古文書を対象にした文字認識の研究

山田 奨治
国際日本文化研究センター研究部
shoji@nichibun.ac.jp

柴山 守
大阪市立大学学術情報総合センター
sibayama@media.osaka-cu.ac.jp

日本語の文字認識技術を，100年以上も前に書かれた文書に対して適用できるのだろうか．古文書の文字には毛筆・つづけ字・くずし字という特徴がみられ，これに起因するさまざまな克服すべき技術的な課題がある．現在までに，研究を進めるためのいくつかの文字データベースが作成されており，あらかじめ完全に切り出されてある文字を使っての認識実験では，限定されたサンプルについてではあるが，良好な結果が得られている．しかし，古文書の文字認識を実用に近づけるには，文字認識プロセスの再検討が必要であることも分かってきた．また，筆者らは，翻刻作業者を支援するための種々のツールの開発も進めている．本稿では，古文書の文字認識研究の意義と現状，これまでの成果，課題について述べる．

◎問題の所在

筆者らは，日本語手書き文字認識を発展的に応用する研究として，古文書を対象にした文字認識の研究，およびそれを可能にするための環境の整備，既存の技術を活用した古文書の翻刻（古文書を読んで活字にすること）支援のシステム化の研究などに取り組んでいる．

古文書とは，狭い意味では差出人がある意思伝達の書類のことであるが，本稿でいう古文書は，他者への意思伝達を目的としない「古記録」や「古典籍」も含めた，広い意味で捉えることにする．時代でいうならば，古代から明治の初期くらいまでのあいだに作成された文書を，古文書と呼ぶことにする．古代から中世までに作成されて現代に伝わっている文書数は，約25万通といわれているが，これに近世のものを加えると古文書は無数にあるといってよい．

これらの古文書の多くは，各地の文書館などに収集され保管されているが，その量があまりに膨大なため，どのような古文書をどれだけ所有しているのかを把握すらできていない文書館もある．ましてや，それらのすべてを翻刻し，あるいは電子化して，歴史研究の史料として利用できるかたちにするまでには，膨大な労力と時間が必要なのが現状である．

古文書に書かれた文字の特徴は，第1にそれらの多くは毛筆で書かれていること，第2につづけ字が多いこと，第3にくずし字が多いことの3点に集約される．もちろん，古文書の様態は書かれた時代や種類によってさまざまであるから，すべての古文書がこれらの特徴を持っているわけではない．古文書のなかにも活字印刷に近いような，読みやすい木版の印刷物もある．しかし，未翻刻のものが圧倒的に多い近世の文書に限っていえば，おおよそ上記のような特徴を持っているといってよいだろう．

第1の毛筆であるという特徴によって，文字を構成する線の「かすれ」や「つぶれ」，運筆による線の濃淡が生じる．とくに線が「かすれ」たり「つぶれ」たりすることは，文字認識の処理を施すうえで重大な問題になる．

図-1
古文書の文字（かすれ，つぶれ，つづけ字，くずし字が同時に現れる例）

第2のつづけ字であるという特徴によって，これまでの日本語手書き文字認識の技術を応用するためには，つづけ字のなかから1文字を切り出す必要が生じてくる．これが第3のくずし字であるという特徴と重なって，文字切り出しだけをとっても容易に解決できない難問が，古文書の文字認識の前に立ちはだかっている（図-1）．

しかしながら，このようにたいへん困難に思える古文書の文字認識にも，研究に着手するためのいくつかの手がかりがある．まず対象とする文書の年代についていえば，未翻刻の文書の多さを考えれば江戸時代の近世文書にターゲットを絞ってよいだろう．近世に書かれた文書にも，江戸幕府の公式記録から個人の日記まで，さまざまなものがある．筆者らは，歴史研究での重要性を勘案して，公的な記録文書を対象にしている．この種の文書は，毛筆書きされたものがほとんどである．おそらく技術的な容易さからいえば，木版刷りの板本を対象にした方がよい成果を期待できるだろう．しかし，筆者らはあえて困難な毛筆手書きの文書の文字認識に挑戦している．

近世の公的な記録は「御家流」と呼ばれるくずし字によって書かれてある．つまり，文字のくずしの作法にはある程度の統一性がある．さらに，文書の種類によっては定型文が頻出する．たとえば，借金証文の場合ならば「申候處実正也」（もうしそうろうところじっしょうなり）といった語句がよく使用され，本文の最後は必ず「依而如件」（よってくだんのごとし）で結ばれる．用紙のどのあたりにどのような情報が書かれているかのレイアウトも，文書の種類によってははっきりとした構造を持っている．

これらのことを手がかりに，古文書の文字認識という遠大な研究に対してどのように取り組んでいるのかを，以下にご紹介したい．

図-2
「伏見屋善兵衛文書」の一部（大阪市立大学所蔵）

◎研究の流れ

幸か不幸か，古文書を対象にした文字認識の先行研究は少ない．1995年に著者のひとりの山田は，高次局所自己相関特徴を使って変体かな（明治初期まで使用されていた，現在のものとは違うひらがな）を文字認識させる試みを行った[1]．また，1998年に帝塚山大学の川口洋らのグループは，古文書から採字した漢数字を試験データにして，ニューラルネットによる文字認識を試みた[2]．

これらの個別の先駆的な研究を受けて，1998年8月に国際日本文化研究センターにおいて「第1回古文書OCR（自動読み取り）シンポジウム」が開催された[2]．同シンポジウムには日本史・古文書学研究者，手書き文字認識研究者ら約60名が参加し，(1) 歴史研究者からみた古文書の文字認識への期待，(2) 古文書の文字認識研究の現況，(3) 日本語手書き文字認識の最先端技術の3つのテーマについて討議が行われ，当面の研究方略として次の4点推進することがまとめられた．

(1) 対象の選択において，書体の安定した公文書で歴史的な価値の高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) 古文書の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

その後，同シンポジウムの参加者ら中心になってHCR（Historical Character Recognition）プロジェクトを結成し，1999年度から科学研究費を獲得して，上記の研究方略に沿って研究を推進している[3]．HCRプロジェクトのこれまでの個別の成果について，(1) 古文書文字データベースの作成，(2) 古文書の文字切り出しと文字認識の研究，(3) 知識による翻刻の支援，(4) 電子

名称	内容	採字元	字種	文字数	画像
HCD1	年齢表記文字	宗門改帳	16	3,066	2値
HCD1a	単位表記文字	宗門改帳	16	3,200	2値
HCD1b	単位表記文字	宗門改帳	8	1,600	2値
HCD1c	親族関係表記文字	宗門改帳	8	1,600	2値
HCD2	借金証文標題行	伏見屋文書	200行	1,378	2値
HCD2a	借金証文標題行	伏見屋文書	200行	1,378	256階調
HCD2b	借金証文標題行	伏見屋文書	200行	1,378	24bitカラー
HCD3	借金証文標題文字	伏見屋文書	183	4,933	2値

表-1
古文書文字データベースHCDシリーズ（2002年8月現在）

くずし字辞典の開発の各項目を取り上げて，以下に述べたい．

◎古文書文字データベースの作成

古文書文字データベースは，文字認識研究の基礎データとしての便宜を考えて，(1) 同一字種について書き手の異なる多くのサンプルが得られるもの，(2) 定型的な表現が頻出するものの2種類の観点から作成している．(1) の観点から作成しているものは，江戸時代の戸籍の一種である「宗門改帳」（しゅうもんあらためちょう）から漢数字や単位表記文字などを抽出したもので，HCD1 (Historical Character Database 1) と称している．HCD1はHCRプロジェクトの一員である，帝塚山大学の川口洋が作成している．(2) の観点からは，江戸時代の借金証文を中心にした文書群で大阪市立大学が所蔵する「伏見屋善兵衛文書」（図-2，以後「伏見屋文書」）の全文字（約243,000文字）をデータベース化するべく作業を進めている．「伏見屋文書」から作成した文字切り出し研究用データベースをHCD2，標題文字データベースをHCD3とそれぞれ称している．「伏見屋文書」のデータベース化は，大阪市立大学と国際日本文化研究センターにおいて行われている．

表-1は，HCD1から3の概要である．HCD1に収録されている画像の一部を，図-3に示した．表-1のデータベースはすべて，HCRプロジェクトのウェブページ[3]からダウンロードすることができる．

◎古文書の文字切り出しと文字認識の研究

仮に，古文書から文字が完全に切り出されていたとしたならば，従来型の日本語手書き文字認識技術で，どの程度の認識率が得られるのだろうか．HCRプロジェクトの一員である東北大学の加藤寧と和泉勇治らは，改良型方向線素特徴量とWeight Decayを適用したニューラルネットワークにより，HCD1データベースに対して96.67%の認識率が出ることを報告した[4]．また，おなじくHCRプロジェクト一員である大阪電気通信大学の梅田三千雄らは，自己想起型ニューラルネットワークにより，HCD1に対して平均99.06%の認識率が得られることを報告した[5]．これらの結果は，限定されたデータベースによるものではあるが，古文書の文字認識の研究を進めることに希望を持たせるものである．

古文書の文字認識のためのデータベースを作るに当たっては，採りたい文字を必ず見つけ出すことができるとは限らないため，1字種について多くのサンプルを実際の古文書から採字することには，たいへんな困難が伴う．そこで，少ないサンプルから多くのサンプルを人工的に作成する必要が生じる．HCRプロジェクトでは，この問題に関しても検討を進めている．

古文書にはつづけ字ではなく，文字をはっきりと分けて書いてあるものも少なくないが，古文書に多くみられるようなつづけ字を認識させようとした場合には，文字認識の前処理として正確な文字切り出しが必要になる．つづけ字の古文書から文字を切り出す方法についても，筆者らは研究を進めているが，いまだ汎用性のある有効な方法は見つかっていない．

つづけ字やくずし字で記述された古文書を人間が解読する場合は，個別の文字の判別や認識とともに，ことばや文章の判別・認識を行って逐次に解読を進める．すなわち，文字・用語・文体の三位一体による解読が必要とされる．さらに文書の解読のためには，文書が記された時代背景や関連する知識が必要であることはいうまでもない．このうち，文字の識別に関する情報のみだけを取り上げても部首，扁，旁，画数，筆順，筆勢，筆圧と多くの要素が存在する．

一方，古文書の文字認識をしようとする場合，入力画像から文字や語の画像特徴を抽出し文字認識へと進

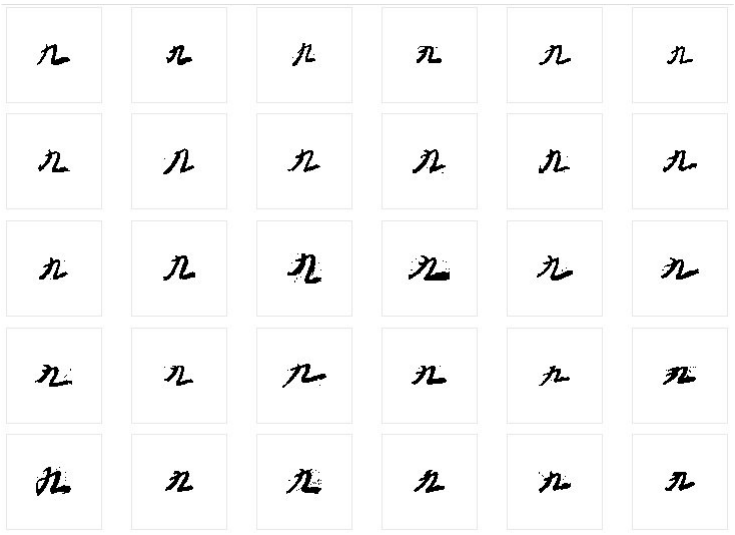

図-3
古文書文字データベースHCD1収録文字の一部「九」
（川口洋氏作成）

める必要があるが，そこでは形状，線，エッジ等による画像からの限られた特徴を基本に考えねばならない．これは，前述した人間が行う文書の認識プロセスとは大きく異なる．さらに，古文書の文字認識を従来型の文字認識モデルに当てはめると，文字の切り出し，正規化などの過程で，前述の人間が解読の際に用いる情報のいくつかが失われると想定される．たとえば，正規化によって文字の形状が変形し，極端な場合，本来の文字とはまったく異なった他の字形との類似度が増すことも分かってきた．

従来の文字認識プロセスでは，(1) 切り出しから認識までが順次処理される，(2) 正規化で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ータにした辞書検索を行っていない，(4) 通常は文字認識の後処理で整合性がチェックされる．筆者らはこうした従来型の文字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ている．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ータが指定できる方法
(3) 後処理から認識へバックトラックする機能
(4) 文字切り出しと認識の同時処理が行われる方法

などである．

人間が行う文字認識では，文字や語の前後関係，共起関係，およびそれらの背景などの知識を持って解読される．したがって，古文書の文字認識においても当然，これらの仕組みを反映させねばならないだろう．筆者らは，n-gramによる文字の共起関係の情報を文字認識にも導入し，実験をしている．また梅田らも文献5)の手法を発展的に応用して，文字認識処理と文字切り出し処理を連携させる方法を研究しており，古文書の文字認識に適した手法として今後の進展が期待できる[6]．

◎知識による翻刻の支援

古文書には多くの種類があるが，近世の借金証文類は様式が比較的一定しており，使用されている用語には定型がある．たとえば，「依而如件」「実正也」などの用語は必ずといってよいほど借金証文のなかに登場する．このように使用される用語に定型がみられる種類の文書については，多くの用例を集めてそこから用語に関する知識を抽出し，知識にしたがって翻刻者を支援する方法が考えられる．具体的な方法としては，n-gramを利用することの有効性が予想される．

筆者らは古文書の証文類を対象に，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を提示する可能性について検討した．証文類の用例データとするために「伏見屋文書」の全文情報を使用した．そして種々の検討の結果，「伏見屋文書」の場合には3-gramと2-gramを併用する方法が有効であることが分かり，その手法を実装した翻刻支援インタフェースを開発した（図-4）．開発し

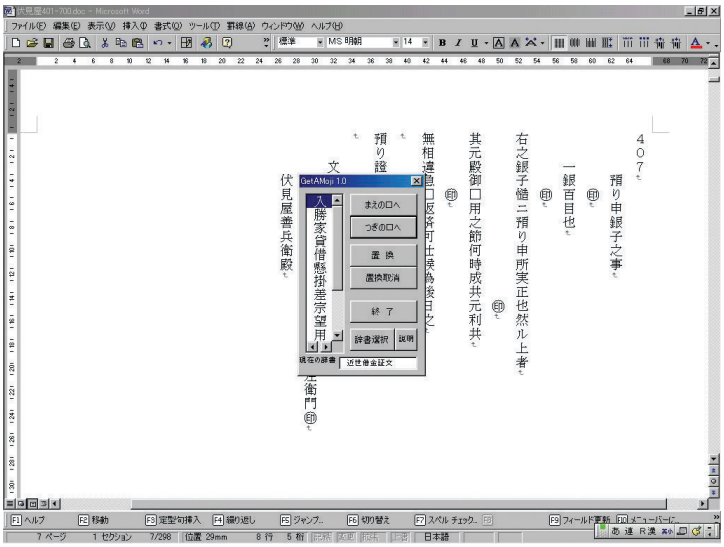

図-4
n-gramによる翻刻支援システム（GetAMoji）

たインタフェースの利用試験を行ったところ，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分かり，その有効性が確かめられた．

この手法は，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．またこの手法は，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．

◎電子くずし字辞典の開発

古文書の翻刻の際に使用される標準的な辞書を電子化し，検索の利便性を高めることができたならば，翻刻作業の大幅な向上が見込まれる．しかしながら，現在のところ，古文書の文字辞典類のなかで電子化されたものはない．辞書の電子化を考えるならば，現在最もよく使われている辞書を対象にするすることが理想である．専門的な翻刻者がよく使用している辞書の1つに，児玉幸多編『毛筆版くずし字解読辞典』(東京堂出版)(以後『くずし字辞典』) がある．しかしながら，この辞書を実際に使ってみると，求める文字にたどりつくにはそれなりの時間がかかり，検索漏れが起こる可能性も高いことが分かる．辞書を電子化して検索の方法を工夫すれば，知りたい文字にたどりつくまでの時間を短縮し，検索漏れを少なくすることができるだろう．

また，辞書を電子化することによって，紙の辞書では到底できない検索方法を実現することができる．それは，ある文字に類似した文字を一覧的に検索することである．類似文字の検索を実現する際に鍵となるのは，文字の特徴量と文字間の類似度の設定方法である．くずし字の特徴量と類似度は，日本語手書き文字認識技術で使用されている手法を応用することによって求めることができる．

『くずし字辞典』の特徴は，第1ストロークの方向によって文字を分類している点にある．すなわち第1画を，(1) 下に向かって連続する点で起筆する「縦点」，(2) 右に向かって連続する点で起筆する「横点」，(3) 右上から左下へ斜めに伸ばす棒で起筆する「斜棒」，(4) 上から下へ伸ばす棒で起筆する「縦棒」，(5) 左から右へ伸ばす棒で起筆する「横棒」の5種類に分けて，その種類ごとに文字が掲載されている．したがって，第1ストロークがおなじ文字間についてのみ文字の類似度を計算することによって，第1ストローク情報による候補の絞り込みができる．いわば，オフライン文字認識手法に，第1ストロークというオンライン的な要素を取り入れた手法といえるだろう．

以上のようなアイディアのもとに，『くずし字辞典』を出版社の許諾を得て電子化し，類似文字検索機能を開発して，電子くずし字辞典を実装した．この電子辞典は，『くずし字辞典』から採字した23,703文字の画像

図-5
電子くずし字辞典（eKuzushi）

とテキスト情報，そしておなじ第1ストロークからはじまるすべての文字間の類似度情報を持っている．現在のところ，検索の入口は文字コードのみとなっている．調べたい文字を日本語入力FEPで入力し，検索ボタンを押すと文字コードに該当するすべての文字画像が一覧で表示される．文字画像をダブルクリックすると，類似文字を検索することができる（図-5）．

この電子辞典は，現在のところワープロ的に入力された文字が最初の入口となっているため，分からない文字が何であるかのおおよその見当を，利用者がつけなくてはならない．将来的には，(1) タブレットなどで手書き入力された文字からの検索，(2) スキャナなどで画像入力された文字からの検索，(3) より深い階層のストローク情報からの検索機能を持たせるべく，研究を進めているところである．

◎今後の展望

以上，古文書の文字認識（HCR）に関する研究の概要について述べた．HCR研究はまだ緒についたばかりで，ようやく研究の基盤になる文字データベースがかたちを整えつつあり，文字切り出し，文字認識，翻刻支援についての問題の所在を明らかにすることと，初期の実験的な検討がなされている段階である．

古文書という現代人にとっては難解な文字と取り組んでいるうちに，次のようなことが実感されてきた．すなわち，古文書のつづけ字から文字を正確に切り出せるということは，その段階ですでに文字認識がなされていることを意味し，逆につづけ字のなかから文字を認識できるということは，その時点で文字切り出しが平行処理されているであろうということである．この文字切り出しが先か文字認識が先かというパラドックスを，なんらかの方法で解く必要がある．1つの解法としては，あらかじめ切り出された文字を前提にするのではなく，文字を認識しながら切り出すような手法を確立することであろう．

HCR研究には多くの技術的な課題が隠されており，また研究を進めることの文化的な意義はたいへん大きい．さらに，HCRの研究課題のなかには，つづけ字・くずし字の文字認識のように，日本語手書き文字認識の実用化にそのまま応用できる課題もある．これまでに作成したデータベースなどは，公開できるものはすべてHCRプロジェクトのウェブページにアップロードしてある．今後，より多くの研究者がHCR研究に参加してくれることを願っている．

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) 「挑戦　古文書OCR」特集号, 人文学と情報処理, No.18 (1998).
3) HCRプロジェクトのウェブページ
http://www.nichibun.ac.jp/~shoji/hcr/
4) 和泉勇治, 加藤　寧, 根元義章, 山田奨治, 柴山　守, 川口　洋: ニューラルネットワークを用いた古文書個別文字認識に関する一検討, 情報処理学会研究報告, Vol.2000, No.8, pp.9-15 (2000).
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) 橋本智広, 梅田三千雄: 天保郷帳における石高表記文字の個別認識, 情報処理学会研究報告, Vol.2002, No.8, pp.55-62 (2002).
（平成14年8月4日受付）