人文科学とコンピュータ　50−2
（２００１．５．２５）

古文書翻刻支援システム開発（ＨＣＲ）プロジェクト報告（２）

○山田奨治 (1), 加藤寧 (2), 並木美太郎 (3), 川口洋 (4)
原正一郎 (5), 石谷康人 (6), 笠谷和比古 (1), 小島正美 (7)
梅田三千雄 (8), 山本和彦 (9), 柴山守 (10)
(1) 国際日本文化研究センター, (2) 東北大学, (3) 東京農工大学
(4) 帝塚山大学, (5) 国文学研究資料館, (6) 東芝, (7) 東北工業大学
(8) 大阪電気通信大学, (9) 岐阜大学, (10) 大阪市立大学

この報文では，平成 11 年度より 3ヶ年の予定で開始した「古文書翻刻支援システム開発（ＨＣＲ）プロジェクト」の概要とねらい，および進行状況について報告する．このプロジェクトは，手書き日本語文字認識技術を発展的に応用して，古文書の翻刻作業を支援するシステムを開発するための諸研究を実施するものである．われわれは現在，(1) 古文書文字データベース，(2) 古文書用例データベース，(3) 古文書文字切り出し，(4) 古文書文字認識，(5) 知識による翻刻支援，(6) 電子化古文書文字辞典について研究を進めている．

Historical Character Recognition (HCR) Project Report (2)

○ YAMADA Shoji(1), KATO Nei(2), NAMIKI Mitarou(3)
KAWAGUCHI Hiroshi(4), HARA Shouichiro(5), ISHITANI Yasuto(6)
KASAYA Kazuhiko(1), KOJMA Masami(7), UMEDA Michio(8)
YAMAMOTO Kazuhiko(9), and SHIBAYAMA Mamoru(10)
(1) International Research Center for Japanese Studies, (2) Tohoku Univ.
(3) Tokyo Univ. of Agriculture and Tech., (4) Tezukayama Univ.
(5) National Institute of Japanese Literature, (6) Toshiba Corp.
(7) Tohoku Institute of Technology, (8) Osaka Electro-Communication Univ.
(9) Gifu Univ., (10) Osaka City Univ.

In this article, we report on the outline and current status of the Historical Character Recognition (HCR) project, which has been conveyed since April 1999. Our project covers various research topics to develop a historical document research supporting system, applying advanced hand-written character recognition technology. We are continuing the following researches for historical document: (1) character database, (2) corpus, (3) character segmentation, (4) intellectual supporting system, and (5) digital dictionary.

1　はじめに

歴史学研究においては，古文書の翻刻が研究プロセスの重要な基礎的作業である．古文書翻刻作業は高度に知的な作業で，歴史の基礎知識，文書の種類やレイアウトに関する知識，定型文言・慣用表現の知識，文字の異体字やくずし方に関する知識と翻刻経験の蓄積が必要であり，人間が古文書翻刻作業をひととおりこなせるようになるまでには，相当の訓練期間を必要とする．古文書翻刻の知的プロセスを解明し，その知見にもとづいて古文書翻刻作業の一部を支援するシステムがあれば，歴史学研究の有効なツールとして活用しうるかもしれない．われわれは古文書翻刻支援をめざしたシステム開発に必要な一連の技術開発を，HCR (Historical Character Recognition) プロジェクトとして立ち上げた [1].

プロジェクトの当面の研究方略は，以下の 4 点である．

1. 対象の選択において，書体の安定した公文書であり歴史的価値のたかいものを対象にする．
2. 文字認識のための辞書構築を進めるために，標準文字データベースを作成する．
3. 古文書読解に関する専門知識を整理し，システム化する．
4. 人間と機械の作業分担を明確化し，両者を円滑につなぐ知的ユーザインタフェースを構築する．

本プロジェクトは，文字のくずしのはなはだしい文書を含むすべての古文書の解読や，古文書解読の完全自動化を目指すものではない．古文書解読プロセスのモデル化とシステムへの実装を通して，古文書解読という高度な知識処理過程を実証的に解明することと，同一文型・書体の文書が大量にあるような古文書の翻刻において，人間の作業負荷軽減に有効なシステム，人間が得意とする作業は人間が，機械が得意とする作業は機械がおこない，両者の円滑なインタラクションが確保できるシステムの開発が狙いである．

2　古文書文字データベース

古文書文字認識の研究を進めるためには，研究者間で共有可能な研究の土台となる文字データベースが必要である．ところが，現状では古文書文字に関してそのようなデータベースは存在しないため，われわれはまずデータベース整備から作業をはじめた．古文書文字認識の試験データとなる文字データベースは，以下の観点から作成している．

1. 用例データとともに文字データが提供でき，知識処理を加えた文字認識の開発に供せられるもの．
2. 歴史研究上の汎用性のたかい文書からの文字．
3. 字種が限られているが，さまざまな筆跡のサンプルが多数得られるもの．
4. 標準的な古文書文字辞典の文字．

1 と 2 の観点からは，大阪市立大学所蔵の「伏見屋善兵衛文書」を取り上げ，そこに登場する全文字の切り出しとデータベース化を進めている．3 の観点からは,「宗門改帳」に記載された文字のデータベース化を進めている．4 の観点からは，古文書翻刻者が利用する標準的な辞書のひとつである，東京堂出版『毛筆版くずし字解読辞典』を選択し，収録されている大部分の文字のデータベース化を完了した．

2.1　「伏見屋善兵衛文書」全文文字データベース

知識処理と組み合わせた古文書翻刻支援を考えた場合，定型文言が頻出するタイプの文書に焦点をあてることが有効である．近世の金子借用証文などは，文書の様式や文言が定型であり，当初の研究対象とするには最適であると判断した．われわれは，上記の条件を満たし種々の権利上の問題もクリアできる研究対象文書として，大阪市立大学が所蔵する「伏見屋善兵衛文書」（以降「伏見屋文書」）（図 1）を選択した．

「伏見屋文書」は，大阪の元伏見坂町（現在の大阪市南区坂町）の茶屋，伏見屋善兵衛家に伝わった文書である．伏見屋善兵衛は，遊興の地である伏見坂町のなかでも最大の茶屋として栄えた．また町年寄をつとめ，芝居興業にも関係し，何軒かの貸家をもち，金融業を営んだ．本文書は，文化から慶応年間にいたる各種の証文類である．芝居関係では，天保年間を中心に歌舞伎役者の芝翫，我童らの手附証文がある．伏見屋の金融・借家，同家内部の親族関係に関する諸証文・議

図 1: 「伏見屋善兵衛文書」

定等も含まれている．文書の総数は，証文類が約 1,300 である．

文書からの文字切り出しとデータベース化は，つぎのような手順で実施している．

1. 原文書をカラーマイクロフィルム撮影
2. カラーマイクロからディジタル化し，紙にプリント
3. プリントされた文書に対し，手作業でカラーマーカーを使って文字ひとつひとつを丸で囲む
4. マーク済みシートをスキャン
5. 自動切り出しソフトで文字を切り出す
6. 文字データと照合しながら校正

手順 3 でマークされたシートは，図 2 のようなものになる．われわれは，このシートから丸で囲まれた領域を自動的に切り出すソフトウェアを開発した．標題部分について文字を切り出し，文字データと照合した結果を図 3 に示した．

平成 13 年 4 月現在,「伏見屋文書」の全標題 4,995 文字の切り出しを完了し，公開に向けたデータベース

図 2: 文字部分をマークしたシート

化の作業を進めている．引き続き,「伏見屋文書」の全文，約 243,000 文字のデータベース化をするべく作業している．全文字のシート上でのマーキング作業はほぼ完了しており，文字切り出し，ノイズ処理とデータベース化を鋭意進めている．

2.2 「宗門改帳」文字データベース

われわれは，字種が限られているがさまざまな筆跡のサンプルが多数得られる文字データベースとして，共同研究者の川口洋が収集した「宗門改帳」記載文字のデータベース化を実施している（表 1）．現在これらのデータを HCD1 (Historical Character Database 1)，1a，1b という名称で公開し，古文書文字認識の基礎実験に供している．HCD1 のシリーズに収録されている字種とサンプル数は，表 2〜4 のとおりである．

表 1: 古文書文字データベース HCD1 シリーズ

名称	内容	字種	文字数	画像
HCD1	年齢表記文字	16	3,066	2 値
HCD1a	単位表記文字	16	3,200	2 値
HCD1b	単位表記文字	8	1,600	2 値

表 2: HCD1 収録の字種とサンプル数

字種	サンプル数	字種	サンプル数
ツ	200	八	200
一	200	九	200
二	200	十	200
三	200	壱	200
四	200	弐	200
五	200	年	200
六	200	拾	200
七	200	廿	66

表 3: HCD1a 収録の字種とサンプル数

字種	サンプル数	字種	サンプル数
田	200	両	200
畑	200	分	200
高	200	朱	200
石	200	家	200
斗	200	軒	200
升	200	間	200
合	200	馬	200
金	200	疋	200

表 4: HCD1b 収録の字種とサンプル数

字種	サンプル数
内	200
男	200
女	200
人	200
〆	200
長	200
横	200
夕	200

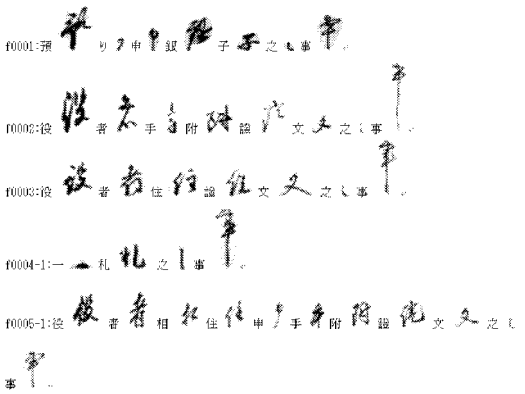

図 3: 文字切り出し結果

2.3 『くずし字解読辞典』文字データベース

「伏見屋文書」や「宗門改帳」といった実際の古文書から採字してデータベース化することも重要であるが，古文書文字辞典に登場するような典型的なくずし字のパターンをデータベース化することも有用であろう．われわれは多くの古文書翻刻者が利用している標準的な辞書のひとつである，東京堂出版『毛筆版くずし字解読辞典』 [2] を選択し，出版社の許諾を得てそのデータベース化を実施した．

データベース化した文字は，同辞典のなかの「付録」を除く本編と増補のかな文字部分全 308 頁に登場する文字と用例，25,202 文字（1 用例も 1 文字とした）である．すべての文字および用例について，画像ファイル名，S-JIS コード，今昔文字鏡コード，読み，今昔文字鏡文字画像への URL を文字データとして作成し，くずし字画像を 400dpi の 2 値で画像取り込みした．

残念ながら，著作権上の理由により当データベースを公開することはできないが，これを活用して後述の古文書文字認識研究，電子化古文書文字辞典の研究を進めている．

2.4 文字切り出し研究用データベース

古文書のつづけ字のなかから 1 文字を切り出すことができたならば，手書き文字認識の技術を適用しやすくなる．ところがつづけ字から正確に文字を切り出すことは，至難である．文字切り出し自体が HCR のおおきな研究テーマでもある．文字切り出し研究を進め

るためには文字の場合と同様，標準的なデータベースを整備して多くの研究者がおなじ土俵で議論ができる環境を整える必要がある．

われわれは，文字切り出し研究用データベースとするために,「伏見屋文書」から標題行を抽出した．ノイズが比較的すくなく 1 行のみからなる標題で，複数の文字から構成され，かつ文字がつづけ字になっている 200 標題を選択して，そのフルカラー画像および翻刻文字をデータベース化した（図 4）．現在，公開にむけた準備を進めている．

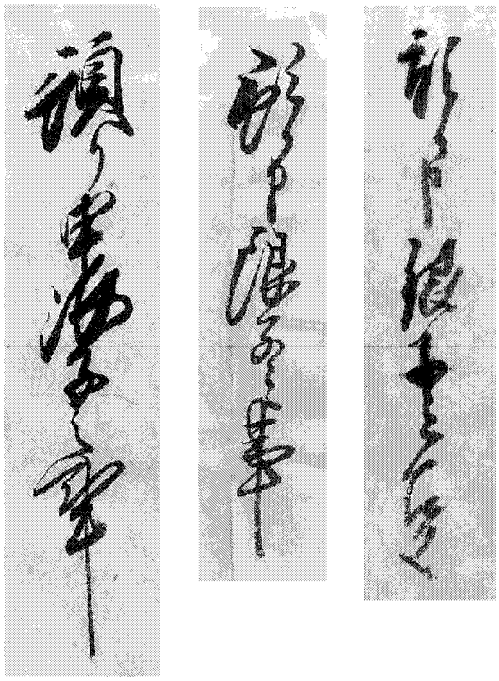

図 4: 文字切り出し研究用データベース収録画像の例

3　古文書用例データベース

古文書に登場する文面の用例を収集することによって，そこから知識を抽出し，その知識を使った古文書翻刻支援が可能となる．またその用例は，定型的な文言が頻出するタイプの文書を収集するのが効果的である．古文書文字データベース作成の対象とした「伏見屋文書」は，そのほとんどが金子借用証文である．証文類は「実正也」「急度返済可申候」「依而如件」などの定型文言が多く見られ，文書の様式も安定しているため，用例データベースの対象として最適である．われわれは，古文書文字データベース作成作業と平行して「伏見屋文書」全文約 243,000 文字を翻刻し，用例データベースとした．作成された用例データベースは，後述の「知識による翻刻支援」研究に利用している．

4　古文書文字切り出し

古文書文字の切り出し，及び文字認識の基礎的研究をおこなうために，古文書標題のみを対象とした文字パターン辞書データベース構築と，関連するユーザインターフェイスの開発を実施した [3]．古文書の形態は縦横の長さ，おおきさが一様でないため，古文書レイアウトの把握や他の古文書との比較が容易にできない．そのため古文書概略画像をピラミッド型の上位層で抽出し，その抽出した抽象化レベルのレイアウトから標題部分だけに着目して原画像から標題部分の抽出をおこなった．

古文書画像のピラミッド型によるレイアウト抽出をおこない，その結果を判断し，標題の抽出を射影ヒストグラム法とラベリング法のふたつの手法を用いておこなった．その結果，78 %の割合で標題抽出をおこなえ，形式が未知である文書の分類が会話型で短時間におこなえるユーサインターフェースを開発した（図 5,6）．しかし，印影や裏写りの影響を受けたものに対しては，本手法では解決されず，また誤って文字の一部分のみ抽出されたものもある．文字の一部分のみ抽出された文書に対する改善は，今後各閾値を一定値から各画像の画素値の分布に対して変化させた実験をおこないたいと考えている．また，古文書画像において，レイアウトを認識するルール，及びその実現する手法について考察した．今後このレイアウト認識の実験もおこないたいと考えている．

5　古文書文字認識

従来の文字認識過程には，つぎのような特徴がある．

1. 切出しから認識までが順次処理される
2. 辞書への正規化では失われる情報がある
3. 文字サイズ・意味カテゴリーなどをパラメタにした辞書検索をおこなっていない

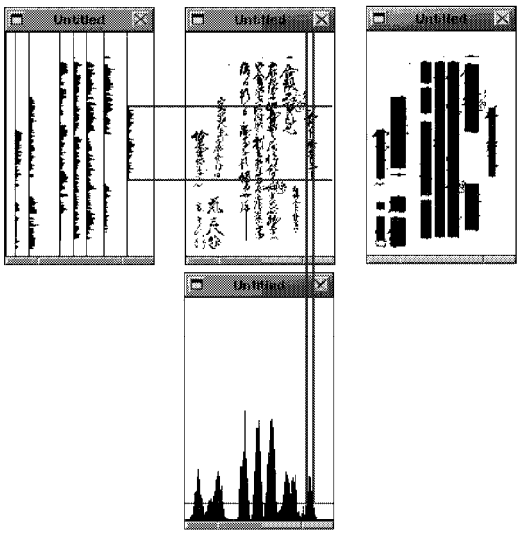

図 5: ヒストグラムによる抽出範囲選択

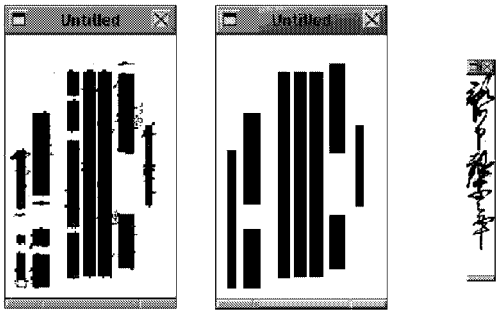

図 6: 文字列の抽出個所及び標題抽出結果

4. 通常は、認識過程の終了後の後処理で整合性がチェックされる

こうした従来型の認識プロセスにおいて，人間の文字認識プロセスに近いモデル化が可能かどうかを検討した [4]．具体的には，

1. 各文字パターンのサイズなどの特徴が失われない方法
2. 辞書検索時にサイズ等のパラメタが指定できる
3. 後処理から認識へバックトラックする機能
4. 文字切出しと認識の同時処理がおこなわれる方法

などを検討する必要がある．

以下に示す文字認識の実験では，上記の 1,4 について実現した．正規化は，認識しようとする対象画像に対して，文字パターン辞書から取り出されたパターンを対象画像のサイズに一致するように変換することである．われわれは，従来の認識プロセスとはまったく逆の発想で検討した．

まず，2-gram を用いた切出し，及び認識プロセスについて検討した．

1. 標題の先頭文字に出現する文字カテゴリーに含まれる 1 文字パターンを辞書から取り出す．
2. つぎに対象画像の文字幅を、辞書から取り出した文字パターン幅に変換する．すなわち正規化する．
3. つぎにマッチングに移行する．マッチングは重ね合わせ法によるが，隣接文字の「侵入」や「連結」を切出すためにマッチングをおこなう範囲を限定しなければならない．このために，マスク処理をおこなう．
4. 対象画像上での探索範囲は，おおむね経験則から文字パターンの高さの 2 倍としている．
5. マッチングにより，両パターンの距離が一定のしきい値以下になったとき，一致したとみなす．
6. 一致したパターンで対象画像のパターンを消去し，これがつぎの対象画像となる．

以上があらたな試みの認識プロセスの概要である．この実験結果から，2-gram を用いて切出し・認識をおこなった場合，約 90 %の認識率を得た（図 7,8）．この方式は，従来の人間の動作に比較してより近いのではないかと考えている。

このほかにもわれわれは，非線形正規化によりすくない文字サンプルから多様な文字サンプルを生成する手法についても研究を進めている．また HCD1 を対象とした自己想起型ニューラルネットを使った古文書文字認識で，未知パターンに対する平均認識率 99.06 %を達成している [5].

6 知識による翻刻支援

翻刻時に遭遇する読めない文字（不明文字）の前後文字から n-gram の情報を使って不明文字の正解候補

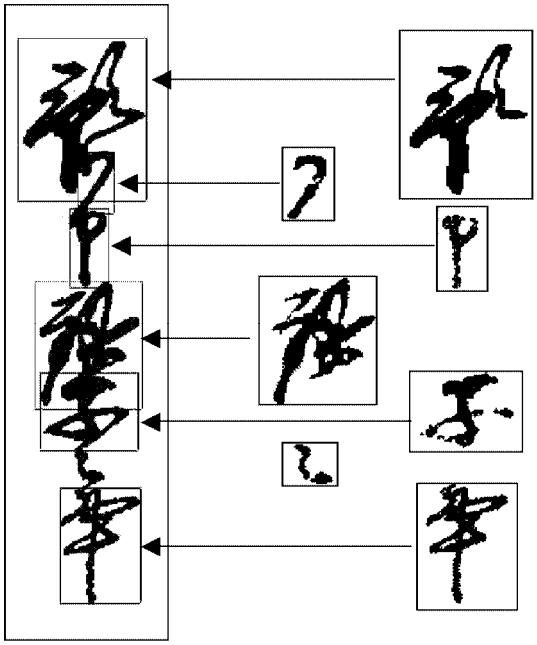

図 7: 切り出し・認識結果の例

を提示する可能性について検討した [6]．用例データとして「伏見屋文書」を使用し，翻刻支援手法の検討と検証をおこなった．その結果，前後の既知文字から 3-gram および 2-gram の情報を使って不明文字の正解を検索する実験により，第 10 候補までで 72.70 %の正解率を得られると推定できた．

本手法を Microsoft Word のマクロとして実装し，GetAMoji マクロの名称で公開している（図 9）．翻刻文を Word に呼び出し，GetAMoji を実行すると「□」文字の部分の正解候補が提示される．GetAMoji の利用試験をおこなったところ，翻刻経験のない初心者が辞書なしで翻刻した結果の正解文字数が有意に増加することがわかり，システムの有効性が確かめられた．

GetAMoji には「伏見屋文書」から作成した近世借金証文用辞書がサンプル辞書として付いているが，利用者が翻刻文の Word ファイルから，自分の辞書を作成する機能も持っている．

本手法は，不明文字の前後の文字が正しいと仮定して，その情報から不明文字の候補を提示するものである．したがって，前後の文字がそもそも誤っていたり，文字数の推定が誤っていたり，不明文字が連続してしまった場合には，正しい候補文字の提示ができない．本手法の応用として，英文のスペルチェックに対応す

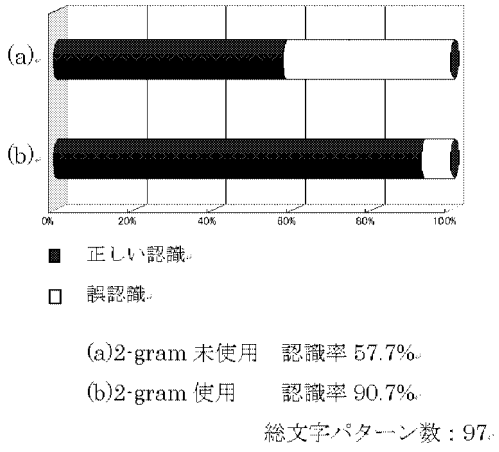

図 8: 切り出し・認識結果

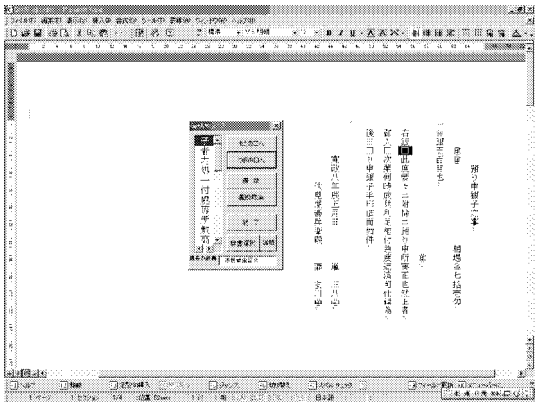

図 9: GetAMoji マクロ

るような，翻刻済み文字に対する検証システムのようなものも考えられるだろう．また本手法は，証文類という一定の表現が頻出するパターンをとる文字列に対して有効な手法であって，その他の種類の文書対してこの手法がどの程度有効であるかは今後の検討が必要である．

7　電子化古文書文字辞典

翻刻者が古文書を翻刻する際には，古文書文字辞典を参照しながら作業を進める．古文書翻刻作業に使われている標準的な辞典のひとつである『毛筆版くずし字解読辞典』 [2] は，文字の第 1 ストロークの方向から検索できるという，ほかの辞典にみられない特長を

有している．しかしながら紙ベースの辞典では，その検索の利便性はかならずしもたかいとはいえない．

われわれは古文書文字データベース作成作業において同辞典をディジタル化している．そこで同時点のディジタル情報を使って，紙の辞典よりも検索性をたかめた電子化古文書文字辞典の開発を進めている．電子化古文書文字辞典では，従来の「漢字」や「読み」からの文字検索に加えて，文字の外形や運筆からの検索を可能にする．

現在，それらの機能を実現するためのデータ作成法や検索アルゴリズムの基礎研究を実施している．将来的には，電子手帳のような携帯型のツールに電子化古文書文字辞典を搭載することを目指している．

8　おわりに

平成 11 年度より 3ヶ年の予定で開始した「古文書翻刻支援システム開発（ＨＣＲ）プロジェクト」の現況について報告した．現在までのところ，古文書文字データベース，古文書用例データベース，および知識による翻刻支援システムについて研究成果を公開するにまで到っている．古文書文字切り出し，古文書文字認識，電子化古文書文字辞典についてもデータを整備と平行して基礎的研究を進めている．しかしながら，古文書文字データベース化を進めている「伏見屋文書」の文字総数が膨大な数にのぼるため，文字の切り出し作業およびノイズ処理作業は難航している．科学研究費が終了する平成 13 年度はひとつの区切りの年となるため，今後の研究の発展に結びつけうる土台作りが年度内に完成できるよう努力している．

HCR プロジェクトのホームページは，

http://www.nichibun.ac.jp/~shoji/hcr/

である．最新の研究成果報告や本報告で述べた成果物の公開は，当ホームページからおこなっている．

謝辞

本研究は，日本学術振興会科学研究費補助金・基盤研究 (B)(1) 一般研究「古文書解読プロセスの知能情報学的解明」（平成 11〜13 年度，研究代表者：山田奨治），同展開研究「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」（平成 11〜13 年度，研究代表者：山田奨治），同一般研究「古文書ＯＣＲの試論的研究」（平成 11〜13 年度，研究代表者：柴山守），同展開研究「古文書解読支援システムの開発と電子辞書技術の応用に関する研究」（平成 12〜14 年度，研究代表者：柴山守）の支援を得て実施しているものである．また「伏見屋文書」の文字切り出し作業に関して，(財）元興寺文化財研究所のご助力を得ている．

参考文献

[1] 山田奨治, 加藤寧, 川口洋, 原正一郎, 石谷康人, 柴山守, 笠谷和比古, 小島正美, 梅田三千雄, 山本和彦：古文書翻刻支援システム開発プロジェクト報告（１）－プロジェクト概要－, 情報処理学会研究報告, Vol. 2000, No. 8, pp. 1–8 (2000).

[2] 児玉幸多編：毛筆版くずし字解読辞典, 東京堂出版, 東京 (1999).

[3] 尾崎浩司, 柴山守, 荒木義彦, 山田奨治：古文書画像の標題文字セグメンテーション, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol. 2000, No. 17, pp. 279–286 (2000).

[4] 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－, 京都大学大型計算機センター第 67 回研究セミナー報告, pp. 70–79 (2001).

[5] 橋本智広, 横田宏, 梅田三千雄：自己想起型ニューラルネットによる古文書文字認識, 平成 12 年度電気関係学会関西支部連合大会 (2000).

[6] 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, 情報処理学会シンポジウムシリーズ, Vol. 2000, No. 17, pp. 185–192 (2000).