情報処理学会論文誌 Vol. 50 No. 4 1444–1455 (Apr. 2009)

古代木簡解読支援のための文字パターン検索

耒代誠仁[†1] 齋藤恵[†2] 戸根康隆[†1]
石川正敏[†1] 中川正樹[†1]
馬場基[†3] 渡辺晃宏[†3]

手書き文字パターン検索における類似度評価は，手書き文字認識手法を用いて実現できる．しかし，木簡の損傷によって欠損した文字パターンをキーとする場合，非線形正規化におけるパターンの過剰な変形が問題となる．本論文では文字パターンの欠損が疑われる部分をグレーゾーンとして専門家に指摘してもらうことで，非線形正規化処理における問題を緩和し，さらに絞り込み検索にも対応する手書き文字パターン検索手法を提案する．評価実験では，グレーゾーンを用いることによる検索精度の改善が見られた．また，提案手法は木簡解読の専門家からも高い評価を得た．

Character Pattern Retrieval to Support Decoding Historical Mokkans

Akihito Kitadai,[†1] Kei Saito,[†2] Yasutaka Tone,[†1]
Masatoshi Ishikawa,[†1] Masaki Nakagawa,[†1]
Hajime Baba[†3] and Akihiro Watanabe[†3]

Similarity evaluation for character pattern retrieval is feasible by using handwritten character recognition methods. However, too much deformation by non-linear normalization becomes a problem when the key of the retrieval is a character pattern on a mokkan and the pattern has missing parts caused by the damage of the mokkan. This paper presents a character pattern retrieval method that eases the problem and provides a search refinement method for archaeologists by employing gray zones that are suspicious zones of the missing parts indicated by archaeologists. Evaluation experiments showed the improvement of the retrieval accuracy by the employment of the gray zone. Also, archaeologists decoding mokkans gave high marks to our method.

1. はじめに

考古学・歴史学などの分野において，古文書の解読によって得られる情報は重要である．特に，日本各地の遺跡から出土する古代木簡の解読結果には大きな注目が集まっている．

木簡とは，木片に文字が記された文書の総称である．国内における木簡の歴史は古く，奈良平城宮跡，福岡太宰府跡などの古代遺跡からはこれまでに 32 万点を超える木簡が出土しており，その歴史的価値が注目されている．しかし，1 千年以上もの間地中に埋没していた古代木簡の多くには汚損，変色，および文字を表す墨の欠落など，解読の障害となる損傷が見られる．このため，古代木簡の解読は考古学の専門家にとっても困難をともなう．現在，記述の一部でも解読結果が示された木簡は 10 万点に満たず，完全な解読が行われたものに至ってはわずかである．

このような中で，コンピュータによる木簡解読支援に注目が集まっている．市販の画像処理システムは解読現場において広く活用されている．古文書解読支援を目指した画像処理手法[1)–3)]，文脈処理手法[4)] の研究も行われている．我々は木簡解読に有効な画像処理と文脈処理を提案し，その効果を示した[5),6)]．

しかし，墨の欠落によって欠損をともなった文字パターン自体の解読については，専門家の手作業に依存している．欠損した文字パターン（以下，欠損文字パターン）を解読する場合，専門家は他の古代木簡，その他史料から類似した部位を持つ文字パターンを探し，豊富な知識と経験を用いて解読を試みる．しかし，史料の量は膨大であり，文字パターンを検索する専門家への負担は大きい．

我々は，古代木簡のデータベースの構築と公開を通して専門家への負担の軽減を目指してきた[7)]．このデータベースは出土場所，釈文の一部，木片の形状などをキーとして古代木簡を検索する機能を持ち，木簡に限らず古文書解読に広く役立つと期待されている．しかし，木簡上の不完全な文字パターンをキーとした検索については実現されておらず，専門家への有効な支援の提供を実現するうえでの課題となっていた．

手書き文字認識手法は，文字パターンをキーとした検索を実現する有効な手段である．し

†1 東京農工大学
Graduate School of Engineering, Tokyo University of Agriculture and Technology
†2 株式会社 ACCESS
ACCESS Inc.
†3 奈良文化財研究所
Nara National Research Institute for Cultural Properties

かし，既存手法を欠損文字パターンに適用した場合，非線形正規化処理による過剰な変形が問題となる．専門家が文字パターンの欠損を補完することも難しい．ただし，豊富な知識と経験を有する専門家にとって，墨の欠落が疑われる部分を指定することは十分現実的である．

そこで，本論文では文字パターンの欠損が疑われる部分をグレーゾーンとして専門家に指摘してもらうことで，非線形正規化処理における問題を緩和し，さらに欠損した情報を補うための絞り込み検索にも対応する手書き文字パターン検索手法を提案する．

2. 古代木簡解読

2.1 古代木簡

古代木簡は漢字によって記述されており，その内容は実務的である．特に荷札として用いられたものが多く発見されており，納税・民間物流などに広く利用されたと考えられている．また，人足・物資の要求など役所間でやりとりされた公文書としての書式も確認されている．**図 1** は遺跡から出土した古代木簡である．

古代木簡の解読作業を通して，当時の日本における政治・物流・地域産業などが明らかになっている．長屋王邸の所在は，木簡によって明らかになった史実の一例である．貴重な古文書として，一部の古代木簡は国の重要文化財に指定されている．以下，本論文では古代木簡を単に木簡と記す．

しかし，現在までに何らかの解読結果が公開された木簡は 10 万点に満たない．その理由として木簡の汚損，変色，および墨の欠落など木簡に見られる損傷があげられる．専門家は豊富な知識と経験，および史料などを駆使して損傷した木簡の解読を試みているが，欠損が著しい文字パターンの解読は容易ではない．

2.2 コンピュータによる解読支援の現状

木簡解読作業におけるコンピュータの活用例としては，画像処理を用いた墨の抽出支援があげられる[5]．木簡を撮影したデジタル画像を画像処理システムで読み込み，カラーチャネルの選択・クラスタリングアルゴリズムの適用などを行うことで，薄くなった墨を強調し，また，木片表面の黒ずみや木目と墨をある程度区別することができる（**図 2**）．

文脈処理を用いた推読支援についても，木簡解読には有効である[6]．しかし，欠損文字パターン自体の解読に対する有効な支援は実現されていない．解読に悪影響を与える欠損は木簡上の多くの文字パターンに見られ，木簡の解読作業を妨げる要因となっている．

近年では，木簡解読におけるデジタルアーカイブの活用に注目が集まっている．欠損をと

図 1　古代木簡
Fig. 1　Historical Mokkans.

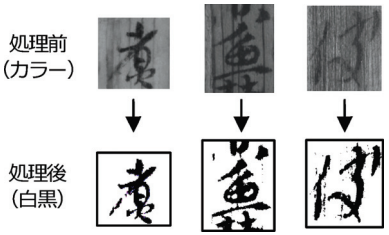

図 2　画像処理の例
Fig. 2　Examples of image processing.

もなう文字パターンの解読を行う場合，類似した形状を持つ文字パターンを他の木簡から探し，その解読結果を参照する方法は有効である．しかし，多数の木簡，およびそれらを掲載した大量の書物から該当する文字パターンを探し出す際には，解読作業，およびそれに関する思考の長時間にわたる中断を余儀なくされる．適切な検索手段を有するデジタルアーカイブは，この問題に対する有力な解決方法である．我々は，平城宮跡をはじめ各地の遺跡から出土した古代木簡をデータベース化するとともに，出土場所，釈文の一部，木片の形状などをキーとする検索機能を実現・公開している[7]．しかし，欠損をともなう文字パターンをキーとする検索（以下，欠損文字パターン検索）は実現できていない．

3. 文字認識手法の課題

3.1 既存のオフライン手書き文字認識

文書に記された手書き文字パターンの検索は，テンプレートマッチングによるオフライン手書き文字認識技術の利用によって実現できる．

オフライン手書き文字認識では，最初に文字を含む画像（入力パターン）に画像処理を施し，2 値画像を得る．2 値画像は，字形を表す黒画素，およびそれ以外を表す白画素からなる．この 2 値画像に対して非線形正規化処理を行い，さらに特徴抽出処理，判別処理（尤度計算）を経て認識結果となる字種を得る．

非線形正規化処理は，黒画素が表す情報を 2 値画像内で均一化することで筆記者，筆記環境などに依存した文字パターンの変動を吸収し，安定した認識精度を実現する．効果的な非線形正規化手法としては，線密度を用いたものがあげられる[8)–10)]．

Liu らが 2 値画像中の画素 (X,Y) に定義した線密度について述べる[10)]．ここでは，2 値画像の大きさを $(x,y)=(Width,Height)$ で表す．また，2 値画像中の画素 (X,Y) は横のライン：$y=Y\{x\,|\,0\leqq x\leqq Width\}$ と縦のライン：$x=X\{y\,|\,0\leqq y\leqq Heigh\}$ にそれぞれ含まれる．このとき，(X,Y) の横のラインに対する線密度 $Dh(X,Y)$ は式 (1)〜(4) で定義される．

(a) 画素 (X,Y) が白画素だけのラインにある場合

$$Dh(X,Y)=0.5/Width \tag{1}$$

(b) 画素 (X,Y) が黒画素の場合

$$Dh(X,Y)=8.0/Width \tag{2}$$

(c) 画素 (X,Y) が黒画素に挟まれた白画素の場合

$$Dh(X,Y)=1.0/Clearance \tag{3}$$

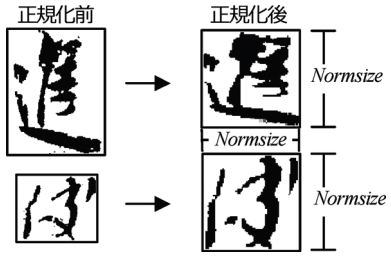

図 3　非線形正規化の例
Fig. 3　Examples of non-liner normalization.

(d) 画素 (X,Y) が黒画素とラインの端点に挟まれた白画素の場合

$$Dh(X,Y)=1.0/(Width+Offset) \tag{4}$$

ただし *Clearance* は (X,Y) を挟む黒画素の間隔，*Offset* は (X,Y) を挟む端点と黒画素の間隔である．一方，縦のラインに対する線密度 $Dv(X,Y)$ の定義は式 (5)〜(8) のとおりである．

(a) 画素 (X,Y) が白画素だけのラインにある場合

$$Dv(X,Y)=0.5/Height \tag{5}$$

(b) 画素 (X,Y) が黒画素の場合

$$Dv(X,Y)=8.0/Height \tag{6}$$

(c) 画素 (X,Y) が黒画素に挟まれた白画素の場合

$$Dv(X,Y)=1.0/Clearance \tag{7}$$

(d) 画素 (X,Y) が黒画素とラインの端点に挟まれた白画素の場合

$$Dv(X,Y)=1.0/(Height+Offset) \tag{8}$$

線密度を用いた非線形正規化では，最初にすべての画素について $Dv(X,Y)$，$Dh(X,Y)$ を求める．次に，式 (9) に示す線密度累積関数を求める．

$$\begin{aligned} Hh(x)&=\textstyle\sum_{y'=0}^{Height} Dv(x,y')\{x\,|\,0\leq x\leq Width\}\\ Hv(y)&=\textstyle\sum_{x'=0}^{Width} Dh(x',y)\{y\,|\,0\leq y\leq Height\}\end{aligned} \tag{9}$$

この累積線密度関数がそれぞれの区間内で単調増加直線となるように 2 値画像に非線形な伸縮を加えることで，2 値画像中の線密度の分布を均一化する．さらに，画像の大きさが縦横ともに *Normsize* になるよう線形変換する．線密度を用いた非線形正規化の処理例を**図 3** に示す．

非線形正規化は，墨の疎な部分と密な部分をならすとともに文字の大きさを一定にするこ

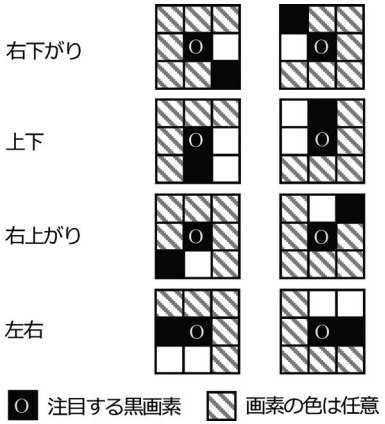

図 4　特徴抽出用の型
Fig. 4　Templates for feature extraction.

とで，字種と関係ない字体の変動を吸収する．

特徴抽出では，非線形正規化後の画像から字体の特徴を抽出する．手書き文字認識に有効な特徴抽出手法として，字体の輪郭線に注目したものが提案されている[11]．注目する黒画素とその 8 近傍画素に**図 4** の型をあてはめ，適合の有無を調べることで輪郭線の 4 方向の成分を抽出することができる．輪郭線に注目した手法は手書き文字認識において広く利用されており，高い精度を実現している．

すべての画素について特徴の抽出が完了すると，非線形正規化後の画像を格子状の区画に分割し，それぞれの区画の中心を頂点とするガウスフィルタを設定する．このガウスフィルタを画素ごとの特徴に乗じて集計したものが，各区画の特徴となる．ガウスフィルタの裾は区画の境界を越えるように設定し，当該区画の外にある画素の特徴についても集計の対象とすることで，非線形正規化だけでは取り除けない字体の変動による特徴の位置ずれをぼかすことができる．**図 5** に区画数が 4×4 の場合の例を示す．ただし，第 i 行第 j 列の区画に対して集計した特徴を F_{ij} とする．区画ごとの特徴を要素とする特徴ベクトルによって，入力パターンはパターン空間上にマッピングされる．区画数が $m \times n$ のとき，特徴ベクトルも $m \times n$ の要素を持つ．本論文ではこれを $m \times n$ 次元の特徴ベクトルと呼ぶ．

判別処理では，前述の特徴ベクトル，およびあらかじめ字種ごとに用意した特徴ベクトル（テンプレート）を用いた尤度の計算（テンプレートマッチング）を行う．尤度の計算では

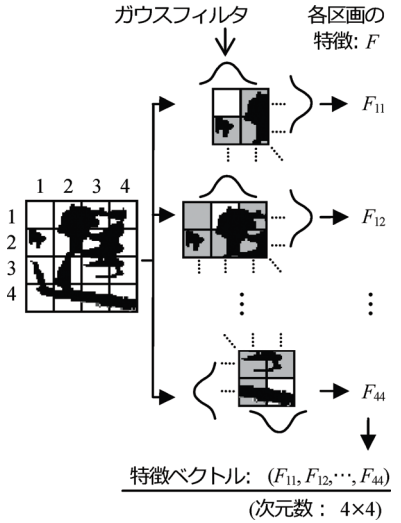

図 5　各区画の特徴からなる特徴ベクトル
Fig. 5　Feature vector consists of regional features.

特徴ベクトル間の距離を用いる．入力パターンとの尤度が高いテンプレートが属する字種を求めることで，認識結果が得られる．

上記の方法を利用した手書き文字パターン検索では，キーとなる文字パターン，および検索対象となる文字パターンからそれぞれ特徴ベクトルを作成し，これらを比較することで，キーに近い特徴を持つ文字パターンを得ることになる．

3.2　文字パターンの欠損による問題

文字パターンの欠損によって非線形正規化は直接的な影響を受ける．

非線形正規化処理は 2 値画像内の黒画素が表す情報に注目した処理である．これは「白画素は字形を表さない無効な領域を表す」との前提に基づく．文字パターンが完全な形を有する場合，この前提は有効である．

しかし，欠損をともなう文字パターンでは，白画素となる部分にも字形が存在した可能性がある．この点を無視して非線形正規化処理を行うと，黒画素が表す部分は欠損を埋めるように拡大される（**図 6**）．これは過剰な変形であり，続く特徴抽出において特徴が抽出されるべき区画にずれを生じさせる．この特徴の位置ずれは欠損が著しいほど大きくなり，平滑

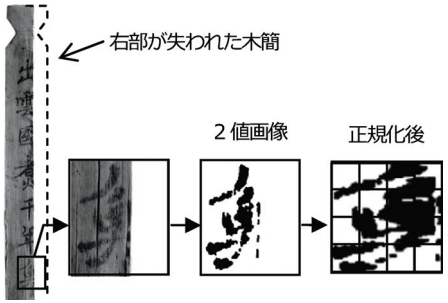

図 6　欠損文字パターンの非線形正規化
Fig. 6　Non-linear normalization for damaged character pattern.

化による吸収も困難となる．その結果，特徴ベクトルを構成する各要素は文字画像中の位置に関する正当性を維持できなくなり，続く特徴ベクトル間の距離計算は信頼性を失う．

4. 欠損文字パターン検索の実現

4.1 多値画像への拡張と絞り込み検索

文字パターンに欠損が生じると，2 値画像上で黒画素となる部分が白画素に反転する．解読支援が必要となるような場合，このような画素の色反転を見抜くことは専門家にとっても困難である．しかし，木簡の損傷状態を参照することで，色反転が疑われる部分を推定することは可能である．

そこで，欠損文字パターン認識においても色の反転が疑われる画素を取り扱う仕組みを実現し，検索精度の向上に役立てることを考える．このためには，白/黒画素しか扱えない 2 値画像を拡張する必要がある．

色の反転が疑われる画素を取り扱う仕組みを以下に述べる．まず，画素の色反転が疑われる領域をグレーゾーンと定義する．グレーゾーンの指示は専門家がペンデバイスなどを用いて行う．指示されたグレーゾーンは白/黒画素と区別するために灰色で表現する．その結果，白/灰/黒画素を含む多値画像ができる．

非線形正規化を行う際，グレーゾーンは黒画素が示す部分の過剰な拡大を抑制する．さらに，グレーゾーン内の灰色の濃度は，色反転した可能性が特に高いと思われる場合は濃く，逆に色反転した可能性が比較的低い場合は薄く，それぞれ専門家が変更できるものとする（**図 7**）．これによって，専門家の推定を反映した絞り込み検索を実現する．

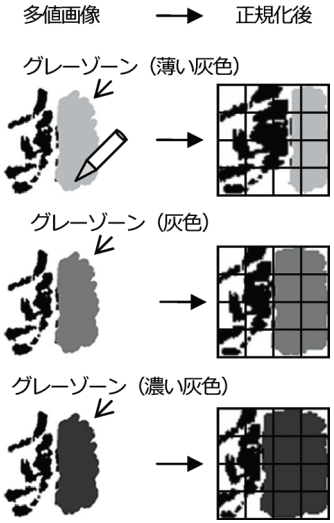

図 7　グレーゾーンと非線形正規化
Fig. 7　Non-linear normalization with gray-zone.

4.2 非線形正規化

非線形正規化では，グレーゾーン内に字形を表す情報が存在したという仮定に基づいて処理を行う．しかし，グレーゾーンそのものは特定の字形を表すものではない．そこで，グレーゾーン内には音声におけるホワイトノイズのように偏りを持たない情報 In_{gray} を考える．In_{gray} はグレーゾーン内に含まれうる情報量の最小値 $|In_{\min}|$ と最大値 $|In_{\max}|$ に対して式 (10) を満たす．

$$|In_{\min}| \leq |In_{gray}| \leq |In_{\max}| \tag{10}$$

この In_{gray} はグレーゾーンの灰色の濃度を反映して決定する．すなわち，白に近い灰色であれば In_{gray} を $In_{\min}$ に近づけ，黒に近い灰色であれば $In_{\max}$ に近づける．これによって，グレーゾーン内で比較的多くの画素の色が反転したと考えられる場合，専門家は濃い灰色を用いることで黒画素が表す部分の拡大を強く抑制することができる．一方で，色の反転した画素が比較的少ないと考えられる場合は薄い灰色を用いることで拡大をある程度許容する．

このような非線形正規化を実現する方法として，本研究では線密度を用いた手法を提案する．この方法では，グレーゾーンをすべて黒画素，または白画素とした 2 つの 2 値画像を

生成し，それぞれに対して線密度を求め，それらをグレーゾーンの色（灰色の濃度）に応じて加重平均したものを多値画像の線密度とすることで非線形正規化を実現する．

4.3　特徴抽出と判別処理

本論文では特徴抽出と判別処理について，グレーゾーン内で字形を表す特徴を推定する方法（以下，特徴推定法），およびグレーゾーンに応じてテンプレートを修正する方法（以下，テンプレート修正法）の 2 つを提案する．以下に詳細を記す．

① 特徴推定法（Feature Estimation）

特徴推定法はグレーゾーンに含まれる字形を表す特徴 Ie_{gray} を推定・抽出する．Ie_{gray} はグレーゾーン内に含まれうる特徴量の最大値 $|Ie_{\max}|$，および最小値 $|Ie_{\min}|$ に対して式 (11) を満たす．

$$|Ie_{\min}| \leq |Ie_{gray}| \leq |Ie_{\max}| \tag{11}$$

このような特徴抽出を実現するため，特徴推定法では欠損した部分と残存する部分が一体の文字パターンであったことに注目して，グレーゾーン以外の領域から輪郭線に注目した特徴の平均値を求め，グレーゾーン内の画素の方向特徴とする．なお，グレーゾーン内の灰色の濃度は先に行う非線形正規化によって正規化されると見なせるため，特徴抽出には反映しない．

判別処理では，上記の方法で入力パターンから抽出された特徴ベクトルをテンプレートと比較して尤度を求める．

② テンプレート修正法（Template Adjustment）

入力パターンから黒画素の特徴だけを抽出するとともに，テンプレートからはグレーゾーンに対応する部分の特徴を取り除くことで，残存する字形に注目した特徴抽出，および判別処理が可能になる．

このように，認識対象に応じてテンプレートに修正を加える方法を文字認識に採用したものとして孫らの手法があげられる[12]．この手法では，部分的な字体の類似性および欠損を統計的に解析し，識別結果を不安定にする特徴をテンプレートおよび入力パターンから取り除くことで認識精度の向上を実現している．テンプレートの修正にともなう計算量の増加に対しては，適用対象を誤認識が発生しやすい字種の判別に限定し，統計的解析に基づいてあらかじめ修正したテンプレートを用意することで対応できる．

しかし，解読支援が必要となるほどの欠損が発生した場合，誤認識を基準に適用対象を制限する効果は期待できない．グレーゾーンの位置・形状を統計的に予測して，修正済のテンプレートを事前に用意することも不可能である．計算量の増加が字体検索処理のレスポンスを低下させ，専門家の思考に停滞を招くことを考慮すると，グレーゾーンに応じた動的なテンプレートの修正を効率的に行う方法が必要である．

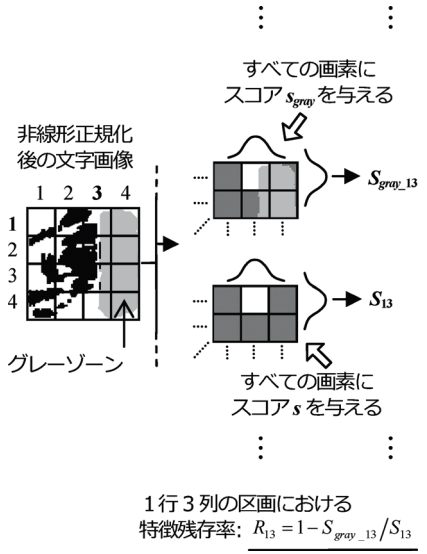

図 8　特徴残存率
Fig. 8　Feature remaining rate.

そこで，本論文では特徴残存率（Feature Remaining Rate）を定義し，これに基づくシンプルなテンプレート修正法を提案する．まず，非線形正規化後の入力パターン（正規化済み入力パターン）に含まれるすべての画素に式 (12) で表されるスコア s_{gray} を与える．ただし，c は 0 より大きい定数である．

$$s_{gray} = \begin{cases} c & \text{(画素は灰色)} \\ 0 & \text{(画素は黒/白)} \end{cases} \tag{12}$$

次に，特徴抽出の際と同じく画像を格子状の区画に分割し，特徴抽出で用いるものと同じガウスフィルタを乗じて区画ごとに s_{gray} を集計する．このとき，第 i 行第 j 列にある区画に対して集計された s_{gray} を S_{gray_ij} とする．続いて，正規化済み入力パターンに含まれるすべての画素に式 (13) で定義されるスコア s を与え，同様に区画への分割とガウスフィルタを用いた集計を行う．

$$s = c \qquad \text{(画素の色は任意)} \tag{13}$$

このとき，第 i 行第 j 列にある区画に対して集計された s を S_{ij} とする（**図 8**）．ここで，第 i 行第 j 列の区画における特徴残存率 R_{ij} を式 (14) のとおり定義する．

$$R_{ij} = 1 - S_{gray\_ij}/S_{ij} \tag{14}$$

テンプレートの修正では，テンプレートとなる特徴ベクトルの各要素 F_{ij} を $R_{ij} \times F_{ij}$ で置き換える．また，正規化済み入力パターンからは黒画素の特徴だけを抽出し，特徴ベクトルを作成する．これらを比較し，尤度を求めることで判別処理を行う．

5. データベースを用いた評価実験

5.1 実 験 方 法

グレーゾーン導入の効果を定量的に検証するため，我々は古代木簡から抽出した文字画像のデータベースを作成し，それを用いた評価実験を行った．

① 評価用データベース

作成したデータベースは 309 字種 2,108 画像からなる．字種は古代木簡に多く見られるものを選択し，字種あたりの文字画像数は 2 以上とした．また，木簡画像からの文字の切り出し，および 2 値化は木簡解読の専門家が行った．なお，実験結果の客観性を確保するため，すべての文字画像は人間が文脈などの補助的な情報なしで解読できるものとした．データベースに含まれる文字画像の一部を**図 9** に示す．

② 欠損をともなうキーの作成

文字画像から疑似的なグレーゾーンまたは欠損を含むキーを作成するために，**図 10** に示す 16 種類のマスクを用意した．なお，図 10 の (5)–(7) は (1)–(3) をそれぞれ 150%に拡大したもの，(9)–(11) は 50%に縮小したものである．

疑似的なグレーゾーンまたは欠損を含むキーの作成方法を**図 11** に示す．ここでは，文字画像と重ね合わせたマスクを灰色とすることで疑似グレーゾーンを含む多値画像を，またマスクを白画素とすることで疑似欠損を含む 2 値画像を，それぞれ得る．

③ 評価尺度

実験では，1 個抜きクロスバリデーション法を用いた[13]．すなわち，1 つの文字画像からキーを生成して残り 2,107 個の文字画像を検索する試行を，データベースが一巡するまで実施した．このとき，各試行においてキーと同じ字種に属する文字画像が尤度上位 10 位以内に含まれる確率を求めた．以下，この確率を検索率（retrieval rate）と呼ぶ．検索率を 10 位まで求める理由は，解読作業に適した小型のコンピュータが持つ低解像度の表示画面において，1 度に表示できる検索結果候補の数が 10 個程度になるためである．なお，提案

図 9　データベースに含まれる文字画像の例
Fig. 9　Examples of character image in database.

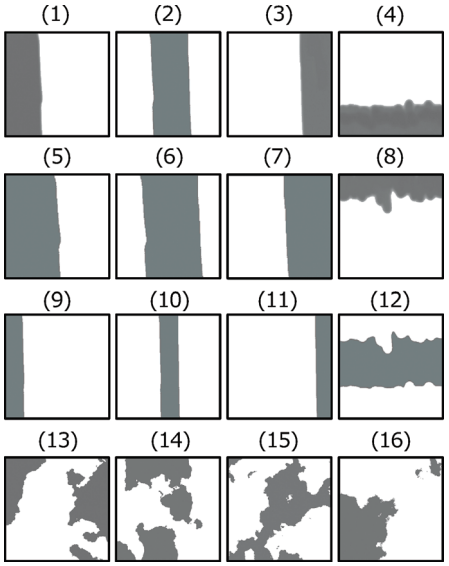

図 10　疑似的なグレーゾーン/欠損用のマスク
Fig. 10　Masks for quasi gray-zone/lack.

手法の用途が文字パターン検索であることを考慮し，同一字種の複数の画像が同時に上位 10 位に含まれる場合でも 11 位以下の画像の繰上げは行わない．

尤度計算における距離尺度としては，特徴ベクトルどうしのシティブロック距離（Cityblock distance）とユークリッド距離（Euclidian distance）を用いた．

④ 疑似グレーゾーンの色濃度

疑似グレーゾーンに使用する灰色の濃度は 0%（白）から 100%（黒）までの間で任意に設定できる．しかし，運用時の利便性を考えると有限個の既定値を用意してユーザに選択させ

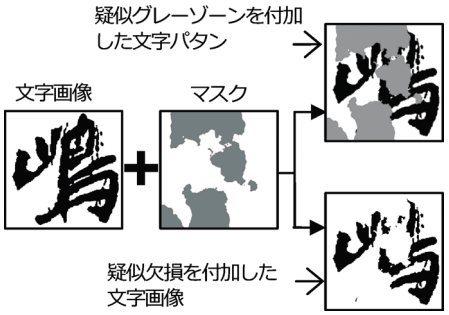

図 11　マスクを用いたキーの作成
Fig. 11　Generating key with mask.

表 1　マスクを適用しない場合の検索率
Table 1　Retrieval rates for non-masked images.

特徴ベクトル	Cityblock	Euclidian
8×8 次元	76.0%(1,601)	75.0%(1,581)
10×10 次元	68.6%(1,445)	71.0%(1,497)

るのが現実的である．そこで，濃度を 50%（白と黒の単純平均）に固定する場合，および 0/25/50/75/100%の中で最も結果が良かったものを採用する場合について実験を行う．前者はグレーゾーン内で色が反転した画素の数をユーザが推定しない場合の既定値であり，最悪のケースにおける結果を示す．一方，後者はユーザが最適値となる濃度を選択した場合の理想的な値である．なお，濃度を 0%とする場合は疑似欠損を付加した文字画像がキーとなる．

5.2 実験結果

はじめに，マスクを適用しない文字画像を用いて，欠損による影響を受けない場合の検索率を示す．**表 1** は非線形正規化後の画像サイズを 64×64 pixel，区画数を 8×8 とした場合（特徴ベクトルの次元数は 8×8），およびこれらの値を 100×100 pixel，10×10 とした場合（特徴ベクトルの次元数は 10×10）の検索率である．それぞれの試行回数は 2,108 回である．なお，表中ではシティブロック距離を Cityblock，ユークリッド距離を Euclidian と略記する，また，表の括弧内はキーと同じ字種に属する文字画像が尤度上位 10 位以内に含まれた回数を表す．

表 2 は，疑似欠損を付加した文字画像をキーとした場合の検索率である．図 10 のマスクを利用して文字画像ごとに 16 のキーを生成することで，文字画像あたりの試行回数は 16 回となり，それぞれの総試行回数は 33,728 回となる．

表 2　疑似欠損を付加した場合の検索率
Table 2　Retrieval rates for quasi-lacked images.

特徴ベクトル	Cityblock	Euclidian
8×8 次元	42.7%(14,414)	43.0%(14,488)
10×10 次元	35.5%(11,963)	38.9%(13,134)

表 3　特徴推定法による検索率
Table 3　Retrieval rates by feature estimation.

特徴ベクトル	Cityblock	Euclidian
色濃度: 50%		
8×8 次元	57.8%(19,478)	53.4%(18,011)
10×10 次元	50.8%(17,142)	48.2%(16,260)
色濃度: 最適値		
8×8 次元	69.2%(23,338)	65.4%(22,063)
10×10 次元	63.6%(21,456)	61.4%(20,717)

表 4　テンプレート修正法による検索率
Table 4　Retrieval rates by template adjustment.

特徴ベクトル	Cityblock	Euclidian
色濃度: 50%		
8×8 次元	63.1%(21,296)	63.3%(21,339)
10×10 次元	55.5%(18.727)	58.5%(19,725)
色濃度: 最適値		
8×8 次元	75.0%(25,305)	75.4%(25,440)
10×10 次元	69.4%(23,402)	72.5%(24,451)

表 3 は特徴推定法を用いた場合について，**表 4** はテンプレート修正法を用いた場合について，それぞれ疑似グレーゾーンを付加した文字画像をキーとした検索率である．

疑似グレーゾーンを付加した文字画像をキーとした場合の検索率は，疑似欠損を付加した場合の検索率を上回った．また，テンプレート修正法は特徴推定法よりも優れた検索率を示した．特に，色濃度を最適値とした場合は距離尺度，特徴ベクトルの次元数にかかわらずマ

表 5　マスクの大きさと検索率
Table 5　Mask size and retrieval rates.

マスク	Cityblock	Euclidian
疑似欠損		
(1)-(3)	39.8%(2,515)	40.2%(2,539)
(5)-(7)	16.9%(1,067)	18.9%(1,197)
(9)-(11)	66.5%(4,208)	64.9%(4,103)
テンプレート修正法（色濃度: 50%）		
(1)-(3)	65.0%(4,111)	65.4%(4,133)
(5)-(7)	51.2%(3,240)	52.3%(3,307)
(9)-(11)	70.1%(4,489)	70.7%(4,469)
テンプレート修正法（色濃度: 最適値）		
(1)-(3)	77.1%(4,877)	77.0%(4,869)
(5)-(7)	69.6%(4,403)	71.3%(4,508)
(9)-(11)	80.1%(5,065)	80.2%(5,071)

図 12　拡大したマスク
Fig. 12　Expanded masks.

スクを適用しない場合と同等の検索率を示した．

マスクの大きさによる検索率への影響を調べるため図 10 のマスク (1)–(3)，(5)–(7)，(9)–(11) を用いた実験を行った．結果を**表 5** に示す．特徴ベクトルの次元数にはこれまでの実験で良好な結果を示した 8×8 を用いた．また，疑似グレーゾーンを付加した多値画像に対してはテンプレート修正法を用いた．それぞれ，総試行回数は $2{,}108 \times 3 = 6{,}324$ 回である．

これらの結果から，マスクが大きくなると疑似欠損を用いた場合の検索率が大きく低下するのに対して，提案手法を用いた場合は検索率の低下が抑えられていることが分かる．

マスクの大きさによる影響をさらに調査するため，マスク (1)–(3) を 200%まで拡大したマスク（**図 12**）を用いた実験を行った．それぞれ，総試行回数は $2{,}108 \times 3 = 6{,}324$ 回である．結果を**表 6** に示す．

このように，マスクが文字画像の多くの部分を覆う場合においても提案手法による効果は

表 6　拡大されたマスクを用いた場合の検索率
Table 6　Retrieval rates with expanded masks.

Cityblock	Euclidian
疑似欠損	
8.2%(516)	9.5%(601)
テンプレート修正法（色濃度: 50%）	
29.8%(1,885)	31.4%(1,985)
テンプレート修正法（色濃度: 最適値）	
54,7%(3,457)	57.5%(3,633)

表 7　3/5 位候補含有率
Table 7　3rd/5th Accumulative Rates.

Cityblock	Euclidian
マスク不使用	
3 位: 66.2%(1,396) 5 位: 71.1%(1,498)	3 位: 65.7%(1,385) 5 位: 70.3%(1,482)
疑似欠損を付加	
3 位: 31.7%(10,705) 5 位: 36.1%(12,183)	3 位: 31.4%(10,573) 5 位: 36.0%(12,130)
特徴推定法，色濃度: 50%	
3 位: 45.4%(15,323) 5 位: 50.6%(17,081)	3 位: 41.3%(13,922) 5 位: 46.2%(15,590)
特徴推定法，色濃度: 最適値	
3 位: 57.7%(19,446) 5 位: 62.7%(21,158)	3 位: 53.7%(18,104) 5 位: 58.5%(19,718)
テンプレート修正法，色濃度: 50%	
3 位: 51.6%(17,391) 5 位: 56.5%(19,062)	3 位: 51.2%(17,268) 5 位: 56.4%(19,022)
テンプレート修正法，色濃度: 最適値	
3 位: 64.9%(21,892) 5 位: 69.4%(23,396)	3 位: 65.1%(21,956) 5 位: 69.6%(23,461)

見られるものの，検索率には大幅な低下が見られる．

最後に，キーと同じ字種に属する文字画像が尤度上位 3，5 位以内に含まれる確率（3/5 位候補含有率）を**表 7** に示す．これは，正解候補が上位に含まれることが利用者のメリットになる点を考慮したものである．特徴ベクトルの次元数は 8×8 とし，マスクは図 10 のすべてを用いることとする．また，これまでの実験と同様に画像の順位の繰上げは行わな

い．なお，総試行回数はマスク不使用の場合が 2,108 回，それ以外が 33,728 回である．

なお，上記の実験はすべて Intel Xeon 3060（2.4 GHz）を CPU とするコンピュータ上で，単一スレッドのソフトウェアを用いて実施した．ユークリッド距離とテンプレート修正法を利用し，特徴ベクトルの次元数を 8×8，疑似グレーゾーンの色濃度を 50%に固定したときの処理時間は試行 1 回につき約 0.13 秒であった．

6. 専門家による評価実験

次に，グレーゾーンを採用した文字パターン検索，およびグレーゾーン内で灰色の濃度値を変更する機能の有効性を実際の環境で示すために，木簡解読の専門家 3 名を被験者とする評価実験を行った．

実験では，提案手法を実装した木簡解読支援システムを被験者に使用してもらい，アンケートに回答してもらう形式を採用した．各種パラメータについては，特徴ベクトルの次元数は 8×8，特徴抽出および判別処理にはテンプレート修正法を利用，灰色の濃度値は 0/25/50/75/100%の 5 段階から選択可能とした．また，検索結果にはユーザの利便性を考慮して字種および文字画像（2 値画像）の両方を表示した．なお，文字パターン検索を含む木簡解読支援システムの使用方法についてはいっさいの制限を行わず，アンケートの実施に先立ち 1 カ月以上の期間を設けて自由に使用してもらった．

アンケートの内容は次のとおりである．

問 1：グレーゾーンを用いた文字パターン検索（文字認識）機能は必要ですか？

問 2：グレーゾーン内の灰色の濃度を変更する機能は必要ですか？

問 3：その他（自由筆記/任意）

問 1，2 への回答は 5 段階（1：不要〜5：必要）とした．その結果，問 1，問 2 ともに全員が 5（必要）と回答した．また，2 名から問 3 に対して以下のコメントを得た．

- グレーゾーンは，欠損したり劣化したりして不完全な場合の多い木簡の文字の解読には欠かせない，画期的な機能だと思います．また，人間と機械の共同作業という点でも意味のある機能だと思います．範囲の自由設定と，5 段階の濃度設定によって，欠損の範囲をどう見るか，またその程度をどう見るか，いろいろ試すことができるのはたいへんありがたいです．
- グレーゾーンによって欠けてしまった部分を示し，認識率を高めることは非常に重要である．また，欠け方はそれぞれの資料に個別的であり，その様子をグレーの度合いで指定できることもまた非常に重要である．

7. 考　　察

データベースを用いた評価実験において，使用したマスクの形状と灰色の濃度値は疑似的なものにすぎず，残存する墨の形状を考慮したものではない．しかし，5.2 節に示した実験結果においては，グレーゾーンを用いることで検索率の低下を抑えることができた．この傾向は，特徴ベクトルの次元数，距離尺度，および特徴推定法とテンプレート修正法の選択にかかわらず表れており，グレーゾーンを用いる効果の普遍性が示されたといえる．また，テンプレート修正法を用いた場合の検索率および 3/5 位候補含有率は，提案手法の有用性の高さを示すものと考える．処理時間の計測には機材の関係で比較的高速なコンピュータを用いたが，処理時間は十分に短く，ラップトップ型のコンピュータを用いた場合でも実用的な解読支援を提供できると考える．

専門家による評価実験の結果は，文字パターン検索および絞り込み検索の機能が専門家に高く評価されたことを示している．古代木簡の解読作業は，失われた墨を復元する作業である．色反転した画素，すなわち失われた字体の特徴を絞り込みの条件とする文字パターン検索は，この作業を直感的に支援する有効な手段といえる．

ただし，グレーゾーンが大きくなると検索率にも低下が見られる．したがって，木簡の欠損が大きな場合には文脈処理の併用などで検索精度の低下を補う工夫が必要である．

8. おわりに

本論文では，グレーゾーンを用いた文字パターン検索手法を提案し，欠損をともなう文字パターンに対する効果を示した．マスクを用いた評価実験では，特徴ベクトルの次元数と距離尺度が異なるすべてのケースにおいて疑似グレーゾーンを用いた結果が疑似欠損を用いた結果を上回った．また，専門家による評価実験においても提案手法は高い評価を得た．

今後の課題としては，グレーゾーンに設定する特徴についての検討があげられる．失われた墨の形状に対する専門家の推定を特徴抽出に反映する手法を実現することで，検索結果の効果的な絞り込みが可能になり，文字パターン検索の有用性が向上すると考えられる．また，破損をともなう古代木簡から定量的評価に適した文字パターンを数多く収集し，文字パターンデータベースを拡張することも今後の課題である．大規模なパターンデータベースによって統計情報を用いた処理が可能となり，検索精度の向上につながると考えられる．

謝辞　本研究は科研費基盤 S-20222002 および若手 B-19720202 の助成を受けたものである．

参 考 文 献

1) Gatos, B., Pratikakis, I. and Perantonis, S.J.: An Adaptive Binarization Technique for Low Quality Historical Documents, *Proc. 6th DAS*, Florence, Italy, pp.102–113 (2004).
2) Yan, C. and Leedham, G.: Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document Images, *Proc. 9th IWFHR*, Tokyo, Japan, pp.239–244 (2004).
3) Shi, Z. and Govindaraju, V.: Historical Document Image Enhancement Using Background Light Intensity Normalization, *Proc. 17th ICPR*, Cambridge, UK, 2aP. Mo-ii (2004).
4) 山田奨治，柴山　守：n-gram による古文書証文類翻刻支援の検討，人文科学とコンピュータシンポジウム論文集，Vol.2000, No.17, pp.185–192 (2000).
5) 齋藤　恵，蜂谷大翼，耒代誠仁，中川正樹，馬場　基，渡辺晃宏：木簡画像から墨の部分を抽出するための画像処理，信学技報，PRMU2004-259, Vol.104, No.742, pp.163–168 (2005).
6) 耒代誠仁，西嶋佳津，齋藤　恵，石川正敏，中川正樹，馬場　基，渡辺晃宏：木簡解読支援のための文脈処理，日本情報考古学会誌，Vol.13, No.1, pp.7–21 (2007).
7) http://jiten.nabunken.go.jp/
8) Tsukumo, J. and Tanaka, H.: Classification of Handprinted Chinese Character Using Non-linear Normalization and Correlation Methods, *Proc. 9th ICPR*, Roma, Italy, pp.168–171 (1988).
9) Yamada, H., Yamamoto, K. and Saito, T.: A Nonlinear Normalization Method for Handprinted Kanji Character Recognition – Line Density Equalization, *Proc. 9th ICPR*, Roma, Italy, pp.172–175 (1988).
10) Liu, C.L., Kim, I.J. and Kim, J.H.: High accuracy handwritten Chinese character recognition by improved feature matching method, *Proc. 4th ICDAR*, Ulm, Germany, pp.1033–1037 (1997).
11) Liu, C.L., Liu, Y.J. and Dai, R.W.: Multiresolution statistical and structural feature extraction for handwritten numeral recognition, *Pre-Proc. 5th IWFHR*, Colchester, England, pp.61–66 (1996).
12) 孫　寧，安部正人，根本義章：部分整合領域の自動学習による手書き文字の詳細識別に関する一手法，信学論（D-II），Vol.J78-D-II, No.3, pp.492–500 (1995).
13) Tukey, J.W.: Bias and confidence in not-quite large samples, *Ann. Math. Statist.*, Vol.29, p.614 (1958) (abstract).

(平成 20 年 5 月 27 日受付)
(平成 21 年 1 月 7 日採録)

耒代　誠仁（正会員）

2004 年東京農工大学大学院工学研究科博士後期課程修了．同年より同大学研究員，助手，助教を経て，現在，同大学大学院工学府特任准教授．手書き文字認識の高精度化，コンピュータと教育，古文書解読支援等の研究・教育に従事．電子情報通信学会，日本情報考古学会，ヒューマンインタフェース学会各会員．工学博士．

齋藤　　恵

2004 年東京農工大学電子情報工学科卒業．2006 年同大学大学院工学府博士前期課程修了．同年株式会社 ACCESS 入社．モバイルコンピューティング，画像処理，手書き文字認識，古文書解読支援の研究に従事．工学修士．

戸根　康隆（学生会員）

2007 年東京農工大学電子情報工学科卒業．現在同大学大学院工学府博士前期課程に在籍．ペンデバイスを用いたヒューマンインタフェース，古文書解読支援の研究に興味を持つ．工学学士．

石川　正敏（正会員）

2000 年奈良先端科学技術大学院大学情報科学研究科博士課程単位取得退学．同年より島根県立大学助手，大阪府立工業高等専門学校講師を経て，現在東京農工大学大学院工学府特任助教．時空間データベース，地理情報システム等の研究に従事．電子通信学会，日本情報考古学会，ACM，IEEE 各会員．工学博士．

中川　正樹（正会員）

1997 年東京大学理学部卒業．1979 年同大学大学院修士課程修了．同在学中，英国 Essex 大学留学（M.Sc. with distinction in Computer Studies）．1979 年東京農工大学工学部助手．現在，同大学大学院工学府教授．手書きパターン認識，手書きユーザインタフェース，教育の情報化等の研究・教育に従事．理学博士．

馬場　　基

1995 年東京大学文学部日本史学専修課程卒業．2000 年同大学大学院人文社会系研究科博士後期課程中退．同年奈良文化財研究所研究員．現在，同研究所都城発掘調査部史料研究室において日本史，特に古代交通，都市，寺院，出土文字史料の研究に従事．史学会，古代交通研究会，交通史研究会，木簡学会各会員．文学修士．

渡辺　晃宏

1982 年東京大学文学部国史学科卒業．1989 年同大学大学院博士課程単位取得退学．同年奈良文化財研究所研究員．現在，同研究所都城発掘調査部史料研究室長．平城宮・京の発掘調査と出土文字資料の研究に従事．木簡学会会員．文学修士．