridge Univ. Press, p.59-70 (1992).

[7] 柳井晴夫, 高木廣文編: 多変量解析ハンドブック, 現代数学社, p.19 (1986).

[8] 岡山大学旧池田家文庫所蔵, 日置流弓道秘歌, 嘉永元年

[9] 笹目蔵之助: 前掲書, p.299 (1981).

[10] W. K. Pratt: Digital Image Processing 2nd. ed., A Wiley-Interscience Pub., p.498-500 (1991).

[11] 浅井潤子, 藤本篤編: 古文書大字典, 柏書房, p.356-364 (1987).

[12] 荒居英次他編: 古文書用字用語大辞典, 柏書房, p.644-652 (1980).

[13] 林英文: 古文書字叢, 柏書房, p.471-477 (1990).

[14] 若尾俊平他編: 近世古文書解読字典, 柏書房, p.309-312 (1972).

[15] 若尾俊平編: 図録古文書の基礎知識, 柏書房, p.118-123 (1979).

[16] 伊木壽一: 前掲書, p.211 (1990).