I-020

古文書標題の一括認識における非線形正規化法
A Nonlinear Normalization Method for Holistic Recognition of Historical String

中山 英久† 和泉 勇治† 加藤 寧† 柴山 守‡ 根元 義章†
Hidehisa NAKAYAMA Yuji WAIZUMI Nei KATO Mamoru SHIBAYAMA Yoshiaki NEMOTO

1. はじめに

古文書史料のデジタル化のため，翻刻作業の支援にOCRを用いるシステムの検討が行われており，古文書の毛筆文字に特有のつづけ字やくずし字など，手書き文字よりもさらに困難な技術的課題を克服すべく，研究が進められている．古文書文字列 (標題) を認識する際にも，従来の手書き文字列認識の技術を応用が試みられているが，個別文字の場合と比較して，なお一層困難である [1].

従来の手書き文字列認識では，切出し，個別文字認識を行ってから，その文字認識評価値を知識処理で統合する方法が一般的である．しかし，このような直列型の文字列認識システムでは，各モジュールのエラーが後段へと積み重なる可能性が高い．このことから，一括認識 (holistic recognition) の方法が検討されている [2]．一括認識とは，文字列全体をひとつの長方形と考え，個別の文字認識と同じストラテジーで認識を行うというものである．

文字列の一括認識を行う際に検討すべき課題は，文字列を長方形枠に拡大する手法 (画像正規化)，および，文字数が増加するにつれて特徴量の次元数が増加すること (特徴量抽出) である．手書き文字列における一括認識に関する従来研究として，文献 [3][4] が挙げられる．これらの文献においてそれぞれ画像正規化および特徴量抽出の問題に取り組んでいるものの，個別文字認識と同程度の性能が引出せているとは言い難い．また文献 [4] では，画像正規化に非線形正規化をそのまま適用した場合 [5]，線形正規化と同程度の正規化の効果しか得られないことが指摘されており，さらに文献 [6] にて指摘されているように，古文書文字列の認識においては，文字の正規化で字形変化が生じ，類似クラスへの誤読が多発するため正規化は容易ではない．

本稿では，個別文字認識で有効な非線形正規化法を，文字列認識へ適用する手法を提案する．非線形正規化の特性値を算出する際に対象区間を設定することで不自然な字形変化を行うことなく，文字列の大幅なゆがみを抑制した．また，正準判別分析を導入することにより，Wilks の Λ 基準による F 検定を用いた変数選択によって次元削減を行った．

古文書標題データベース「HCD2」の 5 クラス 112 サンプルを用いた，Leave-One-Out 法の識別実験により，従来の正規化よりも正読率が向上したことを示す．

2. 文字列一括認識システム

図 1 に古文書標題画像の一例を示す．この例の通り古文書の毛筆文字は，日本語手書き文字よりも文字サイズの変化が大きく，文字間の接触・入り込みが多発している．したがって，射影ヒストグラムによる文字切出しは困難である [6]．そこで本稿では，古文書標題の個別文字切出しは不可能であるという立場をとり，文字列単位での一括認識を行う．

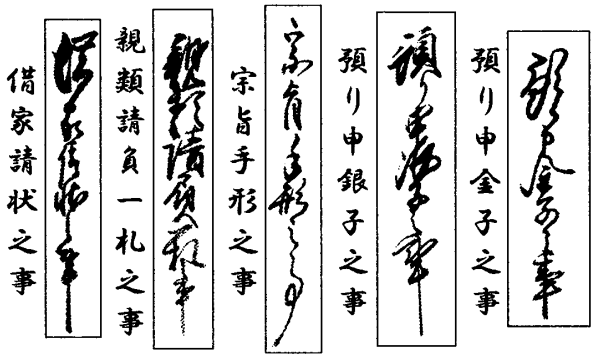

図 1: 古文書標題画像の一例

図 2 に本稿における文字列の一括認識システムの概要を示す．従来の個別文字認識 [7] とほぼ同じ認識システムを用いる．3 × 3 マスクによるスムージング，ノイズ除去 (孤立点除去) の前処理の後，長方形領域へ画像正規化を行い，方向線素特徴量 [8] を抽出する．文献 [8] は個別文字が対象であるため，正規化領域として $(64 \times 64)dots$ の正方形を設定しているが，本稿での対象は文字列であるため，n を想定文字数としたときの正規化領域を $(64 \times 64n)dots$ の長方形とする．よって特徴量の次元数は $d_0 = [7 \times (8 \times n - 1)] \times 4$ となる．本稿では文字数を $n = 7$ と想定し，原特徴量 $d_0 = 1540$ 次元を抽出する．

そして，正準判別分析を用いて 2 次判別を行う．本稿では特徴量の次元数が大きいため，Wilks の Λ 基準 [9] による F 検定を用いた変数増加法によって，d 次元への次元削減を行った (ただし $d \ll d_0$ とする)[10].

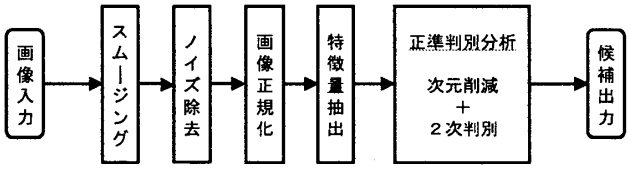

図 2: 文字列の一括認識システム

3. 文字列の非線形正規化

非線形正規化とは，識別に有効でない変形による特徴量への影響を抑える為に，位置や大きさなどの形状を整える画像処理手法のことであり，その手法により特性値から導出された変換関数による正規化領域への

†東北大学大学院情報科学研究科, GSIS, Tohoku University
‡京都大学東南アジア研究所, CSEAS, Kyoto University

拡大や縮小を行う．特性値の定義により様々な非線形正規化法が提案されており，個別文字認識の前処理として有効であることが確認されている．

3.1 非線形正規化

非線形正規化は，以下の手順で行う．変換前の外接方形領域を $x_1 \leq x \leq x_2, y_1 \leq y \leq y_2$ とし，変換後の正規化領域を $X_1 \leq X \leq X_2, Y_1 \leq Y \leq Y_2$ とする．

Step 1 特性値の設定

x 方向，y 方向にスキャンして，各方向別に特性値を設定する．この特性値には各種の定義があり，線形変換では定数 (S0)，文献 [11] ではストローク部分の黒画素 (S1)，文献 [12] ではストローク本数を考慮した線間隔 (S2)，そして文献 [13] ではストローク部分と背景部分の平均的な線間隔 (S3) である．各方向の特性値をそれぞれ $S_x(x,y), S_y(x,y)$ とする．そして，この逆数を線密度と定義し

$$\rho_x(x,y) = 1/S_x(x,y)$$
$$\rho_y(x,y) = 1/S_y(x,y)$$

とおく．なお文献 [14] では，ガウシアンフィルタによる平滑化を施したぼかし線密度が用いられている (S3G とおく)．

Step 2 変換関数の導出

線密度を各方向別に加算して，ヒストグラム $h_x(x), h_y(y)$ を得る．

$$h_x(x) = \sum_{y=y_1}^{y_2} \rho_x(x,y)$$
$$h_y(y) = \sum_{x=x_1}^{x_2} \rho_y(x,y)$$

そしてヒストグラムの累積値から，非線形正規化関数を求める．

$$\phi_X(x) = (X_2 - X_1)\Big[\sum_{k=x_1}^{x} h_x(k)\Big]\Big[\sum_{k=x_1}^{x_2} h_x(k)\Big]^{-1} + X_1$$
$$\phi_Y(y) = (Y_2 - Y_1)\Big[\sum_{k=y_1}^{y} h_y(k)\Big]\Big[\sum_{k=y_1}^{y_2} h_y(k)\Big]^{-1} + Y_1$$

非線形正規化関数は，ディジタル画像上の区分線形関数であるので，実際には線形補完により逆写像を求めることにより，非線形正規化画像を作成する．

$$\Big\{(x,y) \mapsto (X,Y) \Big| X = \phi_X(x),\ Y = \phi_Y(y)\Big\}$$

図 3 に各種非線形正規化画像の一例を示す．左側が変換前の画像，右側が変換後の各種画像 (S0〜S3,S3G) である．

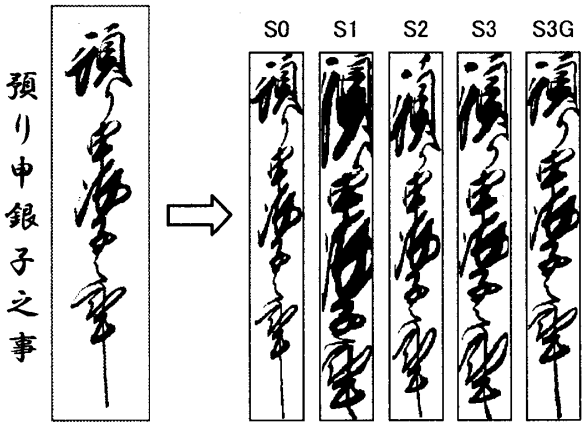

図 3: 非線形正規化画像の一例

3.2 画像分割による特性値の設定

従来の非線形正規化をそのまま用いた場合は，図 3 の S1,S2,S3 に見られるような，長手方向 (y 方向) へ過剰な変形が生じ，字形が崩れるという副作用が生じる．従来は正方形領域を扱っており，等方的であったため問題とならなかった．しかし文字列のような長方形領域を扱う際に字形が崩れるのは，非等方的な特性値を設定することに原因があると考える．よって本稿では，この過剰な変形を抑制するために，特性値を設定する際に画像を分割し，その上で特性値を算出する方法を提案する．提案手法は，以下に示す **Step 1** から **Step 5** の手順で特性値を算出する．

Step 1 平滑化ヒストグラム

y 軸へ射影した黒画素のヒストグラム $f_0(y)$ を作成する．これをガウシアンフィルタによって平滑化し，連続した関数 $f(y)$ を求める．

Step 2 微分ヒストグラム

$f(y)$ を微分し，1 次微分ヒストグラム $f'(y)$ を作成する．このとき微分フィルタとして，以下の局所平均的微分フィルタを用いた．

$$f'(y) = \sum_{y}\sum_{k=1}^{s} \frac{f(y+k) - f(y-k)}{s}$$

ただし，s は y の近傍を表す定数であり，本稿では $s = 5$ とした．同様に $f'(y)$ を微分し，2 次微分ヒストグラム $f''(y)$ を作成する．

Step 3 平滑化ヒストグラムの山部を検出

連続関数 $f(y)$ が $y = a$ で極値を持つならば $f'(a) = 0$ であり，また $f''(a) \leq 0$ ならば，$f(y)$ は $y = a$ 付近で上に凸である．そこで

$$f'(a) = 0 \text{ かつ } f''(a) \leq 0$$

となる点 $y = a$ をヒストグラムの山部であるとみなす．この山部は複数個存在するので，I 個存在するときの山部を $y = a_i \quad (1 \leq i \leq I)$ と表す．

図 4 に平滑化ヒストグラムの作成例と，山部の検出例を示す．

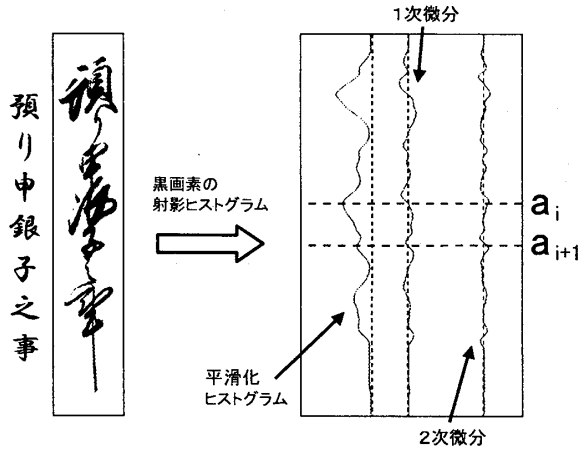

図 4: 平滑化ヒストグラム

Step 4 黒画素のヒストグラムの谷部を検出

平滑化ヒストグラムの I 個の山部 $(y = a_i)$ から，間接的に黒画素のヒストグラムの $I-1$ 個の谷部 $(y = b_i)$ を求める．

$$b_i = \arg_k \left\{ \min_{a_i \le k \le a_{i+1}} f_0(k) \right\} \quad (1 \le i \le I-1)$$

便宜上，y 軸上での画素の上端を b_0，下端を b_I とおく．

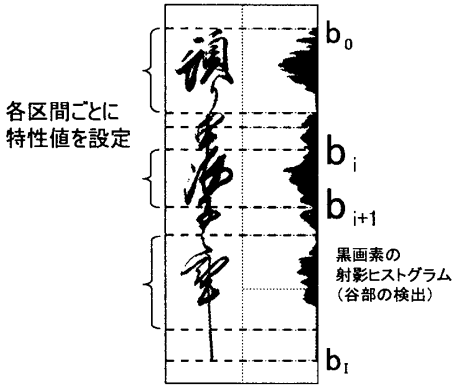

図 5: 特性値を求めるための画像の分割

Step 5 図 5 に示すように，I 個の区間に画像を分割し，各区間ごとに特性値を設定する．$[i-1, i]$ 区間での特性値を $S_x^i(x,y), S_y^i(x,y)$ とすると，以下のように，画像全体での特性値が求まる．

$$S_x(x,y) = \sum_{i=1}^{I} S_x^i(x,y)$$

$$S_y(x,y) = \sum_{i=1}^{I} S_y^i(x,y)$$

そして，3.1 と同様に非線形正規化を行うことが出来る．図 6 に提案手法による非線形正規化画像の一例を示す．左側が従来手法 S2,S3,S3G で，右側がそれぞれ提案手法 S2b,S3b,S3Gb である．特に，従来手法の S2「預」では字が歪んでいるのに対して，提案手法の S2b「預」では字形の過剰な変形が抑制されている．

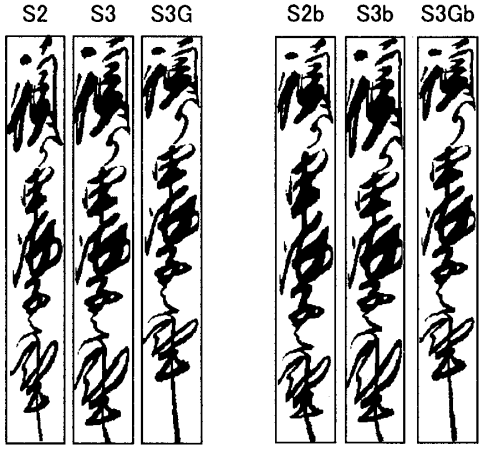

図 6: 非線形正規化画像の比較

4. 認識実験

4.1 データベース

本稿では，古文書標題のデータベースとして，HCR(Historical Character Recognition) プロジェクト [15] において作成された「HCD2」を用いる．翻刻文字列を分析し，1 クラス当たり 10 サンプル以上存在するデータを考慮した結果，本稿では表 1 に示す 5 クラス計 112 サンプルを認識実験に用いる．

表 1: 古文書標題 5 クラス

クラス	翻刻文字列	頻度
A	「借家請状之事」	34
B	「親類請負一札之事」	32
C	「宗旨手形之事」	19
D	「預り申銀子之事」	17
E	「預り申金子之事」	10

4.2 実験方法

2. で示した文字列認識システムにより実験を行った．正準判別分析における次元削減で，選択する次元数を変化させ，Leave-One-Out 法により認識率を求めた．Leave-One-Out 法とは，全 112 サンプルを学習 111 サンプル，未知 1 サンプルに分け，合計 112 回の実験を行い，評価する方法である．

4.3 実験結果

各種の非線形正規化法についての結果を図 7 に示す．図 7 の横軸は，2. の正準判別分析において，d_0 から次元削減した後の次元数 d である．また，各手法において，最高認識率を得たときの結果を表 2 にまとめる．

図 7 において，従来手法では線形正規化 S0 と非線形正規化 S3G が，各次元において比較的高い認識率を

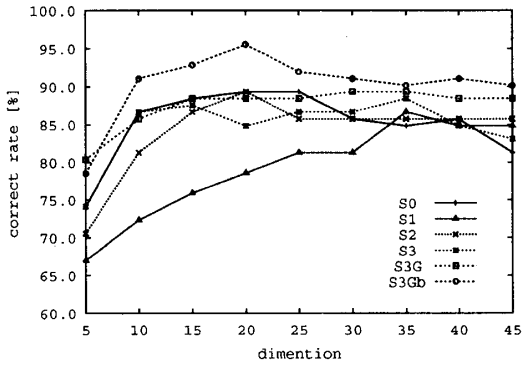

図 7: 各種の非線形正規化に関する実験

表 2: 最高認識率とそのときの次元数

手法	最高認識率	次元数
S0	89.29%	20, 25
S1	86.61%	35
S2	89.29%	20
S3	88.39%	35
S3G	89.29%	30, 35
S3Gb	95.54%	20

示している．S3G は，ぼかし線密度を用いることにより字形の大きな変化を抑えている非線形正規化であり，ぼかしの効果が原因であると考えられる．従来手法での最高認識率は 89.29%(100/112) であったが，提案手法での最高認識率は 95.54%(107/112) であり，差引き 7 サンプル分の改善があった．さらに S0,S3G と S3Gb について，20 次元における各クラスの誤読数を表 3 に示す．すべてのクラスで改善があり，提案手法の有効性を確認した．

表 3: 各クラスの誤読数

手法	A	B	C	D	E	total
従来手法 (S0)	3	2	2	3	2	12
従来手法 (S3G)	4	3	2	3	1	13
提案手法 (S3Gb)	1	2	0	1	1	5

5. まとめ

本稿では，古文書文字列を認識するために，従来の手書き文字列認識技術の応用の一つとして，文字の過剰な変形を抑制した，文字列の非線形正規化法を提案した．従来の非線形正規化法における字形の変化は，特性値の設定に問題があると考えた．そこで提案手法では，画像分割により区間を設定し，その区間ごとで特性値を設定することにより解決を図った．

古文書標題データベース「HCD2」の 5 クラス 112 サンプルを用いた実験により，Leave-One-Out 法での認識率を測定した結果，従来手法の最高認識率が 89.29% であったのに対して，提案手法の最高認識率が 95.54% となり，約 6 ポイントの改善から提案手法の有効性を確認した．

今後は，文字を切出して個別文字認識を行う手法との比較検討が課題である．

謝辞

本研究を行うにあたり，「伏見屋善兵衛文書」標題部分の研究用データベース「HCD2」を提供して頂いた HCR プロジェクトの皆様に感謝致します．

参考文献

[1] 山田奨治，柴山守，“古文書を対象にした文字認識の研究，” 情報処理，no.9，pp.950–955，Sep. 2002.

[2] R.G. Casey, and E. Lecolinet, “A survey of methods and strategies in character segmentation,” IEEE Trans. on Pattern Analysis & Machine Intelligence, no.7, pp.690–706, July 1996.

[3] 堀田悦伸，直井聡，諏訪美佐子，平井淳一，“セグメンテーションの負荷を軽減した手書き住所認識，” 信学技報，PRMU98-161，pp.87–93，1998.

[4] 志久修，中村彰，黒田英夫，宮原末治，“単語全体の形状に注目した手書き日本語単語の認識，” 情処学論，vol.41，no.4，pp.1086–1095，March 2000.

[5] 志久修，中村彰，高比良秀彰，黒田英夫，“パターン整合法による手書き文字列の分類実験，” 信学論 (D–II)，vol.J80–D–II，no.5，pp.1326–1328，May 1997.

[6] 柴山守，“証文類古文書標題の文字認識辞書構築とその利用について — 正規化の問題点と文字認識プロセスの検討 —，” 京大大計センター第 67 回研究セミナー報告，pp.70–79，2001.

[7] 加藤寧，安部正人，根元義章，“改良型マハラノビス距離を用いた高精度な手書き文字認識，” 信学論 (D–II)，vol.J72-D-II，no.1，pp.45–52，Jan. 1996.

[8] 孫寧，安部正人，根元義章，“改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム，” 信学論 (D–II)，vol.J78-D-II，no.6，pp.922–930，July 1995.

[9] C.R. Rao, “Tests of significance in multivariate analysis,” Biometrika, no.1/2, pp.58–79, May 1948.

[10] 柳井晴夫，高木廣文，多変量解析ハンドブック，現代数学社，1986.

[11] 山下義征，樋口浩一，山田陽一，羽下雄之輔，“構造化線素整合法による手書き漢字の大分類，” 信学技報，PRL82–12，pp.25–30，1982.

[12] J. Tsukumo, and H. Tanaka, “Classification of handprinted chinese characters using non-linear normalization and correlation methods,” Proc. 9-th ICPR, pp.168–171, 1988.

[13] H.Yamada, K.Yamamoto, and T.Saito, “A nonlinear normalization method for handprinted kanji character recognition – line density equalization –,” Pattern Recognition, vol.23, no.9, pp.1023–1029, Sep. 1990.

[14] 堀内隆彦，春木亮二，山田博三，山本和彦，“線密度を用いた非線形正規化法の 2 次元的拡張，” 信学論 (D–II)，vol.J80–D–II，no.6，pp.1600–1607，June 1997.

[15] HCR プロジェクトのウェブページ
http://www.nichibun.ac.jp/~shoji/hcr/index.html.