古代木簡解読支援のための画像処理および字体検索の高度化

耒代 誠仁　　中川 正樹　　馬場 基　　渡辺 晃宏
桜美林大学 総合科学系　　東京農工大学 工学府　　奈良文化財研究所 史料研究室

本稿では古代木簡解読を支援する画像処理および字体検索の高度化に関する我々の研究について述べる．画像処理の改善は，汚損・破損した古代木簡の可読性を高め，高精度な字体抽出を実現するために重要である．我々はカラーチャネルおよび周波数に関する分析結果を踏まえ，木目，腐食に有効な画像処理を実現した．また，木簡解読支援システムのユーザインタフェースを改良し，画像処理の実施に必要な操作を簡素化した．字体検索については，テンプレートを増やすと共に形状特徴抽出の改善を行い，検索精度の改善を実現した．

Improvements of image processing and character pattern retrieval methods to support reading historical mokkans

Akihito Kitadai
J. F. Oberlin University

Masaki Nakagawa
Tokyo University of Agri. & Tech.

Hajime Baba　Akihiro Watanabe
Nara National Research Institute for Cultural Properties

This paper presents our research which improves image processing and character pattern retrieval to support reading historical mokkans. The improvement of image processing is important for high readability of the historical mokkans with stains/degradations and for high accuracy character pattern extraction. We performed analyses of color channels and frequencies of mokkan digital images, and implemented image processing methods to remove grains and corrosions on the images. Also, in the improvements of character retrieval, we incremented the templates and upgraded the shape feature extraction method. The upgraded method improves the performance of character pattern retrieval.

１．まえがき

木簡とは木片に墨で文字が記された文書の総称である．国内で広く用いられたのは奈良時代前後とされ，該当する時代の遺跡からは多数の木簡が発掘されている．古代日本で用いられた木簡は古代木簡と呼ばれ，これまでに約 35,000 点が発見されている（図 1）．

古代木簡には荷札・メモ書き・事務連絡などに利用されたと考えられるものが数多く見られる．記録媒体となる数 cm～数十 cm の木片は入手性・耐候性に優れ，また持ち運びにも便利であったことを考えると，これらの用途は合理的であったといえる．そして，その用途故に，古代木簡には物流・地域間交流・人名・地名・日々の出来事など，当時の客観的事象を知るための貴重な情報が記録されることになった．史学・考古学の分野において，古代木簡の解読結果は大きな注目を集めている．

しかし，古代木簡の保存状況は極めて深刻である．古代木簡の大部分は遺跡内の水路，井戸，ごみ捨て場跡など文書の保存には適さない地中から発掘されている．これは，前述した用途で作成・利用された木簡を長期保存する必要性が薄く，利用後には破棄されたためと考えられる．さらに，木片を転用するため，あるいは使用済木簡の誤用を防ぐため，人為的に破壊されたと考えられる古代木簡も多い．このような理由から，古代木簡の解読は専門家にとっても困難な作業となっている．

図 1. 古代木簡（奈良文化財研究所）

腐食・変色など現在進行形で劣化が進む古代木簡の今の姿を記録し，古代木簡が持つ情報の保存・活用を可能にするため，我々は古代木簡デジタルアーカイブ「木簡字典」の構築と拡張を進めてきた[1]．さらに，パターン認識・画像処理・言語処理などの情報技術を利用した古代木簡解読支援ソフトウェア「Mokkanshop」においては，難読字体をキーとした字体検索機能を提供してきた（図 2）．

図 2. 木簡字典と Mokkanshop による解読支援

Mokkanshop が提供する画像処理機能は，古代木簡のデジタル画像に含まれる墨を強調表示または分離/抽出することで，古代木簡の可読性向上，および後述する字体検索の精度向上を図るものである．また，字体検索機能は断片化した墨（字体）の一部または全部をキーとした字体検索機能を提供する．字体検索の結果は木簡字典へのリンクにもなっており，木簡字典に接続することで難読字体の類例となる字体とその原画像を含む様々なメタデータを参照することが可能となっている[2].

Mokkanshop の有用性は，画像処理の使い勝手と墨識別性能，および字体検索の検索対象範囲と精度に大きく依存する．本稿では，古代木簡に見られる汚損・変色に簡単な操作で対応可能な画像処理機能の実装，および字体検索機能で使用するテンプレートの拡充と特徴抽出手法の改善について述べる．

2．画像処理の高度化

汚損，経年変性した古代木簡では，腐食部，木目などの色が強くなり，一方で墨の脱色が進行している．そこで，少しでも多くの墨を他と区別して抽出，強調することが，木簡解読支援のための画像処理の重要な役割となる．ただし，画像処理を施す際の操作が複雑になると，利用者への思考的負担が増大し，解読作業への障害となる恐れがある．そこで，古代木簡に対象を絞った有効な画像処理の選択，およびユーザインタフェースの操作性向上も課題となる．

我々は，これまでの研究を通して様々なカラーチャネルを通した古代木簡画像の分析を行い，有効と判断された手法を実装してきた．これに加えて，最近の研究では周波数分解による効果を調べ，有効性を報告した[4]．しかし，画像処理を直接的に実装した結果，利用者が複数のパラメタを同時に調整する必要が生じ，使い勝手の面では課題が残っていた（図 3）．

図 3. 従来のパラメタ調整用メニュー (Mokkanshop)

利用者への思考的負担を低減するためには，グレースケール画像に対する単純 2 値化のようにスライダバー1 つで操作可能な実装が好ましい．そこで，本研究ではこれまでの成果を精査し，パラメタの調整を簡略化した画像処理を実装した．以下に，新しい実装において利用者に提供する画像処理を示す．

(1) 明度評価法
(2) RGB 最大値評価法
(3) CMY 最大値評価法
(4) 6ch 最大値評価法
(5) YM 比評価法

これらのうち，(1)は従来から提供しているグレースケール画像の単純 2 値化手法である．一方で，(2)～(5)は墨とそれ以外における各色成分

の構成に着目したものである．墨は黒に近い色であり，色成分の偏りが比較的小さい．一方で，墨以外の画素では木片などに由来する特定の色成分が残るため，この違いを利用して墨の判別を行う．

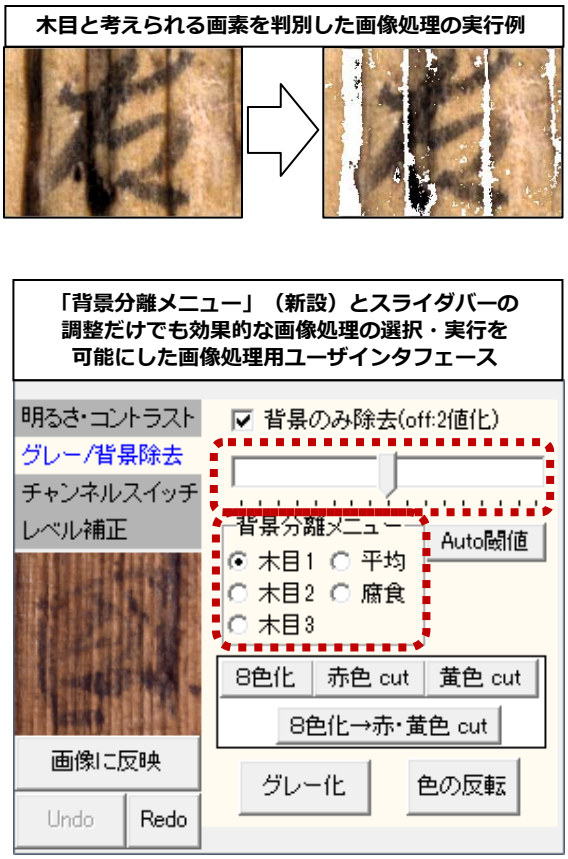

図 4. 新しい画像処理の実装

なお，(5)は木片の茶色が黄色(Y)を強く含むことを考慮した手法である．これら(1)～(5)はすべて単一のパラメタで制御可能なため，必要となるスライダバーは 1 つである．図 4 に(1)-(5)の画像処理を提供する Mokkanshop のユーザインタフェース，および画像処理の例を示す．ユーザインタフェースでは，前述の(2)/(3)/(4)をそれぞれ「木目 1」/「木目 2」/「木目 3」，(5)を「腐食」，また(1)を「平均」というラベルで選択するようにした．これは各処理の特性を目的に合わせて表したもので，例えば木片の腐食部分では色成分の平滑化が進みやすく，茶色に対する判定の有効性が相対的に高くなる．ただし，これは防腐溶液中での保存が主体となる古代木簡に対する考え方であり，他の古文書では異なる手法とラベルが必要になると考えられる．

なお，Mokkanshop では「腐食」と「木目 1」を連続的に適用するといった画像処理の連結により墨の抽出精度を高めることが可能である．また，詳細なパラメタ設定が必要な場合は図 3 のメニューも引き続き利用可能である．

3．字体検索の高度化

字体検索は，前述した画像処理によって抽出された墨の形状を既知の字体と比較し，類似度が高いものを提示する情報検索である．解読困難な古代木簡に残る字体を検索キーとして，デジタルアーカイブに記録された類似度が高い字体および古代木簡の画像，メタデータなどを提示することで，難読木簡の解読を支援する類例検索が可能となる．

字体検索は，キーとなる字体の字種が不明な場合でも機能する．類似度による動的なリンクといえるこの特性を，メタデータによる静的なリンクと組み合わせることができれば，デジタルアーカイブの新しい活用方法に繋がる可能性がある．ただし，動的なリンクの有用性は字体の形状評価に用いるパターンマッチング手法の精度に依存する．我々は，キーの形状欠損を簡単なアノテーションで補完可能にするグレーゾーン法を提案し，文字認識で効果が高い非線形正規化を字体検索に導入することを可能にした．この手法には，グレーゾーンの色濃度を変更することで検索結果の絞込みを行う加重平均算出も含まれる（図 5）．また，グレーゾーンを適用した場合の字体検索精度を向上させるテンプレート修正法を提案・実装し，字体検索の有用性向上を実現してきた[4].

しかし，字体検索の精度に大きな影響を及ぼす形状特徴の抽出手法についての議論は十分とはいえない．これまで我々は，手書き漢字認識で有効性が高い 2 値画像の輪郭特徴（チェインコード）を採用してきた．一方で，近年では 2 値画像を平滑化したグレースケール画像から字体輪郭部の勾配特徴（画素間の輝度差）を抽出・利用する手法の有効性が示されている[5].

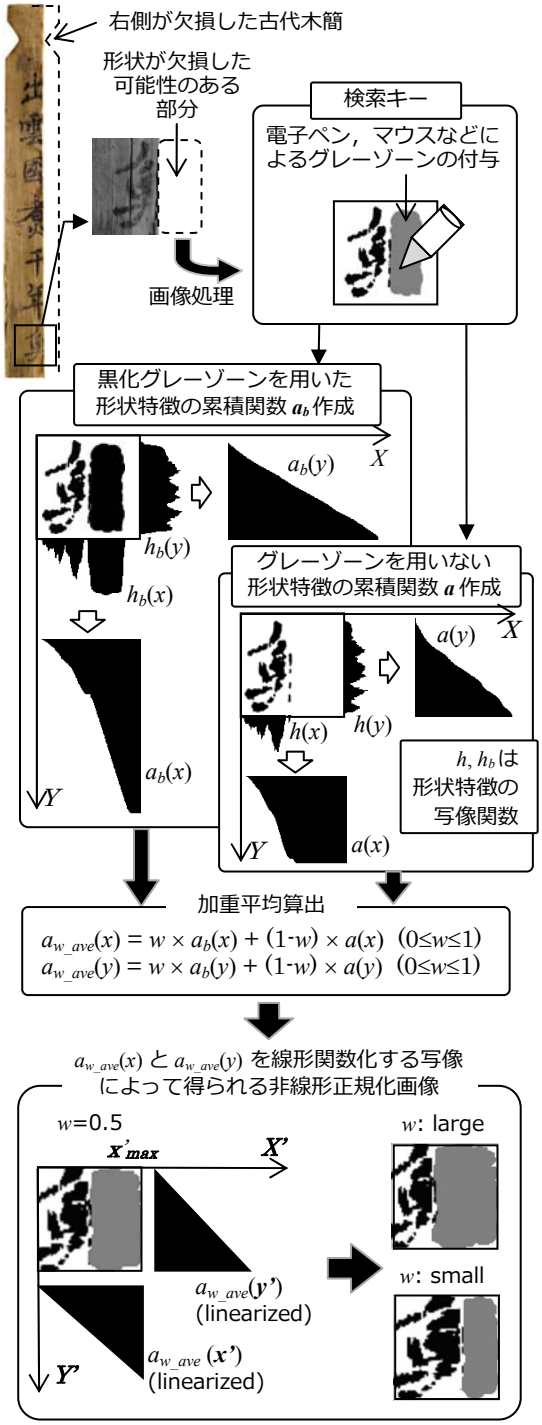

図 5. グレーゾーン法による非線形正規化

勾配特徴は大規模な字体データベースを使った機械学習と併用されることが多い．一方で，形状特徴の欠損を伴う字体の機械学習については現時点では課題が多い．しかし，機械学習を用いない場合でも，グレースケール画像から得られる柔軟な特徴量が，輪郭部分にノイズを多く含む古代木簡上の字体に適している可能性は高い．そこで，本稿では古代木簡から抽出した字体に適用可能な勾配特徴抽出手法を提案し，その評価を行う．

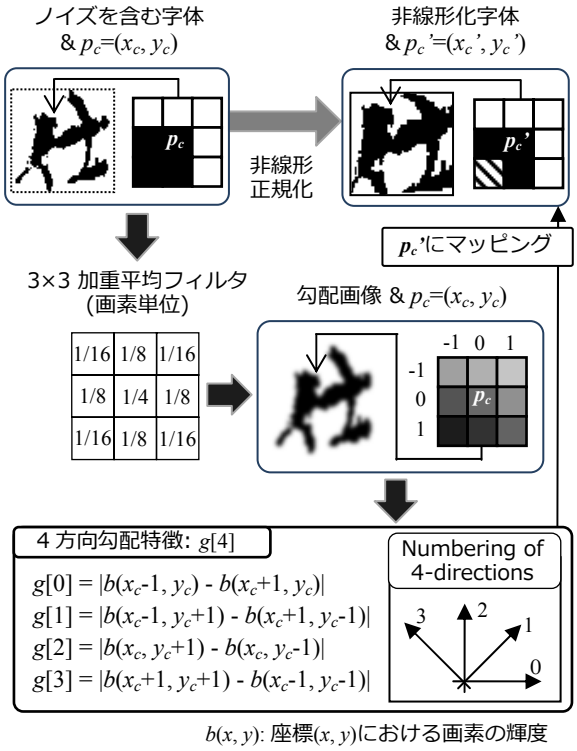

図 6. 勾配特徴抽出処理

図 6 に，輪郭部分にノイズを含む字体の勾配特徴抽出処理を示す．ここで，元画像の輪郭上の画素 $p_c=(x_c, y_c)$は，非線形正規化処理によって $p_c'=(x_c', y_c')$に写像されるものとする．勾配特徴を抽出する画像（勾配画像）には，非線形正規化前の字体（2 値画像）に加重平均フィルタを適用したものを利用する．ただし，加重平均フィルタを一度だけ適用した画像では，輪郭部分に曲線的な勾配が生成される．そこで，加重平均フィルタの適用回数を 3 回とし，輪郭部分の勾配を直線的なものとする．

輪郭上の画素 1 つにつき抽出する勾配特徴は，メモリ使用量を考慮して 4 方向とした．また，画素 p_c において抽出した勾配特徴は，非線形正規化字体上で対応する p_c'にマッピングし，字形評価に用いることとした．なお，チェインコードを用いた従来の手法でも輪郭特徴は 4 方向としているため，今回の提案手法においてもメモリ使用量は変化しない．

評価実験に用いる字体は，木簡解読に関わる専門家が古代木簡から抽出した 4,911 パターンとした．字体の大きさを揃えるため，外接矩形の長辺が 64pixel になるように縦横比を保存した拡大・縮小処理を施した（図 7）．

図 7. 古代木簡から抽出した字体画像

評価方法には Leave-one-out cross validation 法を用いる[6]．すなわち，1 つのパターンをキーとし，残りの 4,910 パターンをテンプレートとした試行を 4,911 回実施する．また，各試行において，キーと同じ字種のパターンが検索候補上位 10 位に含まれた率を検索率とする．表 1 に，チェインコードと勾配特徴を用いた場合の検索率を示す．

表 1. 特徴抽出手法別の検索率

特徴抽出	検索率 （10 位候補含有率）
チェインコード	81.7 % (4,012/4,911)
勾配特徴	82.5 % (4,049/4,911)

次に，字体に形状特徴の欠損が発生した場合，およびグレーゾーンを付与した場合の評価実験について述べる．ここでは，それぞれの試行においてキーとなる字体に 10 個のマスクパターンを適用し，擬似的な欠損およびグレーゾーンを生成する（図 8）．

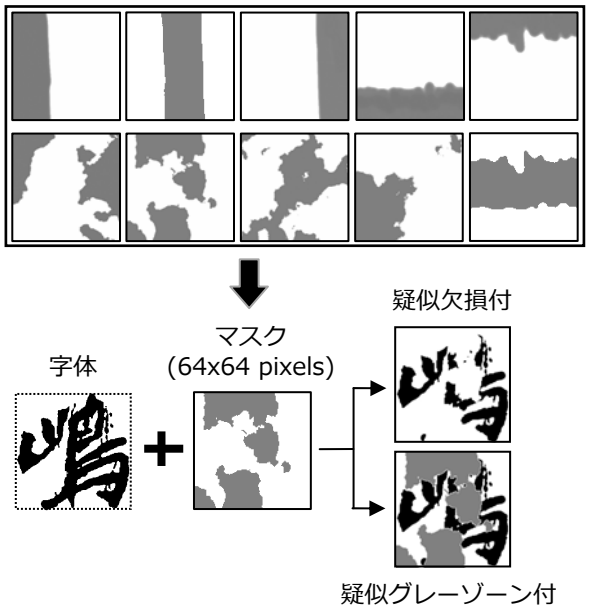

図 8. マスクを用いたキーの生成

生成されるキーに対する総試行回数はそれぞれ 4,911×10=49,110 回となる．ただし，そのうち 94 回で字体を構成するすべての黒画素がマスクに覆われたため，これを除外して 49,016 回の試行に対する検索率を求めた．表 2 に疑似欠損付キーに対する結果を，表 3，4 に疑似グレーゾーン付キーに対する結果を示す．ただし，表 3 ではグレーゾーン法における係数 w を 0.5 に固定（単純平均）しており，また表 4 では各試行において w を 0.0/0.25/0.5/0.75/1.0 の中で最適なものに設定（加重平均による理論値）している．

表 2. 疑似欠損付キーに対する検索率

特徴抽出	検索率
チェインコード	44.2 % (21,667/49,016)
勾配特徴	46.9 % (23,000/49.016)

表 3. 疑似グレーゾーン付キーに対する検索率（w=0.5）

特徴抽出	検索率
チェインコード	60.4 % (29,608/49,016)
勾配特徴	61.8 % (30,307/49.016)

表 4. 疑似グレーゾーン付キーに対する検索率（w=最適値）

特徴抽出	検索率
チェインコード	72.7 % (35,639/49,016)
勾配特徴	74.8 % (36,639/49.016)

以上の結果から，勾配特徴を用いた場合の検索率は，すべての実験においてチェインコードを用いた従来手法を上回った．このことは，古代木簡解読支援のための字体検索において勾配特徴の利用が有効に機能する可能性を示している．ただし，今回の勾配特徴を用いた手法は暫定的な実装であり，今後の継続的な改善が必要と考えられる．

本章の最後に，Mokkanshop で使用しているテンプレートの拡充について述べる．テンプレートは木簡字典に登録されている字体から生成され，字体検索における類似度計算に利用されるとともに，検索結果として Mokkanshop 利用者に提示される．また，当該字体が Web で公開されている場合には木簡字典へのリンクを提供する．我々は木簡字典の拡張と共にテンプレートについても拡充を進めてきた．その結果，2004 年のじんもんこんで公開した初版の Mokkanshop では約 300 字種分であったテンプレートを，現在では古代木簡に多く利用されている 652 字種分まで拡張することができた．

６．あとがき

本稿では，木簡解読支援のための画像処理および字体検索の高度化について述べた．

情報検索は，デジタルアーカイブに蓄積された，そして今後も蓄積される膨大な知識を有効利用する上で不可欠な要素技術である．現在，多くの情報検索はメタデータによる静的なリンクによって実現されている．一方で，画像処理，字体検索などが提供する情報検索は，抽出された墨の状況，付与するアノテーションなどによって変化する動的なリンクを提供する．これらの静的/動的リンクは優劣を比較するものではなく，共存することによってデジタルアーカイブの有用性を高め，活用範囲を広げる可能性を有する共栄共存の関係として今後議論されるべきものと考える．

この点を踏まえ，今後の課題としては，静的/動的リンクの有効な併用方法の提案・実現，および広範囲な時代/文書に適用可能なパターンマッチング技術の提供などが挙げられる．

7．謝辞

本研究は科研費基盤 S-20222002 および若手 B-22720239 の助成を受けたものである．

参考文献

[1] 木簡字典: http://jiten.nabunken.go.jp.
[2] Sherini Somayeh, 来代誠仁, 中川正樹, 馬場基, 渡辺晃宏: 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, 人文科学とコンピュータシンポジウム論文集, Vol.2010, No.15, pp.27-32, 2010.
[3] J. Takakura, A. Kitadai, M. Nakagawa, H. Baba and A. Watanabe: Techniques to Enhance Images for Mokkan Interpretation, Proc. 12th ICFHR, Vol.1, No.1, pp.358-362, Kolkata, India, Novenber 2010.
[4] 来代誠仁, 他: 古代木簡解読支援のための文字パターン検索, 情報処理学会論文誌, Vol.50, No.4, pp.1444-1455, 2009.
[5] C-L. Liu: Handwritten Chinese Character Recognition: Effects of Shape Normalization and Feature Extraction, Lecture Notes in Computer Science, Vol. 4768/2008, pp.104-128, 2008.
[6] J.W.Tukey: Bias and confidence in not-quite large samples, Ann. Math. Statist., Vol.29, pp.614, 1958 (abstract).