文字認識手法による古筆分類の一方法

山田 奨治†, 金子 裕之‡
†国際日本文化研究センター, ‡奈良国立文化財研究所

本研究では，文字認識手法の応用による古筆分類法を提案・検証した．文字特徴量として，射影ヒストグラムに基づく小領域分割と方向線素ヒストグラムに基づくCFV-1特徴量を定義した．百万塔墨書を例に，①同一銘の百万塔はひとりの人物によって墨書されているか，②異体字がある場合に同一人物による文字と判定できるか，の観点から専門家による判定結果とCFV-1特徴量による結果を比較検討した．総サンプル数35件について，専門家による判定と提案手法の一致率88.6％が得られた．これらの結果，提案方法が概ね妥当であり，専門家支援のための自動分類に利用可能なことがうかがえた．

Classifying Early Handwritten Manuscripts Using a Character Recognition Approach

Shoji YAMADA† and Hiroyuki KANEKO‡
†International Research Center for Japanese Studies
‡Nara National Cultural Properties Research Institute

In this article, we proposed and examined a classification method for handwritten characters in historical objects using a character recognition approach. We defined a character feature (CFV-1) based on projection and directional element histogram methods. Using signatures on Nara period stupas known as Hyakuman-To, we compared the assessment of experts and the results of the proposed method on two points: (1) whether or not Hyakuman-To with the same manufacture's names are signed by the same person; (2) whether or not Hyakuman-To with phonetically identical manufacture's names but different characters are signed by the same person. The coincidence rate between the assessment of experts and the proposed method was 88.6 percent for 35 samples. We concluded that the proposed method is supportable and has potential use in an automatic determination system to facilitate and support the assessment of experts.

1 はじめに

和紙や木簡などに墨書された歴史的資料の分析において，その筆跡を鑑別して同一人によるものであるか否かを判定することは，歴史研究の基礎的作業である．署名のない資料はもとより，署名のあるものに関しても模写や贋作の可能性を探るためには，筆跡の鑑別が不可欠である．筆跡鑑別の完全自動化は無理としても，資料が大量に存在する場合には，予備的な分類を自動的におこなって，筆跡鑑別の専門家を支援することは可能であろう．文字の類似度判定には，文字外形の類似性をみることが，ひとつのアプローチとして考えられる．人間が筆跡鑑別をおこなう場合においても，文字外形を重要な情報として用いていることが予想できる．文字外形による自動分類には，文字認識に用いられている技術が応

用可能であると考える．しかしながら，歴史的文書に関する文字認識の研究は，あまり例がない [1][2]．本研究では，実際の歴史的資料に対して文字認識手法を適用し，専門家支援のための筆跡予備分類を自動化可能であるかを検証する．

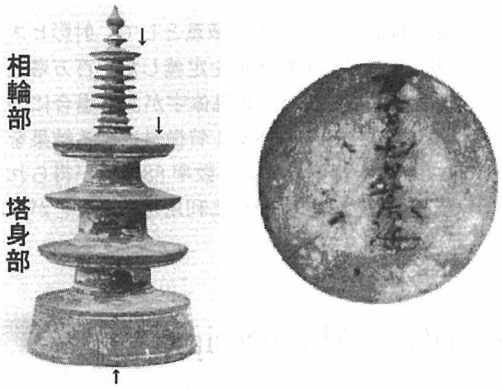

図 1: （左）百万塔の部位名称と墨書位置（矢印），（右）墨書の 1 例，（文献 [3] より改変して引用）

使用する歴史資料は，法隆寺現存の国宝・百万塔（図 1）に記されている製作工人による署名である．百万塔の相輪・塔身の 90 %～ 95 %には，製作工人の署名や日付，工房名「左右」などの墨書がある．工人の署名は，くせ字が多いうえに姓名を書くことは少なく，姓あるいは名前を 1 字に略すことや，音が同じ別字で書く（虎→虎）など無数の形がある．さらに「代筆」もあることが，問題を複雑にさせている．これらの文字をいかに同定するかが，百万塔研究の最大の課題となっている．百万塔製作などにあたった官営工房は，古代国家になくてはならない重要な生産機構である．官営工房の組織・運営方法の解明は，古代国家機構そのものの解明につながる．百万塔をもとに，その工房の実態を明らかにすることが可能である．

本論文では，最初に文字特徴量算出法を定義し，ついで比較文字の選定，類似度の算出と分類をおこない，人間の専門家による分類結果と比較する．それにより，本手法が専門家支援のための自動分類手法として有効が否かを検証する．

2 文字特徴量の定義

本研究においては，文字外形に着目する方向線素ヒストグラム法にもとづいた特徴量を用いる．ここで定義する文字特徴量を，CFV-1(Character Feature Values - 1) とよぶことにする．CFV-1 の算出手順は，①輪郭線抽出，②射影ヒストグラムに基づく小領域分割，③方向線素による特徴抽出，④重みづけ，⑤領域統合と正規化である．

切り出された 1 文字に対して，最初に輪郭線抽出をおこなう．つづいて，1 文字を 6 × 6 の 36 小領域に分割する．小領域分割は，輪郭線化された文字の垂直・水平方向の画素合計（射影ヒストグラム）を均等に 6 分割することでおこなう．垂直・水平方向の小領域境界 b_i は，つぎのようになる．

$$b_i = \max_j(\sum_{k=0}^{j} h_k \leq i \cdot \frac{N_B}{6}), (i = 0, \cdots, 6) \quad (1)$$

ここで，h_k は輪郭線化された文字の射影ヒストグラム，N_B は輪郭線の総画素数である．小領域の総数はこの時点で 36 となる．小領域分割の例を，図 2 に示した．

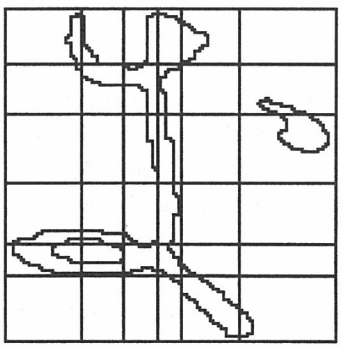

図 2: 小領域分割の例

射影ヒストグラムに基づく小領域分割により，くずしなどの文字変形に頑健な分割をおこなうことができる．図 3 は，くずしの異なる文字を上記の方式で小領域に分割した例である．従来手法によくみられるような，均等に小領域分割する方式と比較して，くずしによる文字変形に応じた分割になっていることがわかる．

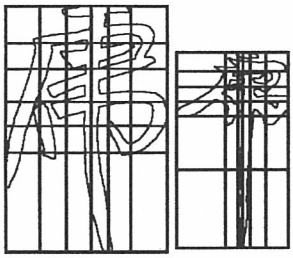

図 3: くずしの異なる文字の小領域分割

特徴抽出は，注目画素近傍の方向線素抽出によりおこなう．文字の輪郭線を構成するすべての画素に関して，黒画素近傍の輪郭線がどの方向に延びているかを図 4 の 4 パターンに分類（方向線素ベクトル化）し，各小領域中のすべての黒画素が 4 パターンのどれにあたるかを算出する．注目画素近傍が 4 パターンのいずれにもあてはまらない場合は，形が近い 2 パターンに対して半分づつの寄与を与える．この処理により 36 小領域すべてについて，4 種類のパターンがいくつあらわれるかが求まり，1 文字が 36 × 4 の特徴ベクトルに変換される．

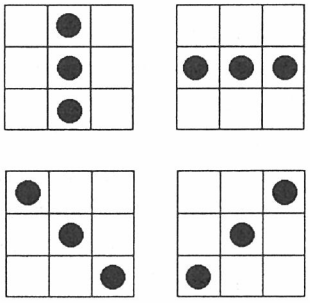

図 4: 方向線素パターン

つづいて小領域の特徴ベクトルに対して，図 5 のような重みをかける．これは百万塔墨書や古文書にあらわれる文字を観察した結果，文字の左上，中心線，右下部に特徴的な差が多いことが見いだされたからである．これは文字の左上には起筆部，右下には終筆部，中心線には文字を特徴づける主要な形状があることに対応している．さらに，近傍 4 小領域で特徴ベクトルを加算し，小領域をひとつづつずらせながら，5 × 5 の 25 領域での特徴ベクトルを求める（図 5 矢印）．これは文字にぼかしをかける処理に相当し，文字形状のバリエーションへの頑健性を高める効果がある．最終的な特徴ベクトル数は，25 × 4 の 100 次元となる．さらに特徴ベクトルを，文字の輪郭線長で割り算して正規化をほどこし，文字の大きさによる影響を除去する．

CFV-1 は，古文書かな文字の認識実験に適用した結果，ある程度の認識率を得られることが確認されている [4].

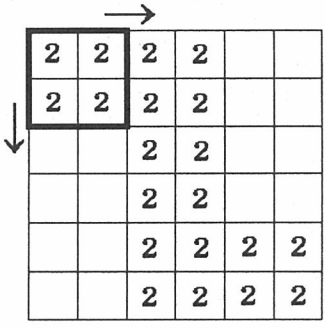

図 5: 特徴ベクトルの重みづけと加算処理

3 百万塔墨書の分析

3.1 対象の選定

実験対象は文字の劣化状態と，比較のために使用可能な墨書内容を検討したうえで決定した．予備的検討の対象としてまず，比較的出現頻度の高い工人名である「荒海」の墨書から，状態の良い数点を選択した．「荒海」に関しては分類性能の評価基準として，専門家 4 名に依頼して文字の似た順に配列してもらった．CFV-1 特徴量による結果と専門家による配列結果を比較することで，手法の妥当性についての検討がおこなえる．

ついで，課題①：同一銘の百万塔はひとりの人物によって墨書されているか，課題②：異体字がある場合に同一人物による文字と判定できるか，の観点から対象とする工人名を選択した．課題①の観点か

らは「和万呂」「龍万呂」「建部雋」「浄人」の工人名が記されたものを，課題②の観点からは「和万呂と倭万呂」の工人名が記されたものをとりあげた．百万塔墨書は劣化がはげしいため，それぞれの工人の墨書について赤外線写真上で肉眼で輪郭を追えるものに対象を限定した．サンプル件数は，「荒海」：4件，「和万呂」：10件，「龍万呂」：4件，「建部雋」：6件，「浄人」：7件，「倭万呂」：1件である．それぞれの墨書識別番号を，表1に示す．

表 1: 墨書識別番号

工人名	墨書識別番号
荒海	1 2 3 4
和万呂	39 184 501 577 617 626 917 1159 1485 2505
龍万呂	52 295 1489 2675
建部雋	492 710 979 1428 1470 2775
浄人	195 367 863 1148 1319 1352 1419
倭万呂	2241

予備的検討対象である「荒海」については，劣化により「海」の文字外形を追うことが困難であったため，「荒」の1文字について検討をおこなうこととした．

課題①において解析対象とする文字は，以下のように限定した．

1. 「和万呂」「龍万呂」については，すべてのサンプルから取得可能な文字である「和」「万」および「龍」「万」のそれぞれ2文字を選択．
2. 「建部雋」については，すべてのサンプルからから取得可能な文字である「雋」の1文字を選択．
3. 「浄人」については，くずしの激しい「人」の文字を除外し，「浄」の1文字を選択．

課題②は，1文字が異体字である「和万呂」と「倭万呂」が，同一人の筆によるものであるか否かという課題である．「和」と「倭」は文字が異なるため，比較対象として用いることはできない．そこで「和万呂」の表記がある百万塔の中で，「倭万呂」の表記があるものと重複が多い文字を検討対象として選択することにした．「和万呂」「倭万呂」と記されている百万塔の墨書文字は，表2のとおりである．これらの中で，識別番号2241「倭万呂」と重複の多い文字である，「左」と「万」の2文字を検討対象とした．したがって「和万呂」の記載があるもので「左」の文字も書かれてある7件 (501, 617, 626, 917, 1159, 1485, 2505) と「倭万呂」1件 (2241) の，合計8件が検討対象となった．

表 2: 百万塔墨書文字

識別番号	工人名	墨書文字	左	和	万
39	和万呂	十月十六日和万		○	○
184	和万呂	和万		○	○
501	和万呂	云二五五左和万	○	○	○
577	和万呂	和万呂		○	○
617	和万呂	左和万	○	○	○
626	和万呂	云二左和万呂	○	○	○
917	和万呂	云二二月三左和万	○	○	○
1159	和万呂	云二四十八左和万	○	○	○
1485	和万呂	云二四廿六左和万	○	○	○
2505	和万呂	左和万	○	○	○
2241	倭万呂	云二五十三左倭万	○		○

画像は，百万塔墨書部を撮影した赤外線写真をデジタル化しデータベース化したものを使用した．写真撮影ならびにデジタル化・データベース化は，奈良国立文化財研究所においておこなわれた．

3.2 処理手順

まず百万塔墨書画像データベースから対象画像を選択し，ハードコピーをとった．ハードコピーをもとに，輪郭を肉眼で確認しながら手作業で文字外形をトレースした．このトレース画をスキャナ取り込みしたものをもとに，①手作業によるノイズの除去，②文字の塗りつぶし，③3×3近傍でのメディアンフィルタリング，④文字回転角度調整，の前処理をほどこした．文字回転角度調整は，文字ごとに基準となる筆画を定めて，サンプル間で基準筆画の方向がそろうように手作業で文字を回転させた．これはCFV-1文字特徴量が，回転に関して不変でないからである．「荒海」の墨書画像例と，「荒」の1文字を取り出して前処理をした結果を図6に示す．

前処理済み画像に対してCFV-1特徴量を算出し，予備的検討対象である「荒海」については主成分分析を，課題①②の工人については，クラスター分析をおこなった．2文字が検討対象となる「和万呂」

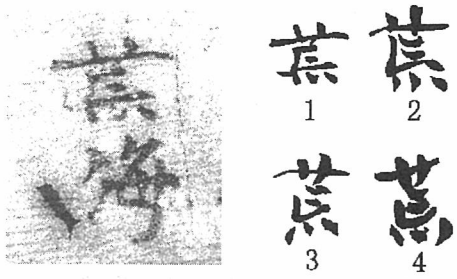

図 6: 「荒海」の墨書画像例と前処理結果

「龍万呂」では，1 文字づつについて求めた CFV-1 特徴量をすべて使用して，200 次元でクラスター分析をおこなった．クラスター分析は，特徴量間の標準化ユークリッド距離によるウォード法を用いた [5].

3.3 処理結果

予備的検討対象の「荒海」について得られた CFV-1 特徴量を主成分分析し，2 次元に配置したのが図 7 である．第 1 主成分の寄与率は 45.0 %，第 2 主成分までの累積寄与率は 76.1 %である．第 1 主成分に注目するならば，文字は 1・2・3・4 の順に並んでいることがわかる．

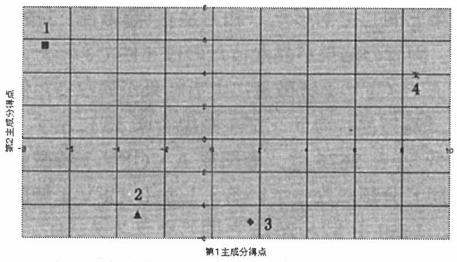

図 7: 「荒海」主成分分析結果

課題①②の「和万呂」「龍万呂」「建部牟」「浄人」および「和万呂と倭万呂」についてクラスター分析した結果を，それぞれ図 8 ～ 12 に示す．

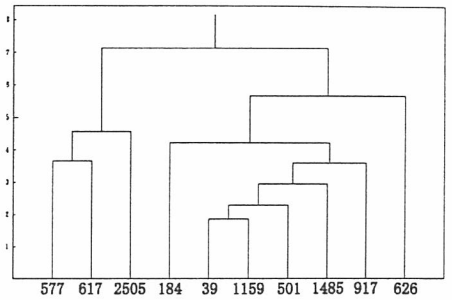

図 8: 「和万呂」(「和」＋「万」) のクラスター

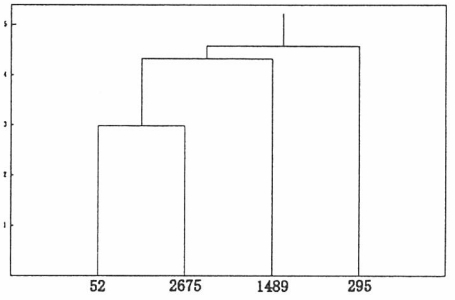

図 9: 「龍万呂」(「龍」＋「万」) のクラスター

4 考察

まず予備的検討として，「荒海」について CFV-1 特徴量による主成分分析の結果あらわれた順序と，専門家によって並べられた似た文字の順序を比較してみる．古代文字の読解を専門とする研究者 4 名を被験者として，4 種類の「荒」の字を似た順序に分類してもらった．分類は文字を 1 次元の主観的な類似度軸上に配置する方法でおこない，類似した文字の順序関係のみに着目して分析した．表 3 がその結果である．Kendall の一致係数は 0.93 であった．表 3 について Friedman 検定の結果，$P < 0.01$ で 4 人の専門家の判断には一貫性があるといえることがわかった．

CFV-1 特徴量による文字類似度の判定は，図 7

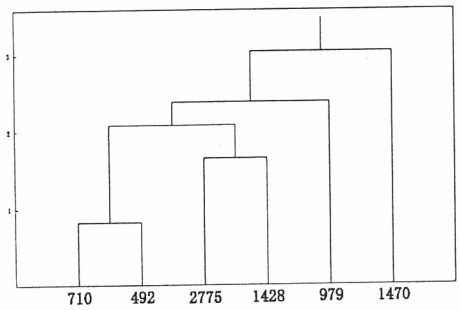

図 10: 「建部庸」(「庸」) のクラスター

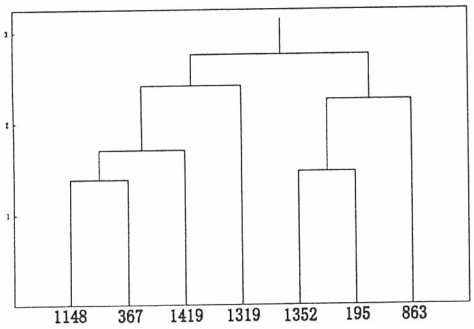

図 11: 「浄人」(「浄」) のクラスター

の第 1 主成分の並びから，被験者 A・D の判断と同じである．主成分分析結果の各成分が，具体的に文字のどのような要素を反映しているのかは，断定できない．一般的に主成分分析では，第 1 主成分に大きさの要因が出ることが多い．本実験の場合は，CFV-1 特徴量算出の最終段階で，特徴量を文字の輪郭線長で割って正規化しているので，単純に大きさの要因が第 1 主成分に表れているとはいえないだろう．以上の結果から，当手法の信頼性はある程度保証されたといえよう．

表 3: 専門家による「荒」の分類

被験者	経験年数	文字の類似順序
A	15	1 2 3 4
B	8	2 1 4 3
C	3	2 3 1 4
D	2	1 2 3 4

Kendall の一致係数 $W = 0.93$

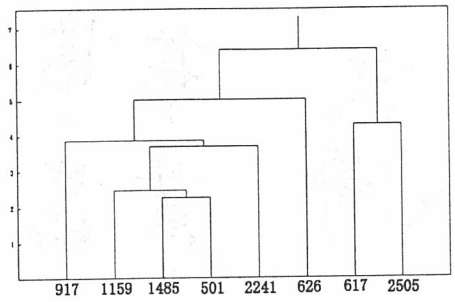

図 12: 「和万呂」と「倭万呂」(「左」＋「万」) のクラスター

ついで，課題①：同一銘の百万塔はひとりの人物によって墨書されているか，課題②：異体字がある場合に同一人物による文字と判定できるか，の考察をおこなう．課題①②の対象について，専門家 A (経験 10 年) に依頼して同筆か異筆かの判定をおこなった．CFV-1 特徴量によるクラスター分析の結果について，どのレベルのしきい値でクラスター分けすればよいかの明確な基準はない．ここでは，専門家 A によって判定されたグループ数に等しいクラスター数が得られるしきい値を採用することとする．専門家による判定結果と CFV-1 特徴量からクラスタリングした結果の比較を，表 4 に示す．

課題①に関して，「建部庸」では CFV-1 特徴量によるクラスタリング結果が，専門家 A による同筆判定結果と同じであった．「和万呂」「龍万呂」「浄人」では，両者の結果が異なったのは 4 件であった．専門家 A と CFV-1 特徴量による結果が異なった例として，「龍万呂」のサンプル画像を図 13 に示す．識別番号 52 と 2675 は，専門家 A・CFV-1 特徴量ともに同じ分類になっている．しかし CFV-1 特徴量では，1489 を 52・2675 と同じクラスターに，295 を別クラスターに分類されたのに対して，専門家 A は逆の分類を示している．

課題②の「和万呂と倭万呂」に関して，表 4 からわかるように，両者ともに「倭万呂」(2241) を「和万呂」のグループ A と同筆と判定している．

これらの結果を総合すると，総サンプル数 35 件に対して結果が異なったのは 4 件であるので，一致率は 88.6 %である．

表 4: 専門家による同筆判定と CFV-1 特徴量によるクラスタリングの結果比較

工人名	専門家 A による同筆判定結果	CFV-1 特徴量によるクラスタリング結果
和万呂	A: 39 184 501 577 917 1159 1485 B: 617 C: 2505 D: 626	A: 39 184 501 917 1159 1485 B: 577 617 C: 2505 D: 626
龍万呂	A: 52 295 2675 B: 1489	A: 52 1489 2675 B: 295
建部牟	A: 492 710 979 1428 2775 B: 1470	A: 492 710 979 1428 2775 B: 1470
浄人	A: 195 863 1148 1352 B: 367 1319 1419	A: 195 863 1352 B: 367 1148 1319 1419
和万呂と倭万呂	A: 501 917 1159 1485 2241 B: 617 C: 2505 D: 626	A: 501 917 1159 1485 2241 B: 617 C: 2505 D: 626

下線は両者の判定結果が異なったもの.

図 13: 「龍万呂」サンプル画像

CFV-1 特徴量による方法の精度を議論するためには，人間の専門家間で判定結果にどの程度の差が生じるのかを検討しておく必要がある．表 5 に，専門家 A に加えて専門家 B（経験 15 年）と C（経験 3 年）の 2 名が協議しながら同筆判定した結果を示した．このように専門家間で判定結果に若干のばらつきがある．表 4 において専門家 A と CFV-1 特徴量によるクラスタリング結果に相違があった「龍万呂」のケースでは，専門家 B+C による判定結果は CFV-1 特徴量による結果と一致している．また，専門家 A と B+C の判断に相違がなかった「和万呂と倭万呂」のケースでは，CFV-1 特徴量によるクラスタリング結果も専門家と同一である．

以上の結果より，CFV-1 特徴量による古筆の分類は専門家による判定結果と比較して概ね妥当であることがわかった．この種のアプローチは，大量に文字資料があった場合の，類似文字の自動分類に利用可能であろう．

5 まとめと今後の課題

以上の結果をまとめると，つぎのようになる．

1. 文字認識手法の応用による古筆分類法を提案・検証した．
2. 文字特徴量として CFV-1 特徴量を定義した．
3. 百万塔墨書を例に，専門家による判定結果と CFV-1 特徴量による方法を比較検討した．

表 5: 同筆判定における専門家間の相違

工人名	専門家 A による同筆判定結果	専門家 B+C による同筆判定結果
和万呂	A: 39 184 501 577 917 1159 1485 B: 617 C: 2505 D: 626	A: 39 184 501 917 1159 1485 B: 617 C: 2505 D: 626 E: 577
龍万呂	A: 52 295 2675 B: 1489	A: 52 1489 2675 B: 295
建部牟	A: 492 710 979 1428 2775 B: 1470	A: 492 710 1428 1470 2775 B: 979
浄人	A: 195 863 1148 1352 B: 367 1319 1419	A: 195 863 1352 B: 367 1148 1419 C: 1319
和万呂と倭万呂	A: 501 917 1159 1485 2241 B: 617 C: 2505 D: 626	A: 501 917 1159 1485 2241 B: 617 C: 2505 D: 626

下線は両者の判定結果が異なったもの.

4. その結果，提案方法が概ね妥当であり，専門家支援のための自動分類に利用可能なことがうかがえた.

以上により，文字認識手法が古筆の分類にも有効であることがわかった．しかしこのアプローチには，特徴量の算出法や分類法の選択によって分類結果に相違が出るという問題点がある．しかしこの問題点は，使用した特徴量と分類法を明確にした上で結果を提示することによって，ある程度解決可能であると考える．どのような方法によって行われたかが明示されることによって，複数の分析結果の相互比較が可能であり，また異なる方法によって試験されることによって数量的分析結果の信頼性を上げることも可能であろう．提案手法は，古筆分類のための最善のものではない．特徴量名称を CFV-1 とした理由もそこにある．今後 CFV-2,3 が提案されることがあっても，手法が明確でありさえすれば，問題は生じないであろう．しかし多くの手法を提示することは，数量的方法になじみの薄い人文科学研究者にかえって混乱を与えることになりかねない．決定的な欠点がない限り，分類手法のバリエーションを多く作ることは適当でないと考える.

今後，百万塔墨書のような古代文字の同定方法が確立できれば，古代官営工房の復元だけでなく，当時の重要な通信手段であった木簡の解読や，木簡の書き手の判別にも利用可能かもしれない.

謝辞

本論文は，文部省科学研究費補助金の平成 9 ～ 10 年度特定領域研究「人文科学とコンピュータ」，同平成 11 年度基盤 (B) 一般研究「古文書解読プロセスの知能情報学的解明」，同平成 11 年度基盤 (B) 展開研究「手書き文字 OCR 技術を援用した古文書翻刻支援システムの開発」の補助を得て実施した研究の成果の一部である.

参考文献

[1] 山田奨治: 高次局所自己相関特徴による古文書かな文字認識, 情報処理学会研究報告, Vol.95-CH-25, pp.21-30, 1995.

[2] 山田奨治: 古文書ＯＣＲ研究の現在, 人文学と情報処理, no.18, pp.2-5, 1998.

[3] 法隆寺昭和資財帳編集委員会: 法隆寺の至宝 百萬塔・陀羅尼経, 小学館, 1991.

[4] 山田奨治: 変体かなの認識実験とその応用, 人文学と情報処理, No.18, pp.71-75, 1998.

[5] 田中豊, 垂水共之: Windows 版統計解析ハンドブック　多変量解析, 共立出版, 1995.