人文科学とコンピュータ　53－8
（2002．1．26）

天保郷帳における石高表記文字の個別認識

橋本 智広　梅田 三千雄
大阪電気通信大学
〒572-8530 大阪府寝屋川市初町 18-8

あらまし:

本論文では，古文書文字列を対象として，古文書特有のつづけ字や食い込みに対処するために，認識処理を援用した文字切り出し手法を提案し，文字列からどの程度個別認識が可能かを検討する．まず，初期文字切り出しとして，連結するそれぞれの領域を矩形で囲み，高さや横幅等の矩形情報を基に各文字パターンを切り出す．そして，得られた文字パターンに対しニューラルネットワークを用いて個別文字認識する．ニューラルネットワークには自己想起型ネットワークを用いた．次に，最適な切り出し位置を得るために，認識処理を援用した再文字切り出しをする．最終的に，得られた切り出し候補に対して個別認識する．本手法において，「天保郷帳」を例にとった 615 個の文字列を対象とした個別文字認識では，総文字数 7987 文字に対する平均認識率は 94.87%が得られた．

キーワード:

文字認識，文字切り出し，自己想起型ニューラルネットワーク，古文書

Recognition of Kokudaka Characters Occurring in Local Tenpo Era Records of Rice Crops

Tomohiro HASHIMOTO , Michio UMEDA
Osaka Electro-Communication University,
18-8 Hatsu-cyo,Neyagawa-shi,Osaka 572-8530,Japan

Summary:

This paper proposes a character segmentation and recognition method of histrical documents. In the segmentation method, the result of character recognition process is utilized to cope with the cursive scripts and the mutual encroachment of characters which are peculiar to the historical documents. In the individual character recognition , autoassociative neural networks are used for flexibility and efficiency. From the recognition experiment applied to 7987 characters include 615 characters strings which are appeared in the local Tenpo era records of rice crops , the correct recognition rate of 94.87% was obtained by using the re-segmentation process.

Keywords:

character recognition , character segmentation , autoassociative neural network , histrical document

1　はじめに

手書き文字認識に関する研究は，様々な研究機関で試みられており，数多くの認識手法が提案され，その技術は実用の段階にある．一方，人文学研究分野では，古文書を対象としたOCRの実現を目指し，古文書に対する認識手法の提案が期待されている [1].

現在，古文書画像データベースの構築においては，史料の解読，文字データ入力に長時間の作業を必要とする．この作業を自動化できれば，飛躍的に作業

時間を短縮でき，多量の古文書史料を短時間で，効率よくデータベース化できる．そこで，古文書を対象としたOCRの研究が進められている[2]-[5].

しかし，古文書を認識対象とすると，文字のパターン数に制限があるため，認識に使用する辞書作成において，十分なデータ採取が困難となる．そのため，認識対象となる文字が限定されてしまう．従って，限られたデータの範囲内で，古文書独自の認識手法を新たに考案する必要がある．また古文書では，認識手法とともに，文字列から個々の文字パターンに切り出す文字切り出しが重要となる．しかし，古文書特有のつづけ字や文字の食い込み等から正確な文字切り出しが困難であり，それに伴い高い認識結果を得にくいなどの問題もある．

本論文では，毛筆で書かれた古文書特有のつづけ字や文字の食い込み等を考慮して，認識処理を援用した文字切り出し手法を提案し，その認識について検討する．

文字列から個々の文字パターンを切り出すために，まず初期文字切り出しとして，文字パターンにおける連結成分を囲む矩形情報を基に統合，分割処理を繰り返し，切り出し候補を得る．個別文字認識においては，特徴抽出に加重方向指数ヒストグラム特徴[6]を用いた．認識処理には，柔軟な情報処理と高い汎化能力を持ち，人間の学習過程をモデル化した，ニューラルネットワーク（以下NNと略す）を利用する．ここでは，古文書文字に対して有効とされる自己想起型NN[4]を用いた．さらに，認識処理結果から初期文字切り出しにおける切り出し失敗矩形を検出し，これらに対して認識処理を援用した再文字切り出しをする．認識処理を繰り返し実行することにより，最適な位置での切り出しが期待できる．そして，再切り出しによって得られた最終的な切り出し候補に対し個別文字認識する．

認識実験では，「天保郷帳」を例にとり，文字列からの個別認識がどの程度可能であるかを検討すると同時に，再切り出しの導入前後の認識率について検討する．

2 システムの概要

本システムの処理手順を図1に示す．まず，対象文字列において文字を個別パターン化するために初期文字切り出しをする．これによって，切り出された各々の文字パターンに対し個別文字認識する．ここでは，前処理，特徴抽出を施し，さらに学習処理によってNNを形成する．初期文字切り出しでは，誤って切り出された文字パターンが存在することが多い．そこで，該当する文字パターンに対し，認識処理を援用した再文字切り出しをする．最終的に，切り出された全ての文字パターンに対しNNを用いて個別文字認識する．

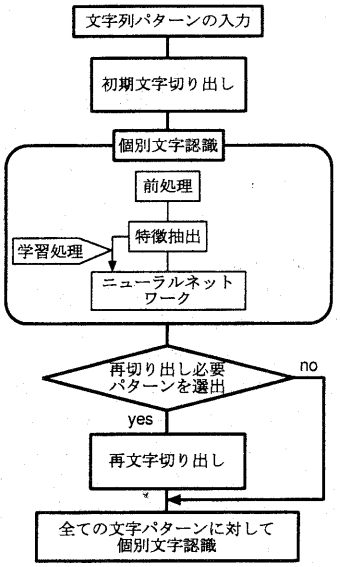


図 1: 処理の流れ

3 初期文字切り出し

文字列から各々の文字を認識するためには，個別に文字を切り出す必要がある．初期文字切り出しは，文字パターンにおける連結成分の高さ（$heigh$）や横幅（$width$），面積（$area$）等の情報に着目した切り出し手法である．

処理手順は，まず，ラベリングにより文字パターンの連結成分に対する外接矩形を求める．外接矩形とは，連結成分を囲む長方形のことである．この時点では，「八」や「三」のような複数の領域から構成される文字は，それぞれ連結性がないために各領域は独立している．そこで，各領域をグルーピングすることで一つの領域とするために統合処理を施す．

統合処理は，4段階に分けて行う．対象文字列が縦書きであるため，先に述べた「八」などの文字は横方向の矩形同士をグルーピングすることで一文字となる可能性が高い．図2に統合処理における各段階での条件を示す．第一段階として，図2(a)のように対象矩形の高さ内に別の矩形があるとき統合する．第二段階は，対象矩形と別の矩形が図2(b)の位置関係にあり，別の矩形の高さ比率が40％以上のとき統合する．この二つの条件は，矩形の高さに着目した統合方法であるが，第三段階以降は矩形同士の重な

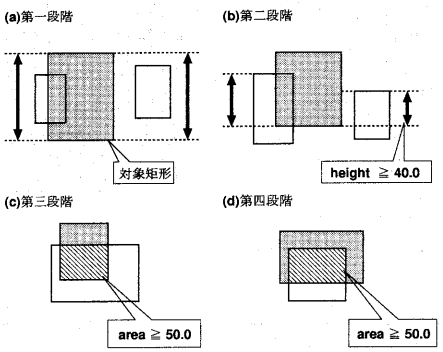


図 2: 統合方法

り合う面積に着目して統合する．第三段階では，図2(c) に示すように，対象矩形が別の矩形の幅よりも小さく，重なり合っている対象矩形の面積比率が 50 %以上で統合する．さらに，第四段階として，図2(d) に示すような別の矩形が対象矩形の横幅よりも大きく，重なりあっている別の矩形の面積比率が 50 %以上のときに統合する．この第四段階に関しては，次に説明する分割処理において，一回目の分割処理が終了した時点で適用する．これらの統合条件の他に，「三」を強制的に統合する条件も与えた．

次に，統合処理によってグルーピングされた矩形の中には誤って統合されたものや，初期の段階でつづけ字や食い込みにより，二文字以上が同一領域であるとみなされた矩形が存在する．そこで，これらの矩形に対し，分割処理を施す．

分割処理は，二度の処理で完了する．まず，それぞれの分割対象矩形を選出するための条件を次のように与える．一回目の条件は，矩形内に 3 個以上の複数文字が存在すると考えられる矩形を選出するものであり，二回目は，矩形内に 2 文字程度が存在すると考えられる矩形を選出する条件である．ここで，全矩形の面積の平均を $area.ave$ とし，全矩形の横幅の平均を $width.ave$ とする．

- 分割対象矩形条件：一回目
  - $height \geq width \times 1.2$
  - $area \geq area.ave$
- 分割対象矩形条件：二回目
  - $width \geq height$
  - $width \geq width.ave$
  - $area \geq area.ave$

これらの条件にあてはまる矩形を図 3 に示すように分割する．まず初めに，$height \div width$ を求め，この値を仮の矩形内文字数 $(moji)$ とする．そして，文字数分の仮の分割ライン (div) を決定する．さらに，仮の分割ラインから前後に div の 20 %の大きさをとり，その区間の射影分布を求める．そして，最小値となるところを最終的に分割ラインとする．二回目も条件は異なるが，同様の処理によって対象矩形を分割する．

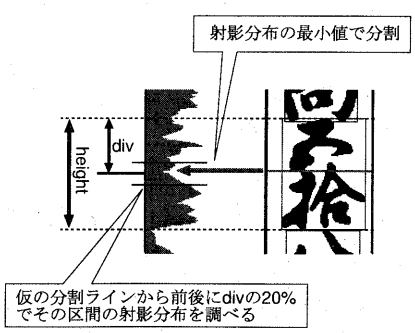


図 3: 分割方法

4 個別文字認識

文字切り出しによって切り出された各々の文字パターンに対し，個別文字認識する．まず，前処理として，孤立点除去，大きさの正規化，スムージングにより文字パターンの均一化を図る．次に，特徴抽出を行う．ここでは，数多く提案されている特徴抽出法のなかから，比較的高い認識率が期待できるとされている加重方向指数ヒストグラム特徴を用いることにした．さらに，得られた特徴量をもとに，NN を形成するための学習処理を行う．そして，形成された NN を用いて個別認識する．

4.1 前処理

切り出した各々の文字パターンは大きさにばらつきがある．そのため，各文字パターンの均一化を図るために前処理を施す．まず，画像に含まれている雑音を除去する孤立点除去，ばらつきのある大きさを均一にする大きさの正規化，さらに，大きさの正規化によって凹凸の激しくなった文字の輪郭部を平滑化するスムージングを施す．図 4 に前処理前後の文字パターン例を示す．

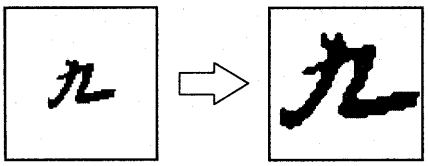


図 4: 前処理

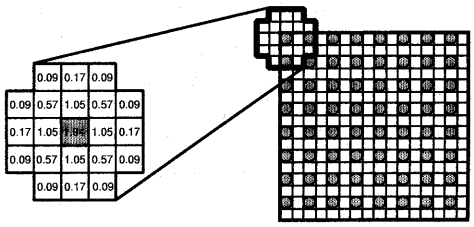


図 6: ガウスフィルタ

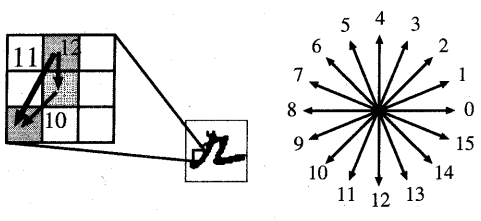
4.2 加重方向指数ヒストグラム特徴

文字の輪郭部に着目した特徴抽出法である加重方向指数ヒストグラム特徴は，次のようにして抽出する．まず，文字パターンに対し輪郭線追跡を行いながら，輪郭部に属する各画素について 16 の方向指数を算出する．方向指数の算出では，図 5 に示すように，注目画素と連結している前の画素から注目画素をみた方向指数と，注目画素から後の画素をみた方向指数から注目画素の方向指数を算出する．

この例では，前の画素から注目画素を見た方向指数は 12 であり，注目画素から後の画素をみた方向指数は 10 となる．そこで，両者の方向指数の平均をとることで注目画素の方向指数を 11 と算出する．

そして，方向指数を算出後，各方向指数に対して方向圧縮する．まず，奇数方向に重み付けし，これらに前後の偶数方向を足しこむ処理により，16 方向から 8 方向へと圧縮する．さらに，反対方向を同一視することにより，4 方向へと圧縮する．

次に，領域圧縮として，96×96 画素の領域について，16×16 領域に分割してヒストグラムを求める．さらに，図 6 に示すような，局所的な位置をぼかす働きをもつガウスフィルタを用いて領域圧縮する．ガウスフィルタは画素一つおきにフィルタリングする．これにより，8×8 領域 ×4 方向からなる 256 次元の特徴量を得る．



(a) (b)

図 5: 方向指数の算出

4.3 自己想起型ニューラルネットワーク

認識には，その柔軟で，かつ高い汎化能力から文字認識において利用されることが多い NN を使用する．ここでは，特に NN の中でも古文書に対して有効とされる自己想起型 NN を用いることとした．

自己想起型 NN は入力層と出力層のユニット数が等しく，入力パターンそのものを理想出力とするネットワークである．従って，教師信号には入力パターンそのものを与える．学習には，バックプロパゲーション法（BP 法）を用いて，各ユニット間の重みを変更していくことで NN を形成する．図 7 に，ここで使用したネットワーク構成を示す．各層のユニット数は，入力層と出力層が 256 個，中間層は 50 個とした．

文字認識においては，カテゴリを出力ユニットに対応付けるネットワークを使用することが多い．この NN はひとつのネットワークで全てのカテゴリに対応することができる．しかし，カテゴリ数の増減により改めてネットワークを再形成する必要がある．これに対して，本ネットワークはカテゴリごとに形成することから，カテゴリ数が変化した場合でも容易に対応することができる．すなわち，既存のネットワークはそのまま利用でき，新たに増加したカテゴリに対するネットワークを形成すれば良いので，学習時間の短縮が可能となる．また，それぞれが単独で学習してネットワークを形成することから，他の文字の影響を受けない学習が可能である．

4.4 自己想起型ニューラルネットワークによる認識処理

あらかじめ認識候補となる文字のネットワークを学習処理によって形成しておき，これらを用いて対象文字を認識する．

まず，切り出しによって得られた文字パターンに対し，各々の特徴量を順に認識候補となる NN へ入力し誤差を算出する．誤差とは，出力層のニューロン値 O_i と理想的な出力である教師信号 T_i との差の

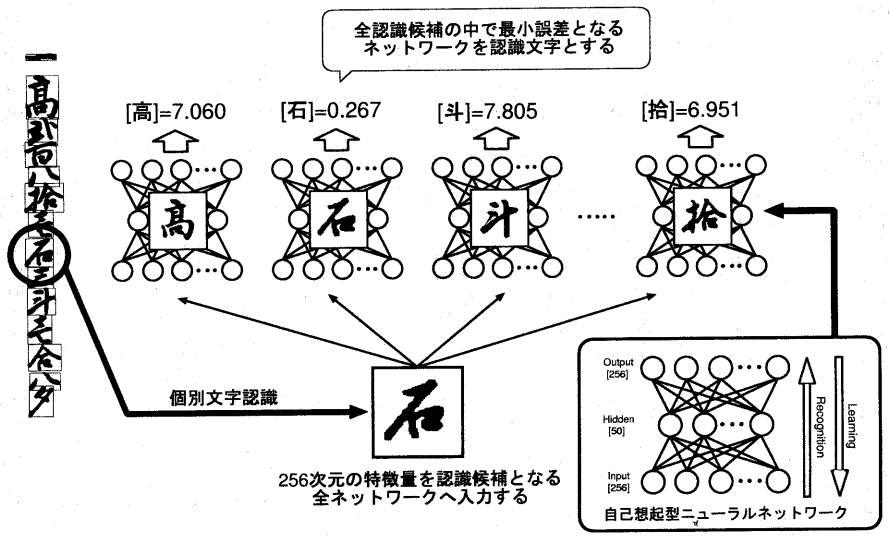


図 7: 自己想起型ニューラルネットワークによる認識処理

二乗和であり，

$$e = \sum_i (T_i - O_i)^2 \qquad (1)$$

で定義される．そして，各ネットワークにおける誤差を比較して最小誤差となる NN を第一位認識候補として認識文字とする．

図 7 に認識処理の流れを示す．例えば，文字列中の「石」を認識対象とした場合，このパターンから抽出した特徴量を全てのネットワークへ入力し，それぞれで算出される誤差を比較する．このとき，「石」に対するネットワークでの誤差が最小となれば，正しく認識されたことになる．

5 認識処理を援用した再文字切り出し

初期文字切り出しでは，文字パターンの連結成分における矩形情報のみを利用するため，文字が完全に個々のパターンに切り出されていないことが多くある．ここで提案する切り出し手法は，古文書特有のつづけ字や文字の食い込みなどを考慮したものであり，認識処理結果に基づき再文字切り出しすることによって，高精度な切り出しを実現する．

本手法では，次に示す条件を基に再切り出し候補矩形を求める．

- 条件 α : NN における二乗誤差 ≥ 2.5

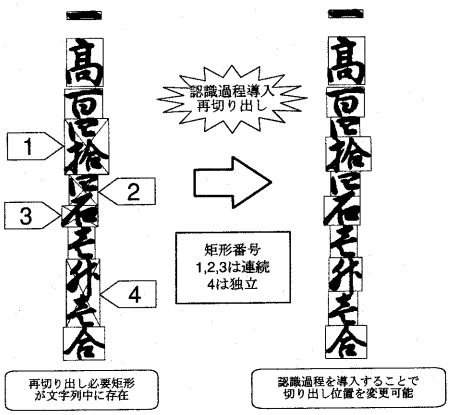


図 8: 再切り出し必要矩形

- 条件 β : NN における二乗誤差 ≥ 1.5

まず，切り出した各矩形に対して個別文字認識する．このとき，NN における誤差をもとに，条件 α を満たす矩形は誤って切り出されたとして再切り出し必要な矩形とする．ここで，全矩形において再切り出しが必要とされる矩形の位置関係は，次のことが挙げられる．

- 再切り出しが必要とされた矩形の上下矩形のどちらも再切り出し不要とされる場合（再切り出し必要矩形が独立）

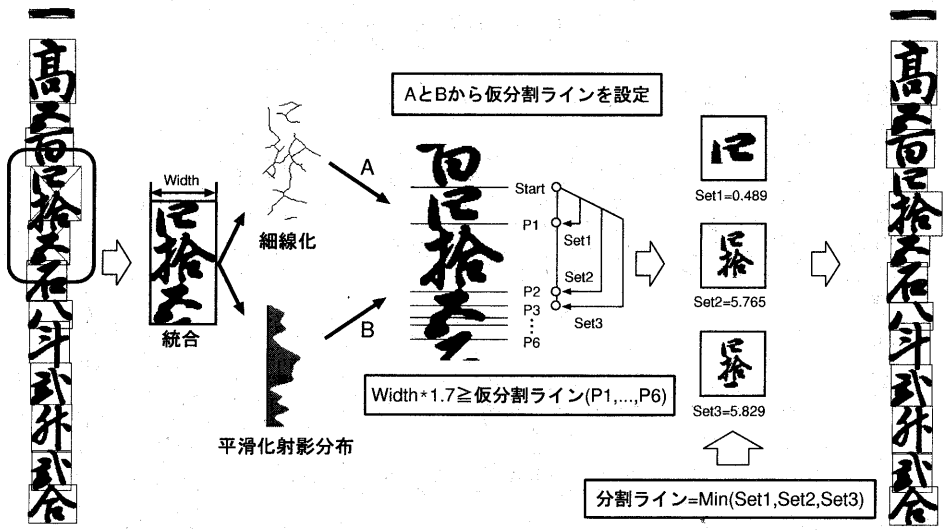


図 9: 分割ラインの決定方法

- 再切り出しが必要とされた矩形の上下矩形のいずれかが再切り出し必要とされる場合（再切り出し必要矩形が連続）

図 8 に再切り出し必要とされた矩形例を示す．ここでは，初期文字切り出しの段階で得られた切り出し候補に対し，条件 α を満たす矩形 4 個に目印となる”×”が印されている．それぞれを矩形番号 1,2,3,4 とすると，1,2,3 においては連続していることがわかる．一方，4 は上下の矩形には目印が付いていないため独立している．そこで，再切り出し必要矩形に対し再文字切り出しを適用する．

ここで，連続している矩形の位置関係から互いに少しでも離れている場合は独立しているとする．それ以外のときは，強制的に連続している矩形を統合する．図 8 では，1,2,3 の矩形の位置関係は離れていないために全て統合する．そして，統合後の矩形に対して個別文字認識する．このとき，条件 α を満たさなければ統合することによって適切に切り出された矩形とする．しかし，条件 α を満たす矩形に対しては再分割する．これ以降の処理は再切り出し必要矩形が独立している場合も同様にする．

図 9 に再分割処理手順を示す．まず，対象となる矩形に対し，Hilditch の細線化処理 [7] によって矩形内の文字パターンを線幅 1 で表し，横方向の射影分布を求める．このとき，射影値 1 となるところは文字同士が食い込んでいないところである．そこで，射影値 1 となる部分をピックアップし，これを仮の分割ライン A とする．

しかし，これだけでは文字の食い込みに対応できない．そこで，原画像における射影分布を求め，分布の平滑化を図るとともに，射影値を一定の割合で小さくする．そして，あらかじめ対象矩形において，収縮法 [8] により文字の線幅を求めておき，この線幅を閾値として，射影値が閾値より小さい部分を仮の分割ライン B とする．

以上の処理により仮の分割ラインを決定する．しかし，これでは，仮の分割ラインが対象とする矩形によっては複数になることが考えられる．そこで，仮の分割ラインが連続している場合はその中間点を最終的な仮の分割ラインとする．また独立している場合はそのままとする．これにより仮の分割ラインを最小限に抑えることができる．

仮の分割ラインから分割ラインを決定するために，矩形の最上部から各々の仮の分割ラインまでの領域に対し個別文字認識する．その際，認識における誤差が最小となるところを分割ラインとする．

まず，文字列中の一文字あたりのサイズには限界があるため，対象矩形の高さに着目して，最上部から矩形幅の 1.7 倍以内にある仮の分割ラインについて誤差をそれぞれで求める．図 9 では，仮の分割ラインが 7 本ある．そこで矩形幅 $width$ の 1.7 倍以内にある分割ラインを調べると，Start から P3 までが該当する．そのため，Start を基準として P1,P2,P3 の分割ラインまでの領域をそれぞれ Set1,Set2,Set3 として認識する．そして，認識における誤差が最小となるところを分割ラインとするため，Start から P1 までの領域が一つの文字パターンであるということに

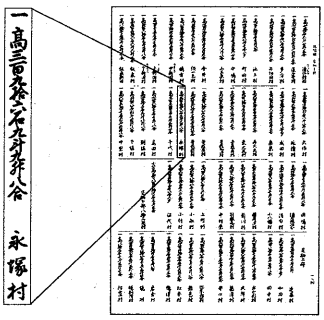


図 10: 対象データ

なる．それ以降は基準を分割ラインとした地点，すなわち P1 を Start として以後同様の処理で分割ラインを決定していく．

最後に全矩形に対して個別文字認識し，条件 β に該当する矩形を見つける．そして，上下の矩形で同様に該当しているものがあれば仮統合し，認識する．その際，条件 α を満たさないならば統合する．そして，最終的な切り出し候補を得る．

6 認識実験

認識実験の対象データとして，内閣文庫の書物「天保郷帳」[9] に収められている相模国に該当する，当時の各村における石高を表わした文字列 615 個を用いた．これらをイメージスキャナにより解像度 500dpi で採取した．一文字列あたりの画像サイズは 1140 × 100pixel である．図 10 に文字列例を示す．各文字列は前半に石高，続いて該当する村が記されている．また，対象文字を文字列から任意に 100 パターン選出し，ニューラルネットワーク形成のための学習パターンとして使用する．なお，総文字数が少ない文字種については，「千」と「才」が 20 パターン，「夕」は 39 パターンを学習パターンとした．ここでは，文字切り出しの段階で比較的文字パターンが正確に外接矩形で囲まれている（切り出されている）ものを選出した．なお，学習回数は 200 回とした．

認識実験では，文字列パターンの石高表記部のみに着目した．表 1 に石高表記部に含まれる全 20 文字種についての認識実験結果を示す．ここでは，全文字列 615 個に含まれる各々の文字総数も同時に示す．これより，文字同士にほとんど間隔が存在しない文字列において，多くの文字種で 90%以上の高い認識率が得られたことがわかる．しかし，文字の食い込みが他の文字と比べて激しい「夕」と「才」については低い認識率となった．これらの文字では，切り出しで得られた対象矩形において，他の文字の一部が同一矩形内に存在することが多くある．このことが，認識において NN で算出される誤差を大きくする要因となり，誤認識が増加したといえる．

表 1: 石高表記部の個別認識結果

文字種	文字数	認識数	認識率 (%)
一	615	615	100.00
高	615	614	99.84
壱	269	241	89.59
弐	437	414	94.74
三	412	366	88.83
四	368	346	94.02
五	397	372	93.70
六	367	332	90.46
七	347	320	92.22
八	322	297	92.24
九	323	304	94.12
拾	552	504	91.30
百	545	517	94.86
千	35	34	97.14
石	615	604	98.21
斗	546	541	99.08
升	531	516	97.18
合	535	520	97.20
夕	117	87	74.36
才	39	33	84.62
平均	7987	7577	94.87

次に，再文字切り出し導入の効果を調べるために，導入前後における認識率の違いについて検討する．図 11 に導入前後の認識率を示す．これより，多くの文字において再切り出しを導入したことで，認識率が向上していることから，その効果が見受けられる．しかし，「夕」は変化がみられず，「才」に関しては，導入したことで認識率が低下した．これは，初期文字切り出しで得られた矩形のなかから再文字切り出し

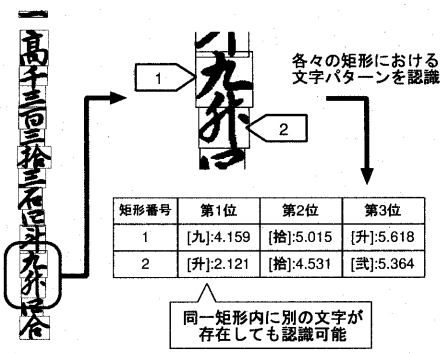


図 12: 切り出し失敗矩形の認識成功例

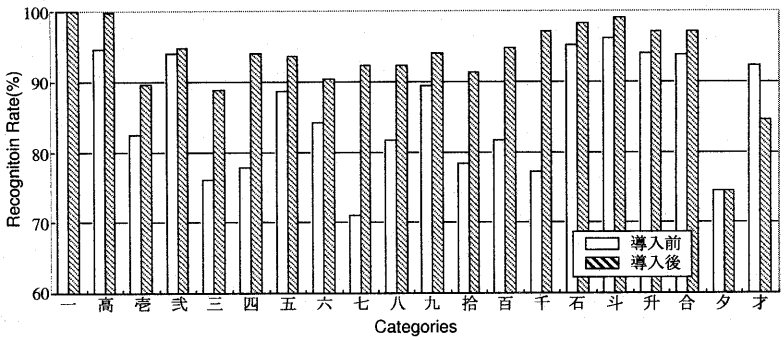


図 11: 再切り出し導入前後の認識率

を適用する矩形を選出するときに，誤って選出されたためであると考えられる．また，「才」は文字の食い込みが他と比べると激しいために誤って切り出されたといえる．他の文字に関しては食い込みやつづけ字がみられる場合でも切り出しに成功したことで認識率も向上した．このことは，切り出しの段階で認識処理を導入したことの効果であるといえよう．

さらに，切り出しに失敗した文字パターンが正しく認識された例を図 12 に示す．これは，認識候補となる NN での誤差が大きくとも，他の NN よりも小さければ正しく認識したことになる．そのため，多少切り出しに誤りが生じても認識が可能となる．また，認識候補数が 20 文字と比較的少ないために正しく認識されたとも考えられる．

7 おわりに

本論文では，つづけ字や食い込み等が原因で，前後の文字が互いに影響しあう古文書文字列に対し，これらを考慮した文字切り出し手法を提案し，切り出された各々の文字パターンがどの程度認識できるかを検討した．認識には自己想起型 NN を使用した．

その結果，提案手法では平均個別認識率は 94.87% が得られた．これより，文字切り出しの段階で認識処理を援用したことが，切り出し率の向上つながったことがわかる．また，これに伴って高い認識率が得られたといえよう．

文字列から切り出された各々の文字を認識するためには，正確に切り出せていなくても認識が可能であることがわかった．しかし，文字列を認識するためには全ての文字に対し切り出せなければ認識が不可能である．そのため，文字列から文字数分の切り出し候補を得る必要があるといえる．今後は，文字列として認識するためにも，高精度な文字切り出し手法について再検討が必要となる．また，キャラクタスポッティング [5] を応用し，キー文字抽出による認識精度の向上につなげる検討を加える必要がある．また，他の書体の古文書にも応用できるように検討しなければならない．

参考文献

[1] 山田奨治 "古文書 OCR 研究の現在" 挑戦古文書 OCR，人文学と情報処理，No.18，pp.2-5，1998.

[2] 日置慎治，上原邦彦，川口洋 "「宗門改帳」に記録された年齢表記の認識" 挑戦古文書 OCR，人文学と情報処理，No.18，pp.35-42，1998.

[3] 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.

[4] 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ワークによる古文書文字認識" 電気関係学会関西支部連合大会，G13-14，2000.

[5] 橋本智広，梅田三千雄 "古文書文字列に対するキャラクタスポッティング" 人文科学とデータベースシンポジウム，No.7，pp.29-38，2001.

[6] 鶴岡信治，栗田正徳，栗田昌徳，原田智夫，木村文隆，三宅康二 "加重方向指数ヒストグラム法による手書き漢字・ひらがな認識" 電子情報通信学会論文誌，Vol.J70-D-II，No.7，pp.1390-1397，1987.

[7] 手塚慶一，北橋忠宏，小川秀夫 "ディジタル画像処理工学"，pp.139-142，日刊工業新聞社，1985.

[8] 加藤寧，根元義章 "ストローク情報に基づく手書郵便宛先の切り出しと認識" 画像ラボ，Vol.8，No.8，pp42-45，1997.

[9] 内閣文庫所蔵史籍叢刊 55「天保郷帳 (一)」