古代木簡解読支援システムにおける字体検索の高性能化

Sherini Somayeh
東京農工大学 工学府

未代 誠仁
桜美林大学

中川 正樹
東京農工大学 工学府

馬場 基, 渡辺 晃宏
奈良文化財研究所

本稿では古代木簡解読支援のための字体検索技術，およびその実装について述べる．古代日本で多数使用された木簡の多くは汚損，破損によって専門家にとっても解読が困難な状態にある．本研究の目的は，コンピュータによるパターンマッチング技術を活用して，古代木簡上の破損した字体の解読を支援する情報検索を実現することにある．実用に耐える検索精度を実現するために，我々はパターンマッチングの精度を向上させる改良グレーゾーン法およびテンプレート修正法の設計と実装を行った．また，専門家のアイデアを検索結果に反映する対話的な検索インタフェースを，我々が開発してきた木簡解読支援システムの上に構築した．

Performance improvement of character pattern retrieval for support system to read historical mokkans

Somayeh Sherini
Faculty of Engineering
Tokyo University of A&T

Akihito Kitadai
J.F. Oberlin University

Masaki Nakagawa
Faculty of Engineering
Tokyo University of A&T

Hajime Baba, Akihiro Watanabe
Nara National Research Institute
for Cultural Properties

In this paper, we describe a technology of character pattern retrieval and its implementation to support reading historical mokkans. Since many mokkans commonly used in ancient Japan have been stained and degraded, it is hard to read them even for the expert readers. The aim of this research is to constructing the information retrieval technology for damaged character patterns on the mokkans by utilizing pattern recognition methods on computers. To achieve sufficient accuracy of retrieval, we have designed and implemented the integrated gray-zone method and the template modification method. Also, to refine the results of the retrieval by reflecting the idea of the expert readers, we have developed the interactive user interface on our support system to read the mokkans.

1. まえがき

本稿では，古代木簡上の破損した字体の解読を支援する字体検索機能の高性能化と木簡解読支援システム上での実装について述べる．

木簡とは木片に墨で文字が記された文書の総称である．特に，古代日本で作成，利用された木簡を古代木簡と呼ぶ．奈良平城宮跡などから出土した古代木簡の総数は 320,000 点を超え，当時を知る貴重な史料となっている（図 1）．

古代木簡の多くは汚損，破損などによって解読が困難となっている．我々は，古代木簡解読支援システム「Mokkanshop」の実現を通してこの問題の解決を目指してきた．Mokkanshop は古文書の解像能を高める画像処理機能，不完全な釈文の補完を支援する文脈処理機能，および破損字体認識機能とデジタルアーカイブ参照機能を連動した字体検索機能などを提供し，古代木簡解読の専門家を支援する[1].

これらのうち，字体検索機能については古代木簡の解読に有効な類例検索を実現する手段となる．難読木簡に対して古代木簡デジタルアーカイブ内の類例を検索・照合することができれば，汚損・破損によって失われた情報の補完，字体・文法の検証などを効率的に行うことが可能となる．

我々は，グレーゾーン法による字体認識技術を実現し，一般的な手書き文字認識手法では対応できなかった破損字体の検索を可能にした．しかし，字体検索機能を実用的なものにするためには，検索精度の向上，絞り込み検索の改善などが不可欠であった．

本稿では，字体検索の精度を向上する改良グレーゾーン法とテンプレート修正法の設計と

Mokkanshop での字体検索の実装，および絞り込み検索の精度を向上させる筆跡追跡機能について述べる．

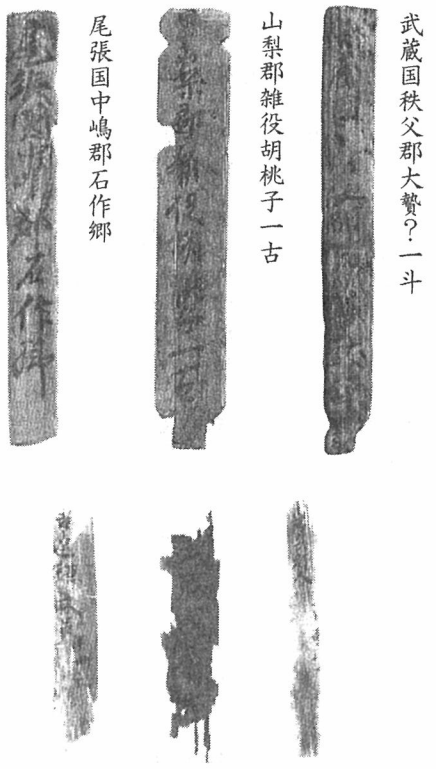

図 1. 古代木簡（奈良文化財研究所）

2．字体検索の精度改善手法について

2.1 グレーゾーン法による非線形正規化

非線形正規化は，字体に含まれる形状情報の分布を均一化することで，個人差・筆記環境に起因する字体の癖を吸収し，パターン認識の精度を向上する字体変形処理である[2]．しかし，破損字体に適用すると，形状情報の欠損により過剰な変形が起こる．

我々は，字体の破損が生じたと考えられる部分（グレーゾーン）をユーザが指示するグレーゾーン法を提案した．グレーゾーンは疑似的な形状情報となり，過剰な変形を抑制する．ただし，字形（黒）と背景（白）の 2 値画像を扱う非線形正規化は，グレーゾーンを直接処理できない．そこで，グレーゾーン法ではグレーゾーンを白画素化/黒画素化した 2 つの画像に対して特徴の分布 h_b と h_w を求め，それらの累積関数 a_b と a_w の加重平均 a_{wave} を線形化する変形を求めることで破損字体の非線形正規化を実現する（図 2）．

グレーゾーン法では，加重平均の係数 w を変化させることで，グレーゾーン内で失われた形状情報の推定値を変化させ，字体の変形を調整することができる．しかし，グレーゾーンを暗い灰色と明るい灰色で塗り分けた場合，すなわちグレーゾーンの一部では多くの形状情報が失われ，他の部分では少ない形状情報が失われたとユーザが仮定した場合に対応できない．

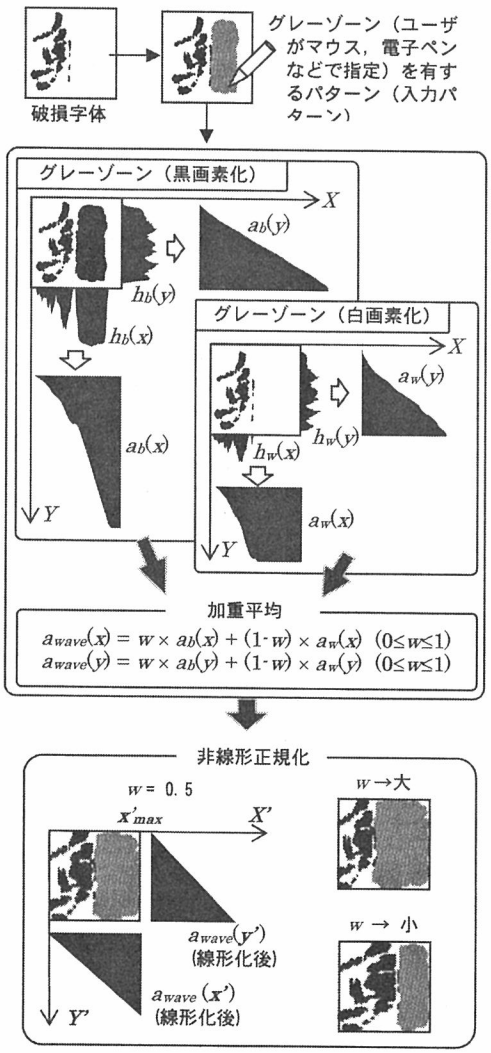

図 2. グレーゾーン法による非線形正規化

2.2 改良グレーゾーン法

そこで，我々は塗り分けに対応した改良グレーゾーン法を実現した．改良グレーゾーン法で

は，先の h_b と h_w に対して差分 $h_{dif}=h_b-h_w$ を求める．また，明度が異なる灰色で塗り分けられたグレーゾーンに対して，グレーゾーンに含まれる画素の平均明度で塗り直された平均グレーゾーンを定義し，それぞれに対する形状情報の分布 h_g，h_{gave} を求める（図 3）．

さらに，下の数式を用いて h_b を再定義する．

$$h_b(x) = h(x) + \{ h_{dif}(x) \times h_g(x) / h_{gave}(x) \}$$
$$h_b(y) = h(y) + \{ h_{dif}(y) \times h_g(y) / h_{gave}(y) \}$$

その後，グレーゾーン法と同様に h_b（再定義）および h_w に対して累積関数 a_b と a_w を求め，加重平均 a_{wave} を線形化する変形を行うことで，改良グレーゾーン法による非線形正規化が完了する（図 3）．係数 w の影響についてもグレーゾーン法に準ずる．

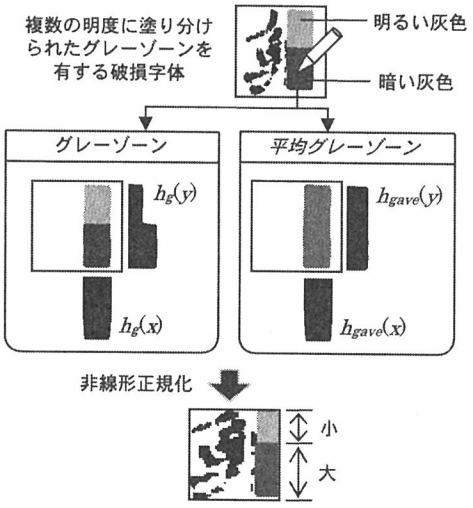

図 3. 改良グレーゾーン法による
非線形正規化

2.3 テンプレート修正法

非線形正規化と並んで問題となるのが，字体の破損した部分に対する類似度評価である．

字体の破損は，字体から抽出できる特徴（文字パターンの形状に関する情報）の欠損に直結する．パターンマッチングに基づく識別器は基本的に完全なパターンの類似度を算出するために設計・実装されており，特徴が欠損した部分については字体が存在しない背景として扱う．これによって類似度計算におけるパターン間距離は不適当に拡大し，字体検索の精度を悪化させる要因となる．

この問題を解決するアプローチとしては，①欠損した情報を何らかの方法で補完する，②識別器が有する知識情報（テンプレート）に意図的な欠損を生じさせて欠損部分に対する不適当な類似度評価をキャンセルする，の 2 つが考えられる．これらのうち①については，解読が困難な字体において形状情報を復元することが現実的でないことを考慮すると，特徴補完に対する自動的な支援が必要となる．そこで，我々は特徴推定法と呼ぶ手法を実現した．この方法では，字体の破損部と残存部が同一のパターンを構成することに着目し，残存部の特徴を破損部にも適用することでユーザへの負担を軽減するものである．しかし，字体の破損が大規模になると特徴の推定誤差が拡大し，有効な特徴の補完が難しくなることが問題であった．

そこで，我々は②のアプローチに基づくテンプレート修正法を実現した[3]．テンプレート修正法の基本的な考え方は，破損した字体（入力パターン）に追加されたグレーゾーンをテンプレートにも重畳し，テンプレートの擬似的な特徴の欠損を生じさせることでパターン間距離の不適当な拡大を是正するというものである．

しかし，一般的なテンプレートは字体（墨部を黒，背景を白とする画像）ではなく，予め字体から抽出されたベクトル情報（特徴ベクトル）である．そこで，画像情報であるグレーゾーンをベクトル情報に適用するためのモデルが必要となる．

特徴ベクトルを作成する際には，字体を格子状の小領域に分割し，それぞれの小領域の中央を頂点とするガウス関数を乗じながら輪郭線の方向特徴を抽出する．第 i 行第 j 列にある小領域の特徴を f_{ij} とするとき，パターン全体の特徴 F を抽出する例を図 4 に示す．

テンプレート修正法ではグレーゾーンによって失われる特徴の割合を下記の方法で推定することで小領域ごとの特徴残存率を算出し，それをテンプレートの各特徴 f に乗ずることでグレーゾーンに応じたテンプレートの修正を効率よく実施できるようにしている．まず，グレーゾーンを含む非線形後の破損字体（正規化後パターン）に含まれるすべての画素に対して，下記の数式で表されるスコア s_{gray} を与える

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

ただし，c は 0 より大きい定数である．次に，正規化後パターンを特徴抽出と同じ小領域に分割し，同じく特徴抽出と同じガウス関数を乗じながら小領域ごとの s_{gray} を集計する．このとき，第 i 行第 j 列にある小領域から得られた集計結果を S_{gray_ij} とする．続いて，正規化後パターンのすべての画素に対してスコア c を与え，先の s_{gray} と同様にガウス関数を乗じながら小領域ご

とに集計する．このとき，第 i 行第 j 列にある小領域から得られた集計結果を S_{ij} とする．

ここで，第 i 行第 j 列における特徴残存率 R_{ij} を次のように定義する．

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

テンプレート修正法では，テンプレートの各特徴 f_{ij} を $R_{ij} \times f_{ij}$ で置き換える．一方で，正規化後の破損したパターンからは黒画素の特徴だけを抽出して特徴ベクトルを作成する．これらを比較して距離を算出することで，計算量の増加を抑えながら類似度評価の精度を高めることが可能となる．

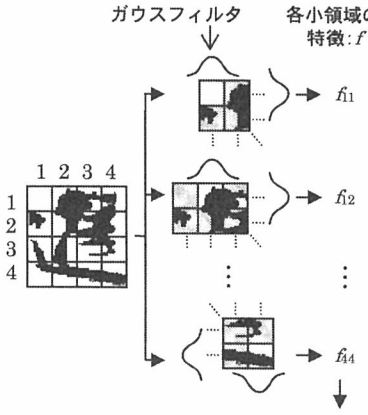

図 4. ガウス関数を用いた特徴抽出
（小領域：4×4）

3. Mokkanshop における字体検索機能の実装

図 5 に，Mokkanshop が持つ字体検索機能とデジタルアーカイブ参照機能のそれぞれのウィンドウ，およびグレーゾーンの指定（塗り分けを含む）を行う画像処理ウィンドウを示す．

グレーゾーンを指定する際には，非線形正規化の結果に注意しながら灰色の濃度を決定する必要がある．しかし，利用者が木簡の解読に集中できる環境を実現する上で，濃度を絶対値で入力/選択するユーザインタフェースは適当とはいえない．そこで，グレーゾーンを指定する際には薄い/濃いといった相対的かつ曖昧な濃度の指定だけを行い，検索時に濃度の微調整を行いながら検索のやり直し，検索結果の絞り込みを行える仕様としている．

また，字体検索の結果については字種だけでなく字体の表示が行えるようにしている．字体検索機能の目的は専門家が解読（釈文作成）を行う上で有用な情報を提供することにある．特に古代日本では部首の位置を変化させた字体，現在のフォントでは表現できない異体字などが多数利用されており，字種を識別するための情報だけでは検索結果として不十分である．そこで，字体検索機能を司る破損字体認識ウィンドウの下部にテンプレートの元になった古代の字体（モノクロ画像）とそのメタデータを表示する領域を設けている．さらに，出典となる古代木簡の情報が参照できる場合は，奈良文化財研究所が公開している古代木簡デジタルアーカイブ「木簡辞典」をインターネット経由で参照することにより，木簡画像（RGB，赤外），発掘場所などの情報をデジタルアーカイブ参照ウィンドウに表示することが可能である．

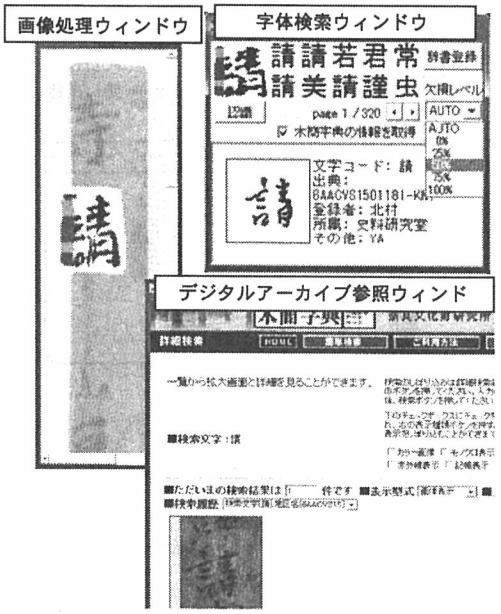

図 5. Mokkanshop における字体検索機能の
ユーザインタフェース

しかし，ユーザの思考を字体検索に反映し，検索精度を高めるためには，グレーゾーンに関する支援だけでは不完全である．そこで，我々は専門家が字体解読の際に行う筆跡追跡の作業に注目し，簡単な操作で欠損した筆跡の追加が行える筆跡追跡機能を実装した．

木簡の解読を行う専門家は，木簡表面および残存する墨の状態を見ながら失われた筆跡を追跡・推定し，高い精度で字体の解読を行う．コンピュータを用いた字体検索においても，推定された筆跡を画像処理による墨抽出の結果に加筆することで，検索時に利用する情報を増やし

た有効な絞り込み検索を実現できる可能性がある．ただし，加筆の際には筆の太さに注意する必要がある．

加筆に用いる筆の太さは，字体本来の筆の太さと一致させることが望ましい．これは，パターンマッチングにおける非線形正規化，類似度評価の手法が筆の太さの変化に対して必ずしも頑健ではないためである．しかし，適切な筆の太さを選択しながら筆跡の追跡・推定にも意識を集中することはユーザへの思考的な負担が大きく現実的ではない．実用的な観点からすれば，加筆を行う際の筆の太さに依存しない手法を実現する必要がある．

そこで，我々は細線化と拡張（dilation）を用いて筆の太さの正規化を行い，この問題に対する解決を試みた．この方法では字体に対して細線化，拡張の順番で処理を行い，細線化によるヒゲの影響を抑えながら筆（ストローク）の太さを正規化する（図 6）．これにより，字体本来の筆の太さと加筆時の筆の太さの差をある程度吸収することが可能となる．その上で非線形正規化処理を行い，破損した字体の特徴ベクトルを作成する．また，テンプレートとなる特徴ベクトルを作成する際にも同様に細線化と拡張を行うことで類似度評価の精度に配慮している．

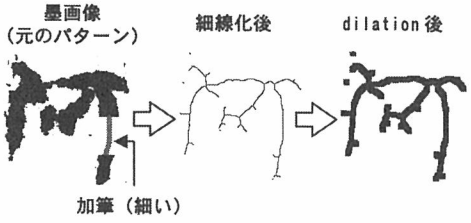

図 6. 細線化と dilation 処理の例

図 7 に筆跡追跡（加筆）のユーザインタフェースを，図 8 に加筆による検索結果の変化を，それぞれ示す．

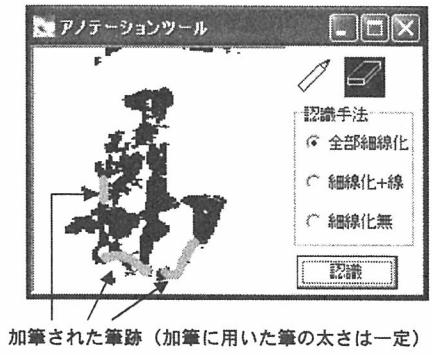

図 7. 細線化と dilation 処理の例

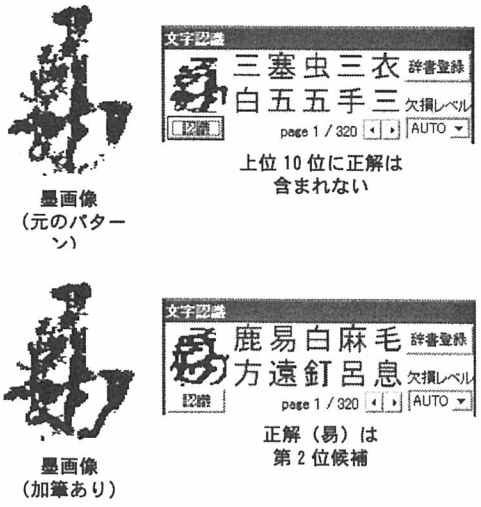

図 8. 加筆による検索結果の変化

なお，加筆を行わない場合の細線化＋dilation 処理による影響については現在検証中である．図 9 に，加筆とは無関係に，細線化＋dilation の有無で検索結果が変化する例を示す．

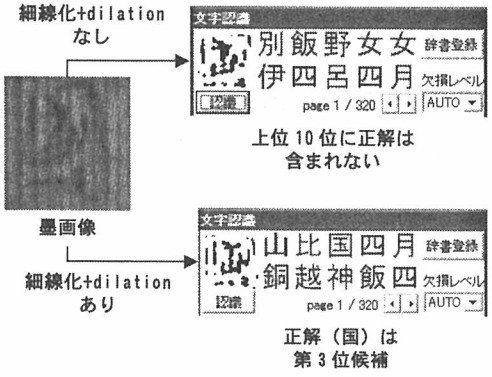

図 9. 細線化＋dilation の有無（加筆なし）と検索結果

4. あとがき

本稿では，古代木簡上の破損した字体の解読を支援する字体検索機能の高性能化と木簡解読支援システム上での実装について述べた．

字体検索は，テキスト情報が得られない難読文書をキーとしてデジタルアーカイブを検索する有効な手段であり，その高精度化はデジタルアーカイブの有効利用を促進する上で欠かせない．

なお，本稿で述べた技術のうち細線化＋dilationによる検索精度への影響，および専門家による評価実験については今後の課題である．

参考文献

[1] 高倉純, 他: 木簡解読支援のための情報検索, 人文科学とコンピュータシンポジウム論文集, Vol.2008, No.15, pp.75-80, 2008.

[2] H. Yamada, et al.: A Nonlinear Normalization Method for Handprinted Kanji Character Recognition ---Line Density Equalization---, Proc. 9th ICPR, pp.172-175, 1988.

[3] 耒代誠仁, 他: 古代木簡解読支援のための文字パターン検索, 情処論, Vol.50, No.4, pp.1444-1455, 2009.

[4] http://jiten.nabunken.go.jp