人文科学とコンピュータ57－1
（２００３．１．２４）

文字切出しを前提としない古文書標題認識

近藤 博人† 松本 隆一† 柴山 守†† 山田 奨治††† 荒木 義彦†
†立命館大学 ††大阪市立大学 †††国際日本文化研究センター

古文書画像を対象にした翻刻支援システムの構築を行っている。本稿では、文字認識の対象となる標題画像の射影ヒストグラムから推定した探索範囲に対して、文字パターン辞書から取り出した文字パターンを探索範囲内の最大文字幅で正規化しテンプレートとしてマッチングを行う、切り出しを前提としない認識手法について述べる。本手法を用いた実験では、近世の借金等証文類を中心にした『伏見屋善兵衛文書』（約 1,900 点、大阪市立大学所蔵）から 200 標題（及び、辞書に存在しない文字、又はサンプル数の少ない文字が含まれる標題を除く 151 標題）を対象として認識実験を行い、翻刻結果とする候補文字の抽出を行った。結果は、認識後の候補文字の抽出における認識率は、59.5%（69.7%）の結果であった。そこで設定に失敗した探索範囲を分析し、文字パターン辞書に含まれる特異な形状をもつ文字種に対する正規化、および先頭文字における適切な探索範囲を再設定する改良によって、候補文字の抽出においては 70.4%（83.1%）の結果が得られた。

Character Recognition without Segmentation for Title in Historical Document Images

Hirohito KONDO† Ryuichi MATSUMOTO† Mamoru SHIBAYAMA††
Shoji YAMADA††† Yoshihiko ARAKI†
†Ritsumeikan University ††Osaka City University
†††International Research Center for Japanese Studies

We have developed a transliteration assisting system which recognizes the character in the document written by calligraphic brush in the historical materials. This paper describes new recognizing scheme which tries to recognize the character without segmentation in the search area estimated from the projection histogram in a title image. A template image, which is a character pattern image extracted from the character pattern dictionary, before template-matching is normalized to be adjusted to a width of character pattern in the searching area after extracting from the dictionary. In an experiment for recognizing 200 titles（151 titles for eliminating them with few character patterns in the dictionary） in the Fushimiya Document, the recognizing rate was 59.5%(69.7%). Furthermore, in the experiment by improving the appropriate normalization for some characters with special shape, and the connection for joining divided searching areas at first character in title image, the result of the recognizing rate was 70.4%（83.1%）.

1.はじめに

全国にはいまだに解読されず、手付かず状態になっている古文書が多く存在する。これらの古文書を解読し、データベース化することによって、多くの人が必要な古文書を容易に参照・閲覧できるようになる。これは歴史学、人文学の研究分野だけでなく、古文書の解読に関心のある一般市民にも大いに役立つと考えられる。しかしながら、解読作業をすべて手作業で行うには、膨大な時間と費用、高度な専門知識を必要とする。そこで解読作業の自動化、もしくは作業者を支援するシステムの開発が求められている。

筆者らは、こうした古文書解読を支援するために、古文書翻刻支援システムの開発を行ってきた[1]。古文書翻刻支援システムの開発では、古文書がくずし字やつづけ字で書かれることから、従来の文字認識技術を用いることは難しい。これは認識を行うために、あらかじめ文字列からの文字切出しを前提としているためである。

そこで本研究では、従来の文字認識過程とは異なり、文字認識の対象となる標題画像の射影ヒストグラムから推定した探索範囲に対して、文字パターン辞書から取り出した文字パターンを探索範囲の文字幅で正規化しテンプレートとしてマッチングを行う、切り出しを前提としない認識手法について提案し、その有効性について検討する。

2.文字切出しを前提としない文字認識手法

2-1　従来の文字認識過程

従来用いられてきた一般的な認識過程を図１に示す。まず認識対象となる文字列に対して、ノイズ除去、スムージングなどの前処理を行う。次に文字列から各文字や語単位で文字を切出す。そして切出した文字を辞書の文字パターンに合せるように正規化し、認識を行う。

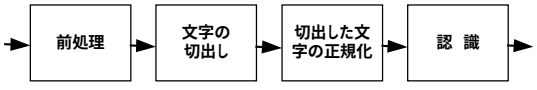


図１．従来の文字認識過程

従来の認識過程を古文書に適用させた場合、各文字や語が適切に切出せるかが問題になる。なぜなら図２のようなくずし字やつづけ字が多い文字列から切出しでは、良い結果が得られていない[1]。

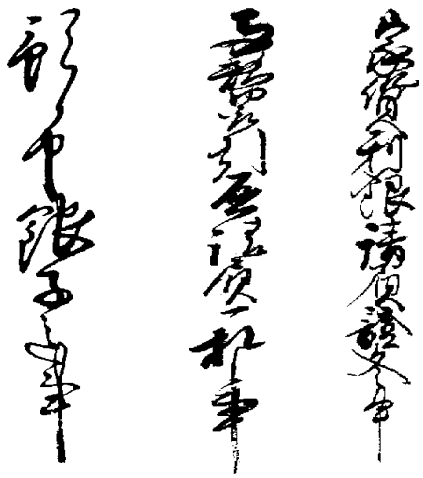


図２．古文書文字列

そのため切出した文字には、上下文字や他行からの接触や侵入などの影響によるノイズが含まれていることが多い。これらのノイズを除去できなければ、認識精度の低下につながることが予想される。

2-2　本手法の文字認識過程

本手法では、まず認識対象となる文字列に対して、用意した文字パターン群（以下文字パターン辞書という）とのマッチングを行う範囲（以下探索範囲という）を設定する。次に探索範囲内の文字とマッチングを行うために、文字パターン辞書から取り出した文字パターンを、探索範囲内の文字の大きさに合わせるように正規化する。そして探索範囲内で、文字パターンを左上から右下へと走査させながらマッチングを行う。

本手法では認識部の前に文字の切出し過程を必要としない、つまり前提としていないのが特徴である。また辞書から取り出した文字パターンを、探索範囲内を走査させながらマッチングを行うので、探索範囲に上下文字との接触や、他行からの

侵入などのノイズが含まれていてもマッチングの結果に影響を及ぼしにくい。

例えば図３において、探索範囲内には「預」と「り」の２文字が存在するが、文字パターン「預」を走査させた場合、探索範囲内の「預」の場所で最大のマッチング結果が得られる。つまりノイズの影響を受けにくいのが分かる。

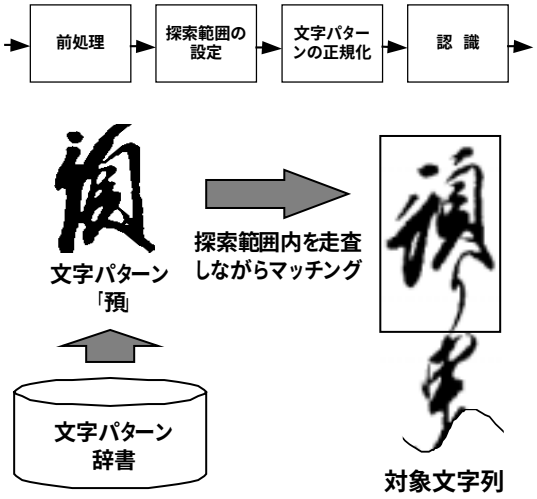


図３．本手法の文字認識過程

3.探索範囲と文字パターン辞書の正規化

3-1　ヒストグラム

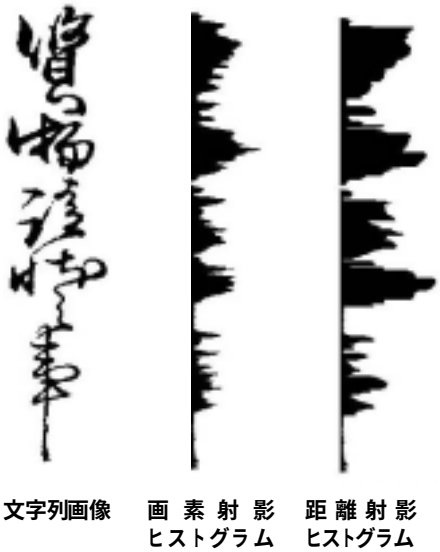


図４．ヒストグラム

探索範囲の設定にはヒストグラムを用いる。従来から文字列の特徴を把握するのに、水平方向画素値に基づく射影ヒストグラムが用いられる[2]。しかし毛筆のつづけ字の多い古文書では、ヒストグラムの起伏や切れ目が判断しづらく、特徴を把握しにくい。そこで、最左端の画素から最右端の画素までの距離（文字幅）をヒストグラム化することにより特徴が掴みやすい（図４）。

3-2　探索範囲の設定

まず文字列からストローク幅推定値[3]を求める。ストローク幅推定値というのは、文字列に含まれる線幅の推定値のことである。

次にこの値を閾値とし、ヒストグラムの閾値以下の部分を除去する。これにより、ヒストグラムをいくつかの塊に分割する事が出来る。そして、分割したヒストグラムの上端から下端までの範囲を探索範囲として設定する（図５）。

このとき、上端から下端までの距離が短い場合、つまりあまりに小さく分割されてしまったヒストグラムの塊は、ノイズとして無視する。

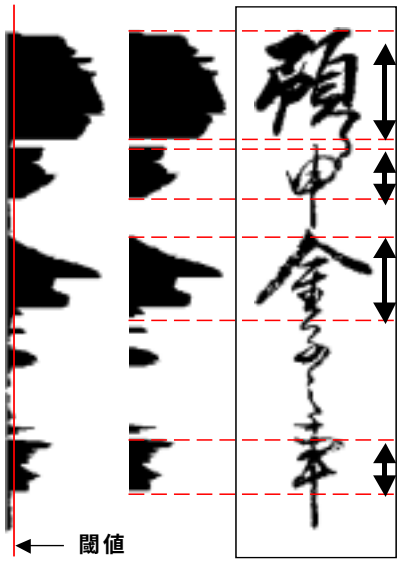


図５．探索範囲の設定

3-3　文字パターン辞書の正規化

設定した探索範囲内の文字と、文字パターン辞書の文字の大きさは異なる。そのため、文字パターン辞書の文字を探索範囲の文字の大きさに合う

ように正規化を行う。

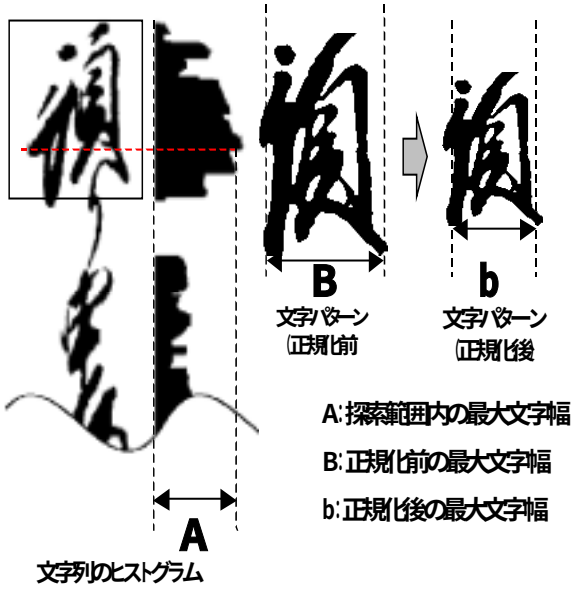


図６．文字パターン辞書の正規化

まず探索範囲内の文字から最大文字幅を検出する。次に文字パターンに対しても、同様に最大文字幅を求める。そして、探索範囲の最大文字幅と文字パターンの最大文字幅の長さが等しくなるように、文字パターン辞書を拡大、または縮小する（図６）。

4.候補文字の抽出実験

4-1　実験方法

本手法を用いた候補文字抽出実験行った。実験対象となる文字列は、「伏見屋善衛兵文書」[4]の200 標題とし、文字パターン辞書として 4420 個の文字パターン（143 文字種）を用意した。ともに「古文書翻刻支援システム開発プロジェクト」[5]のホームページで公開されており、標題画像は「HCD2」、文字パターン辞書は「HCD3」である（付表参照）。

マッチング手法はテンプレートマッチングとし、残差割合の小さい文字から順に、第 10 位まで候補文字として抽出する。そして探索範囲内の文字が、候補文字として抽出できれば正解とした。

4-2　実験結果

200 標題に含まれる総文字数 1378 に対して、設定できた探索範囲は 814 である。そしてこの探索範囲を対象とした候補文字の抽出では、59.5%の累積正解率が得られた（図７）。

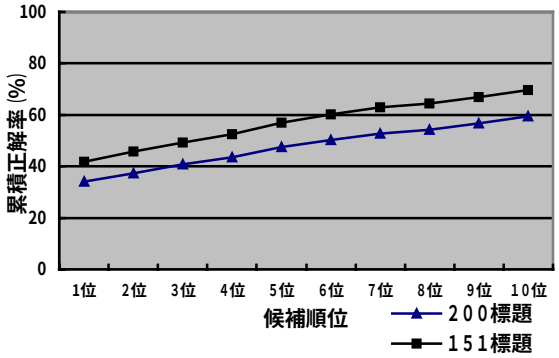


図７．候補順位別累積正解率

今回用意した標題文字列の中には、文字パターン辞書に存在しない文字や、サンプルの少ない文字が含まれており、その文字が探索範囲に設定される場合があった。そこで「辞書に存在しない文字」または「サンプル数の少ない文字」が、探索範囲として設定された 49 標題を除いた場合の結果についても述べる。これは今回マッチング手法として用いたテンプレートマッチングでは、ある程度のサンプル数が必要なためである。そこで 49 標題を除いた 151 標題を対象とした場合では、候補文字の抽出において 69.7%という正解率が得られた。

4-3　考察

今回の実験では、151 標題を対象とした場合でも 69.7%という正解率しか得られなかった。これは設定した探索範囲の中に、文字の一部分がはみ出しているものや、ひとつの文字に対して複数の探索範囲を設定してしまったもの、また全く文字を含んでいない探索範囲が存在するために、マッチングの精度が低下してしまったからである。

これらの設定に失敗した探索範囲は、図８のように文字の上側が外れるパターン(a)、文字の下側

が外れるパターン(b)、文字の上下両側が外れるパターン(c)、そしてそれ以外のその他のパターン(d)に分類できる。

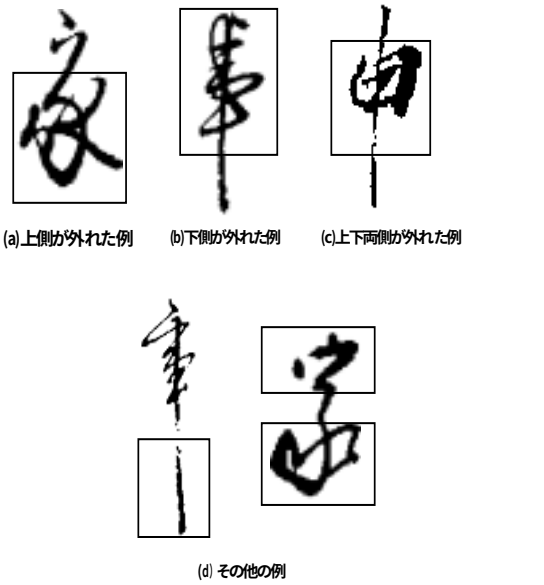
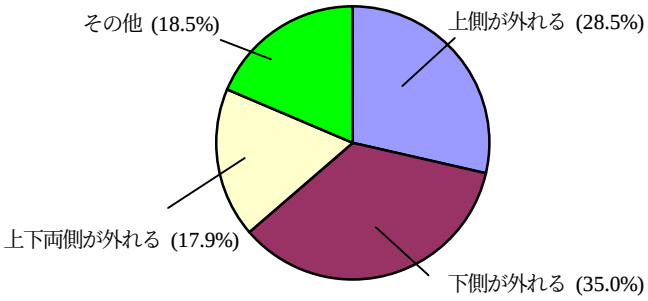
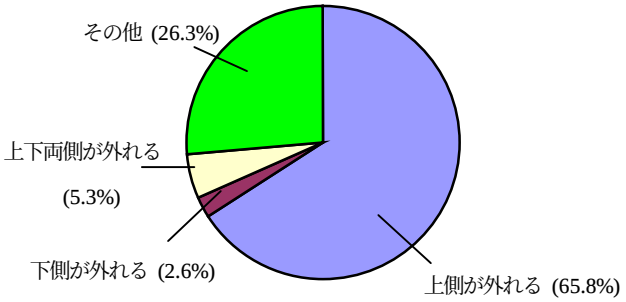


図８．設定に失敗した探索範囲



（a）すべての探索範囲



（b）先頭の探索範囲のみ

図９．設定に失敗した探索範囲の要因

(a),(b),(c)のパターンは、探索範囲設定において、標題文字列から求めたストローク幅推定値を閾値として用いたため、文字の縦線のみが現れる部分でヒストグラムが分割されてしまうのが原因である。このような例は、「事」や「申」のような文字に起こりやすい。そして(a),(b),(c)のようなパターンは、設定に失敗した探索範囲の 81.4%を占めている（図９(a)）。

そこでこの問題を解決するために、あらかじめ辞書内の文字パターンに対して、上下のストローク幅を切除するという前処理を行う。この処理によって、たとえ文字の一部分がはみ出ている探索範囲であっても、候補文字として抽出できるのではないかと考えられる。

次に先頭の探索範囲に注目した時、探索範囲設定に失敗した場合は図８(a)のパターンであることが多い（図９(b)）。これは「家」、「永」、「座」、「親」などの書き出しの点が孤立するために、探索範囲の設定に失敗しやすい文字が、先頭文字となる標題がいくつか存在するからである（図 10）。

そこで一番上の探索範囲に限り、探索範囲を上方に拡張する処理を行う。

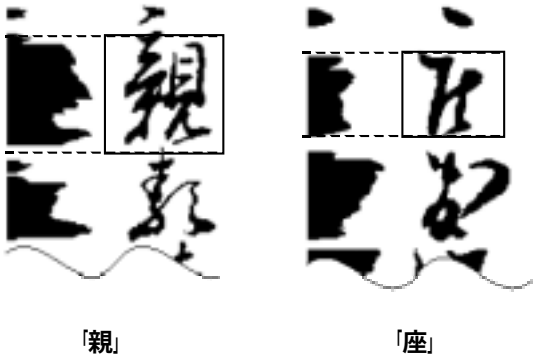


図 10．先頭探索範囲設定の失敗例

5.探索範囲の拡張と文字パターンに対するストローク切除

5-1　先頭探索範囲の拡張

先頭の探索範囲に限り、範囲を上方へ拡張する。探索範囲の上側に、探索範囲の設定時にノイズと

みなされたヒストグラムが存在する場合、そのヒストグラムの上端までを、新たな探索範囲として設定する（図 11)。

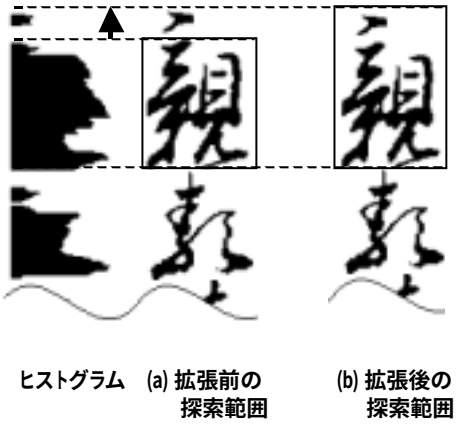


図 11. 先頭探索範囲の拡張

5-2 文字パターンに対するストローク切除

辞書内の文字パターンに対して、文字幅のヒストグラムを求める。次にその文字パターンのストローク幅推定値を求め、閾値とする。そしてヒストグラムを上下双方から走査し、はじめて閾値に達する場所までを切除する（図 12)。辞書内のすべての文字パターンに対して同様の処理を行う（図 13)。

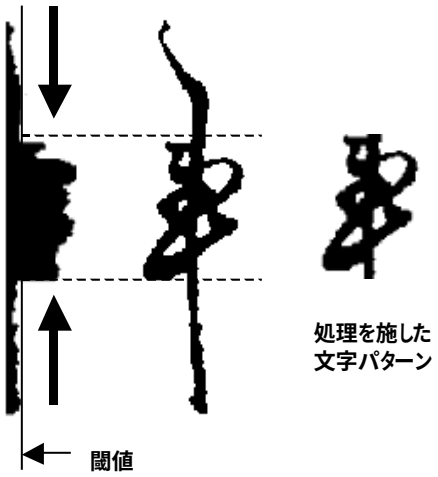


図 12. 上下部分のストローク切除

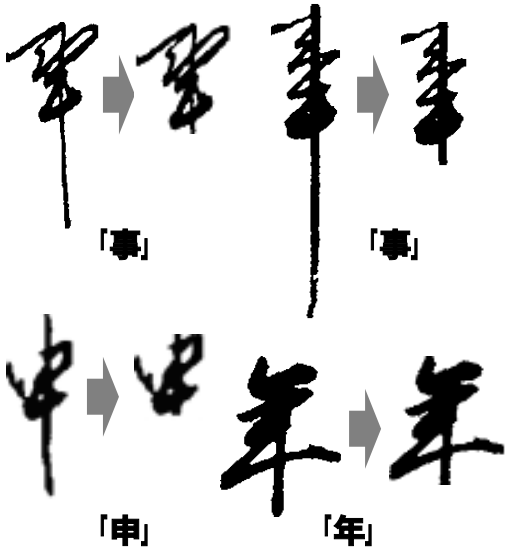


図 13. 切除後の文字パターン

5-3 再実験

探索範囲の拡張と、文字パターンに対するストローク切除の前処理を行ったうえで、再度同様の実験を行った。そして前処理を行った場合（処理あり）と、行わなかった場合（処理なし）の実験結果を比較する。

４章の実験では、設定したすべての探索範囲に対して候補文字の抽出を行った。しかし本稿では図７(a),(b),(c)の失敗パターンを対象として、正解率を向上させるために前処理を行った。そこで今回の実験では、(d)のパターンについては候補文字抽出の対象としないこととした。

まず前処理の有効性を確かめるために、図８(a),(b),(c)の失敗パターンのみを対象とした場合の、処理の有無による抽出成功数を表１に示す。

設定に失敗した探索範囲であっても、200 標題で 44、151 標題で 37 の探索範囲について、新たに正解候補を抽出する事が出来た。

表１．処理の有無による抽出成功数

	対象探索範囲	処理あり	処理なし
200 標題	277	123	167
151 標題	210	106	143

表２．正しく設定できた探索範囲の抽出成功数

	対象探索範囲	処理あり	処理なし
200 標題	474	361	363
151 標題	340	310	314

文字パターン辞書に対して前処理を行うことにより、少なからず字形が崩れることになる。これにより、正しく設定された探索範囲の抽出成功数が低下するのではないかと思われる。そこで正しく設定された探索範囲を対象とした場合の処理あり、処理なしの抽出成功数を表２に示す。

処理の有無でほとんど結果が変わらず、悪影響を与えるどころか、微数ながらも抽出成功数が増加しているのが分かる。これらの結果から、本手法を用いた文字認識において、今回行った前処理が有効である事が分かる。

最後に今回行った実験による、処理の有無による累積抽出成功数を表３に示す。

表３．処理の有無による累積抽出成功数

		総文字数	探索範囲数	対象探索範囲数	抽出成功数
200 標題	処理なし	1378	814	751	484
	処理あり				529
151 標題	処理なし	1054	597	550	416
	処理あり				457

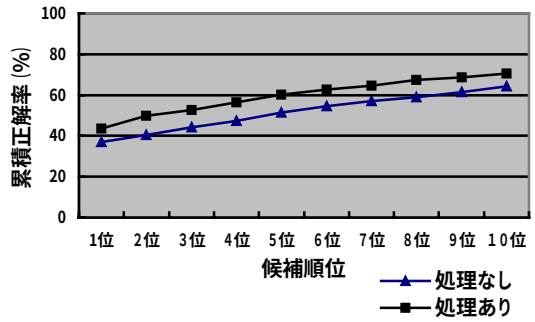


図 14. 処理の有無による候補順位別累積正解率
（200 標題）

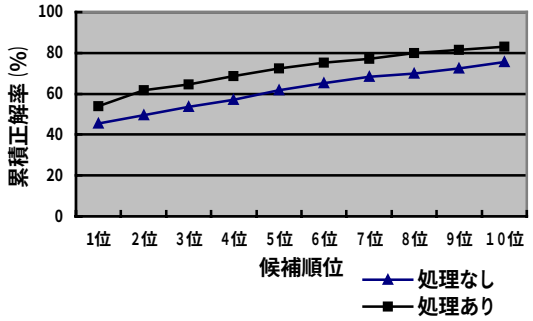


図 15. 処理の有無による候補順位別累積正解率
（151 標題）

第 10 候補までの累積正解率では、200 標題の場合で 70.4%、151 標題の場合で 83.1%の結果が得られた（図 14,15）。どの候補順位においても、処理ありの方が良い結果が得られているのが分かる。そして第 10 候補まで結果では、処理を行うことにより 200 標題で 6.0%、151 標題で 7.5%正解率を向上させることが出来た。

6.おわりに

従来の文字認識過程と異なり、対象文字列からの文字切出しを前提としない文字認識手法を提案した。そして正解率低下の原因である探索範囲設定の失敗パターンを分析し、先頭探索範囲の拡張処理と、文字パターン辞書に対する上下のストローク幅切除という、前処理を行う事で正解率の向上を試みた。その結果 200 標題の場合で 6.0%、151 標題の場合で 7.5%累積正解率を向上させることが出来た。しかし図７(d)のパターンについては、今回改善を行えなかったので検討していく必要がある。さらに他の古文書文献に対しても、同様の実験を行って行きたいと考えている。

また更なる正解率の向上のためには、知識ベースの導入が有効であると思われる[6],[7],[8]。候補文字抽出の際や、抽出後の候補順位の入れ替えなどに知識ベースが利用できれば、処理時間の短縮や、正解率の向上が期待できる。

今後は正解率の向上を目指すだけでなく、GUI によるユーザインターフェースを作成し、対話型システムの検討を行いたいと考えている。

参考文献

[1] 柴山 守:「古文書の文字切出しを考える」、人文学と情報処理第 18 号 特集挑戦古文書 OCR、勉誠出版、pp.57-63、1998

[2] 尾崎浩司、柴山 守、荒木義彦:「古文書画像のレイアウト認識と標題抽出」、情報処理学会研究報告、2000-CH-47、Vol.2000、No.67、pp.47-54、2000

[3] 井野英文、猿田和樹、加藤 寧、根元義章:「ストローク情報に基づく手書き郵便宛名の切出しに関する一手法」、情報処理学会論文誌、Vol38、No.2、pp.280-288、1997

[4] 大阪市立大学学術情報総合センター所蔵、近世資料

[5] 古文書翻刻支援システム開発プロジェクト「HCR Project」
http://asagi1.nichibun.ac.jp/~shoji/hcr

[6] 笠谷和比古:「古文書における文字認識」、人文学と情報処理第 18 号 特集挑戦古文書 OCR、勉誠出版、pp.13-18、1998

[7] 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

[8] 尾崎浩司、柴山 守、山田奨治、荒木義彦:「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 2000 論文集、情報処理学会、2000

付表

名称	内容	採字元	画像
HCD1	年齢表記文字	宗門改帳	2値
HCD1a	単位表記文字	宗門改帳	2値
HCD1b	単位表記文字	宗門改帳	2値
HCD1c	親族関係表記文字	宗門改帳	2値
HCD2	借金証文標題行	伏見屋文書	2値
HCD2a	借金証文標題行	伏見屋文書	256 階調
HCD2b	借金証文標題行	伏見屋文書	24bit カラー
HCD3	借金証文標題文字	伏見屋文書	2値