背景領域の細線化に基づく古文書の文字切り出しと認識

梅田 三千雄† 橋本 智広†

本論文では，古文書文字列を対象として，古文書特有の文字の接触や食い込みに対処するために，背景領域に着目した文字切り出し手法を提案する．まず，対象文字列とその鏡像パターンを結合した合成パターンの背景領域に対して細線化処理を施し，基本パターンを生成する．次に，基本パターンに対してラベリング処理によりパターン内で区分けされている各々の領域を求め，これらに対し個別文字認識する．この認識結果から，文字領域と判断できない領域を検出し，認識処理を援用した領域確定処理を適用する．領域確定処理では，2段階で分割経路を変更し，隣接する複数の領域を組み合わせながら認識処理を繰り返すことで，最適な文字領域を求める．そして，得られた各領域から抽出した特徴量を自己想起型ニューラルネットワークに入力することで認識結果を得る．「天保郷帳」を例とした615個の文字列に対する認識実験により，本手法によって個別文字認識率は98.52%，文字列認識率は90.24%が得られ，文字部の画素に着目した従来手法と比較して，その有効性が確認された．

Character Segmentation and Recognition of Ancient Documents Based on Thinning of Background Region

Michio Umeda† and Tomohiro Hashimoto†

This paper proposes a character segmentation and recognition method of ancient documents. The segmentation method is based on thinning the background region of a compound pattern in order to cope with the cursive scripts and the mutual encroachment of characters which are peculiar to the ancient documents. The compound pattern is generated from the original characters string pattern and two mirror patterns. In the segmentation process, candidate dividing points are extracted from the thinning pattern and the segmented regions are gradually determined by using a recognition processing. In the recognition process, autoassociative neural networks are used for flexibility and efficiency. From the recognition experiment applied to 615 character strings which appear in the local Tenpo era records of rice crops, the correct character recognition rate of 98.52% and the correct string recognition rate of 90.24% were obtained by the proposed method. Therefore it is clarified that the method is effective in the recognition of characters such as ancient documents.

1. はじめに

手書き文字認識に関する研究は，これまでに多くの研究機関で試みられ，様々な認識手法が提案されたことで，その技術は実用の段階にある[1]．一方，人文学研究の分野では，古文書データベースの作成を支援するOCRの実現を目指し，その認識手法の提案が期待されている[2]．

現在，古文書データベースの構築においては，人間が自ら作業を行うために，史料の解読や文字データ入力，編集などに多大な作業時間を要している．この作業が，コンピュータによって自動化可能となれば，飛躍的に作業時間を短縮することができ，多量の古文書史料を短時間で，効率良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[3)~7)]．

古文書を認識対象とすると，史料に含まれる文字パターン数には限度があり，認識で使用する辞書作成において，十分なデータ量を採取することが困難となる．したがって，限られた範囲内でも高精度の認識が実現できる，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，古文書を認識するためには，文字を正確に切り出すことが重要である．しかし，毛筆で筆記されていることにより，古文書特有のつづけ字や食い込みなどが頻出し，これまでに提案された文字切り出し手法をそのまま適用しただけでは，正確な切り出しが困難であり，高い認識精度が得られないなどの問題もある．

従来の文字切り出しでは，文字部を構成する黒画素

† 大阪電気通信大学大学院工学研究科
Graduate School of Engineering, Osaka Electro-Communication University

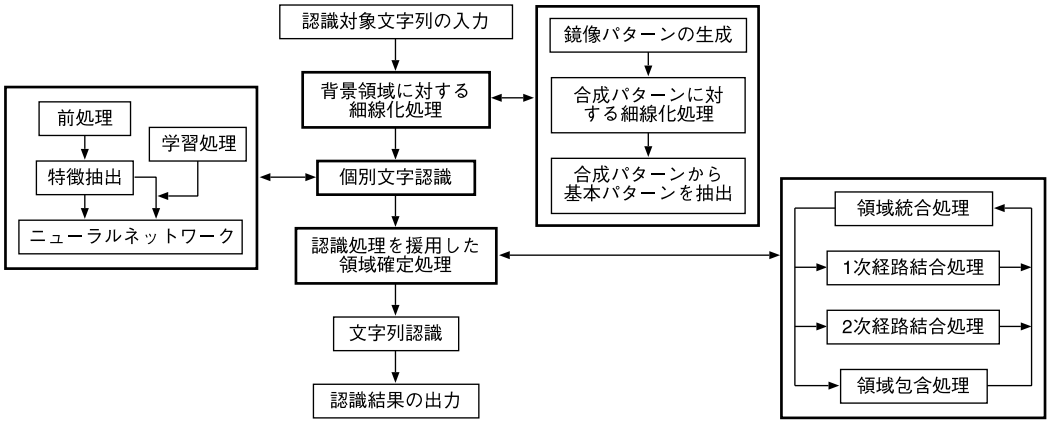

図 1 文字切り出しと認識処理の流れ
Fig. 1 Flow of character segmentation and recognition processing.

の射影ヒストグラムを求めて，その形状や変化量から文字列パターンを線形に分割し，切り出したパターンを矩形で囲み，その矩形の面積や縦横比から判断して切り出す方法が提案されている[6)~9)]．また，文字のつながりあう接触部分に着目して，その形状や接続状態から分割する手法[7),10),11)]などもある．いずれの手法も文字部を対象領域として着目した切り出し手法である．しかし，古文書のような食い込みが激しい文字パターンに対して，このような手法では期待する位置での正確な切り出しは不可能である．たとえば，パターンを矩形で囲む方法[6),7)]では，古文書のように大きな食い込みが生じると，他の文字の一部がその矩形内に侵入してしまう．その結果，この余分な領域が整合処理における誤差に影響を及ぼし，誤認識を招くことになる．

本論文では，文字列パターンに存在する文字領域ではなく，背景領域に着目し，かつその細線化パターンを利用することで，非線形な分割を実現し，文字の接触や食い込みなどにも対処可能な文字切り出し手法を提案する．

文字列の構造に応じた切り出しを実現するために，背景領域に対して Hilditch の細線化処理[12)]を施し，基本パターンを作成する．次に，この基本パターンから初期の分割領域を求めて，各領域に対し個別文字認識する．個別文字認識では，加重方向指数ヒストグラム特徴[13)]を用いて特徴抽出する．さらに，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）を使用する．ここでは，認識カテゴリの変化に容易に対応できる自己想起型 NN[14)]を用いた．次に，認識処理結果から文字領域と判断できない領域を検出し，認識処理を援用した領域確定処理を適用する．そして，この処理を繰り返すことで最終的な文字領域を確定し，認識結果を得る．

認識実験では，古文書として「天保郷帳」[15)]を例にとり，文字列中の石高表記部に存在する全 20 字種に対する個別認識と文字列認識について，この方法により，どの程度認識可能であるかを検討する．さらに，黒画素に着目して線形に切り出す従来の手法[6),7)]と比較することで，本手法の有効性を検討する．

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，認識対象文字列から左右に鏡像パターンを作成し，原画像と結合して，合成パターンを得る．次に，合成パターンの背景領域に対して細線化処理を施す．得られた細線化パターンから対象文字列に該当する領域だけを抽出し，これを基本パターンとして処理を進める．また，この細線化パターンを経路として文字切り出しを実行することから，これを分割経路と呼ぶことにする．

次に，得られた基本パターンに対してラベリング処理により，明らかに独立している領域（以下初期領域と呼ぶ）を求める．そして，この初期領域を対象として，個別文字認識する．認識処理では，前処理として，孤立点除去，大きさの正規化，スムージングを施し，正規化後のパターンから特徴抽出によって特徴量を算出する．特徴抽出には，加重方向指数ヒストグラム特徴を用いた．さらに，学習処理において，バックプロパゲーション法により認識対象文字ごとに自己想起型 NN を形成しておく．

基本パターンから得られた初期領域は，文字領域が正確に抽出できていないことが多い．そこで，認識処

理を援用しながら文字領域を確定していく．この領域確定処理では，2 段階で分割経路を変更し，最適な文字領域を求める．1 次経路結合処理では，基本パターンにおける分割経路の端点を検出し，各端点を中心とした一定の円内に他の端点が存在すればこれらを結合する．2 次経路結合処理では，1 次経路結合処理で結合されなかった端点について，はじめに着目した端点が開始点か終了点のどちらから出発した経路であるかを調べ，進行方向に対して，一定角度内に存在する経路画素と結合する．さらに，最終処理として，対象文字列に存在する文字数と切り出し文字数が同一となるように，文字領域でないと判断した単独領域に対して，隣接領域との包含処理を施す．

以上の処理によって文字切り出しを完了し，得られた文字領域から抽出した特徴量を自己想起型 NN に入力することで認識結果を得る．

3. 背景領域の細線化による分割基準点の設定

従来の文字切り出し手法の多くは，水平または垂直方向に線形に切り出すものである．この手法では，文字と文字とに間隔が存在するときは問題ないが，文字の食い込みがみられる場合には，隣接文字の一部が侵入してしまう．その結果，侵入した部分も切り出された文字領域の一部となり，認識に大きく影響する．また，文字数だけの切り出しが実現できないのも問題である．

これらの問題点に対処するには，文字列の構造を反映した文字切り出しを実現する必要がある．つまり，文字どうしに十分な間隔がある部分では線形に，また食い込みなどが存在する部分では非線形に切り出すことを意味する．これに基づく切り出し手法も提案されているが，いずれも文字部に着目したものである．ここでは，切り出しに必要な情報が文字の背景部に存在することに着目し，背景領域の細線化に基づく切り出し手法を提案する．

まず，背景領域に対して細線化処理を施す．このとき，対象文字列パターンのみに対して処理を適用すると，**図 2** のように細線化パターンの外周が凹凸形状となる．そのため，1 文字ごとへの区切りに必要な分割基準点だけを検出することが難しくなり，処理が複雑になる．

そこで，基準点の設定を容易にするため，対象文字列に対する鏡像パターンを作成し，これらを結合した合成パターンに対して細線化する．そして，得られた細線化パターンから対象文字列となる領域だけを抜き出して基本パターンとする．

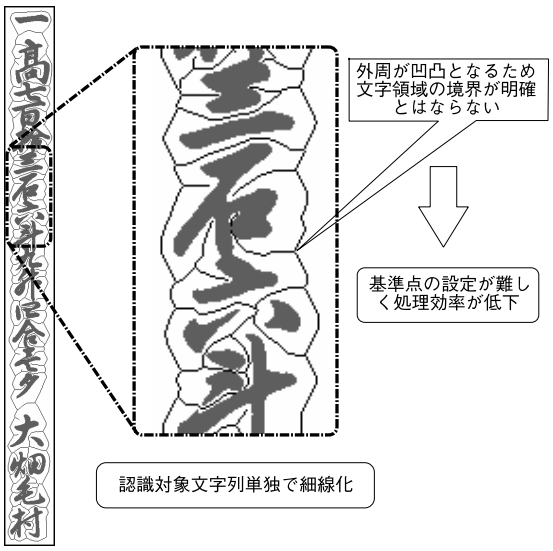

図 2 文字列の背景から得られる細線化パターン
Fig. 2 Thinning pattern generated from background region of character string pattern.

図 3 にその処理例を示す．まず，対象文字列に対して左右に鏡像パターンを作成し，合成パターンを得る．そして，このパターンの背景領域を対象とした細線化処理により分割経路を求める．合成パターンにおける中央部の領域が本来の対象文字列であるため，縦の直線で得られる分割経路とその内部だけを抽出し，基本パターンとする．これを図 2 の結果と比較すると，合成パターンを使用することにより，明確に左右の基準点を設定できることが分かる．以後，この基準点を開始点，終了点と定め，基本パターンをもとに処理を進める．

4. 個別文字認識

文字切り出しによって得られた各文字パターンに対して，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を均一化する．次に，比較的高い認識率が期待できるとされる加重方向指数ヒストグラム特徴により特徴抽出する．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．

4.1 前 処 理

切り出した個々の文字パターンは大きさにばらつきがある．そこで，文字パターンを均一化することを目的として前処理を施す．まず，画像に含まれている雑音を除去する孤立点除去，大きさにばらつきのある文字パターンの 2 次モーメント

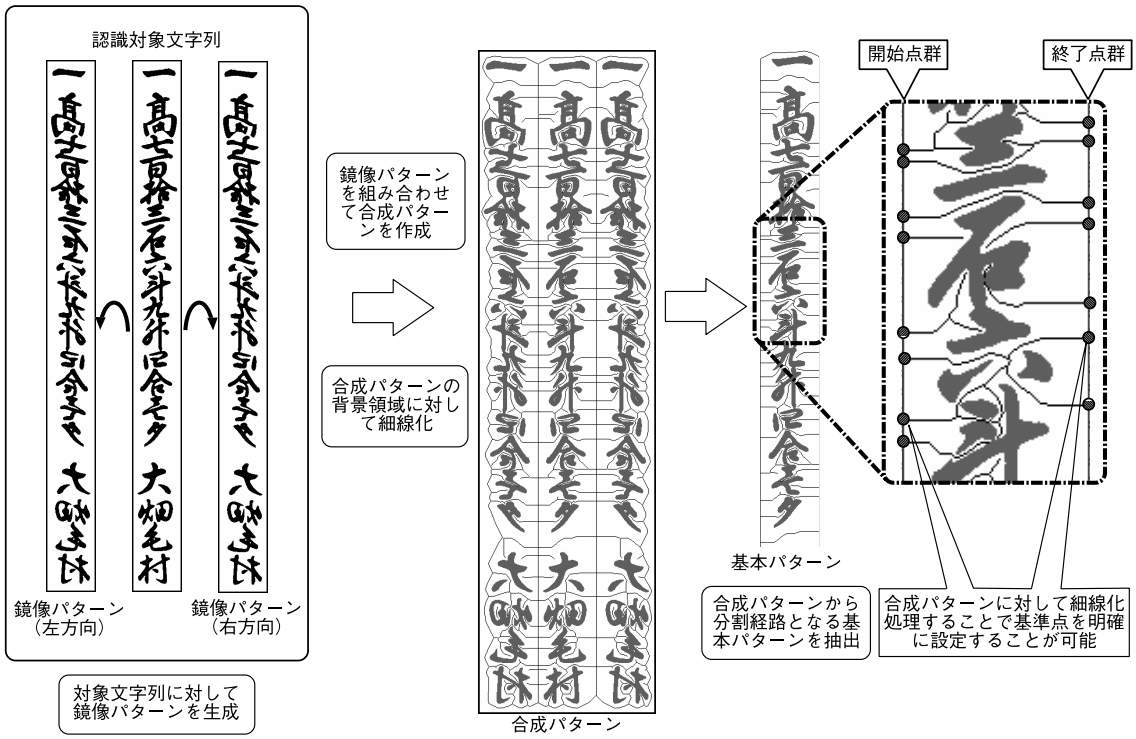

図 3 背景領域に対する細線化
Fig. 3 Thinning applied to background region in this method.

$$r_m = \frac{\sum f(x,y)\cdot\sqrt{(x-X_m)^2+(y-Y_m)^2}}{\sum f(x,y)} \quad (1)$$

$f(x,y)$：文字パターン　r_m：2 次モーメント

X_m, Y_m：文字パターンの重心

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処理を施す．

4.2 加重方向指数ヒストグラム特徴

特徴抽出法には，文字の輪郭部に着目した加重方向指数ヒストグラム特徴を用いた．文字パターンに対して輪郭を追跡しながら，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 4** (a) に示すように，注目画素に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みた方向指数は図 4 (b) から 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．方向指数を算出すると，次に各方向指数に対して方向圧縮する．まず，奇数方向を中心に 1：2：1 の重みを付け，前後の偶数方向を足し込むことにより，

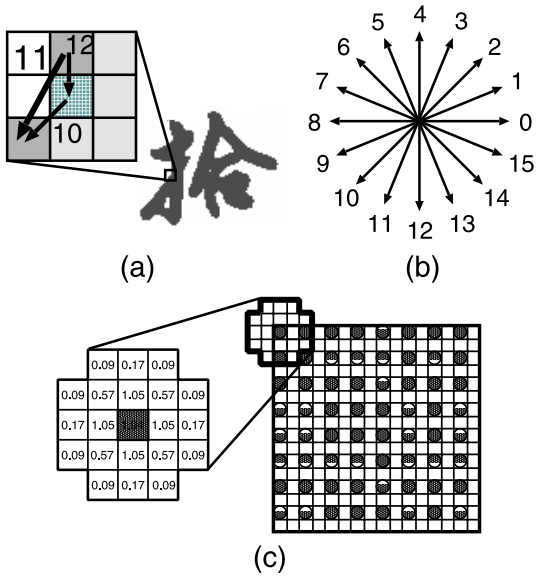

図 4 加重方向指数ヒストグラム特徴
Fig. 4 Weighted direction index histogram feature.

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 方向まで圧縮する．一方，領域圧縮として，96×96 画素の領域に対し，16×16 領

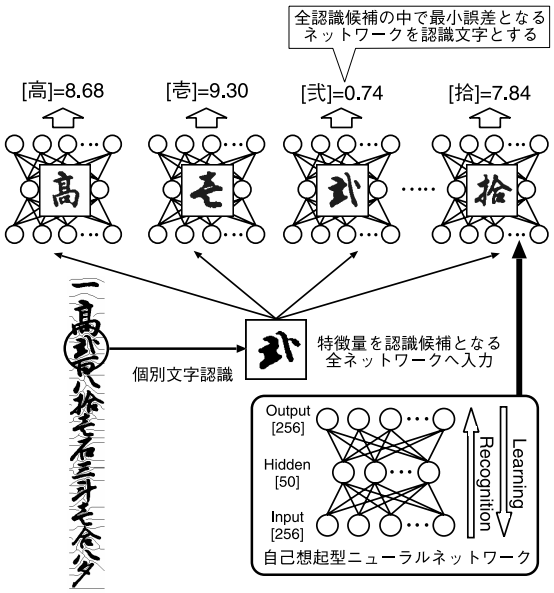

図 5 自己想起型ニューラルネットワークによる認識処理
Fig. 5 Recognition processing by autoassociative neural networks.

域の小領域に分割してヒストグラムを求める．さらに，局所的なぼかしの働きを持つ図 4 (c) に示す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 1 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．

4.3 自己想起型ニューラルネットワーク

ここでは，古文書のような対象カテゴリ数を規定しにくい文字認識をするうえで，高い精度と柔軟な対応が期待できる自己想起型 NN を用いることとした．これ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．したがって，教師信号には入力する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）により，教師値と出力値の誤差が小さくなるように，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．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値 T_i との差の二乗和であり，

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

で定義される．**図 5** 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，中間層は予備検討により 50 個とした．

文字認識においては，その柔軟で，かつ高い汎化能力から NN が利用されることが多い．しかし，そのほとんどは出力ユニットにカテゴリを対応付ける階層型ネットワークである．これは，1 つのネットワークで全カテゴリの認識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ネットワークを形成することから，対象カテゴリ数が変化した場合でも容易に対応することが可能となる．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だけを形成すれば済むので，学習時間の短縮が図れる．また，それぞれがカテゴリごとに独立に学習してネットワークを形成することから，他の認識対象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

この NN はカテゴリごとにネットワークを形成することから，あらかじめ認識対象として定めた文字のネットワークのみを学習処理によって形成しておき，これらを認識処理に使用する．まず，切り出された文字パターンから抽出した特徴量を順に認識対象となる全ネットワークへ入力して，誤差を算出する．そして，各ネットワークにおける誤差を相互に比較し，最小誤差となる NN のカテゴリを認識結果，あるいは第 1 位認識候補とする．

この認識処理例を図 5 に示す．たとえば，文字列中の「弐」の部分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すべ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，「弐」に対するネットワークでの誤差が最小であれば正しく認識できたとする．

5. 認識処理を援用した領域確定処理

背景領域に対する細線化によって得られた基本パターンから，4 連結ラベリング処理によって基本パターン内で区分けされている領域，つまり初期領域を求める．そして，これらに対し個別文字認識する．このとき NN によって求められた認識誤差から，文字判別しきい値処理によって

認識誤差 ≤ 1.25

を満たす領域は文字領域であるとし，これを確定領域と呼ぶ．なお，しきい値は予備実験により設定した．この時点では，各領域がどの字種に認識されたかは考慮しない．一方，この条件を満たさない領域は，その領域単独では文字領域をなさないと見なす．これを未確定領域と呼ぶ．これらの未確定領域に対して，次の領域統合処理をはじめとする領域確定処理を適用する．

5.1 領域統合処理

未確定領域に対して領域統合処理を施す．これは未

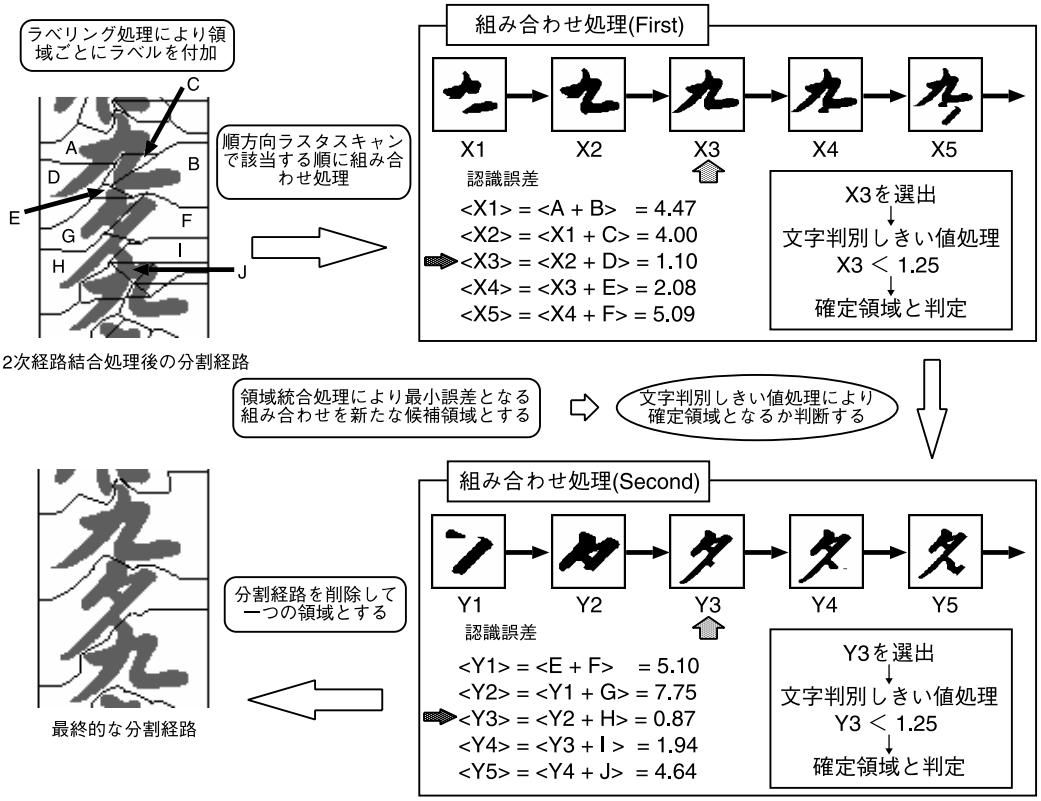

図 6 領域統合処理
Fig. 6 Region integration processing.

確定領域どうしや未確定領域と確定領域を統合することによって新たな領域を形成し，複数の領域を順次組み合わせながら認識することで，確定領域を得るものである．この処理は，初期領域に対して適用し，かつ次に述べる 2 種類の経路結合処理後に得られた領域に対して適用する．

まず，順方向ラスタスキャンで該当する領域を順次結合し，NN を用いて認識誤差を算出する．次に，各領域を結合することで得られる認識誤差を比較し，誤差が最小となる領域を新たな候補領域とする．そして，文字判別しきい値処理によって確定領域となるかを判断する．

図 6 に領域統合処理例を示す．これは 2 次経路結合処理後の処理例である．まず，順方向ラスタスキャンにより領域 A が未確定領域として選出され，次に領域 B が選出される．そして，これらを組み合わせて領域 X1 を仮形成し，NN によってこのときの誤差を求める．次に，領域 X1 と領域 C を組み合わせて領域 X2 を仮形成し，誤差を求める．以下同様に各領域を組み合わせていき，図の例では，最小誤差となる領域 X3 を新たな候補領域として確保する．そして，文字判別しきい値処理により領域 X3 の誤差が条件を満たせば確定領域とする．領域 E 以降についても同様の処理により，1 文字ずつ領域を確定していく．すべての領域が確定されると，余分な分割経路を削除し，文字切り出しのための新たな分割経路を得る．

5.2 1 次経路結合処理

1 次経路結合処理では，基本パターン内に存在する分割経路の端点を選出し，この端点と別の端点の位置関係からこれらを結合する．基本パターンに存在する分割経路は，開始点から終了点まで基準点間を結ぶ 1 つの経路として存在することが望ましい．しかし，**図 7** のように，文字どうしのつながりや重なり合う部分では，分割経路が途中で途切れて結合しない．すなわち，経路が途中で切断され，領域を分割する適切な経路として成立しない箇所が多く見られる．

1 次経路結合処理では，この問題に対処するため基本パターンにおける分割経路の端点を検出し，近隣の端点を検出して双方を結合する．

まず，検出した端点を中心として半径 10 pixel の円内に他の端点が存在する場合は，これらを結合する．半径の大きさは，任意に指定することにより，結合範

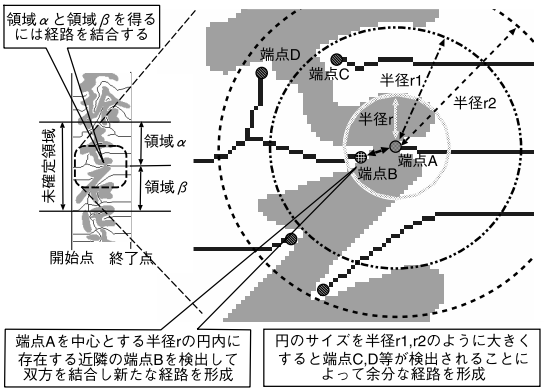

図 7 1 次経路結合処理
Fig. 7 The first processing of route connection.

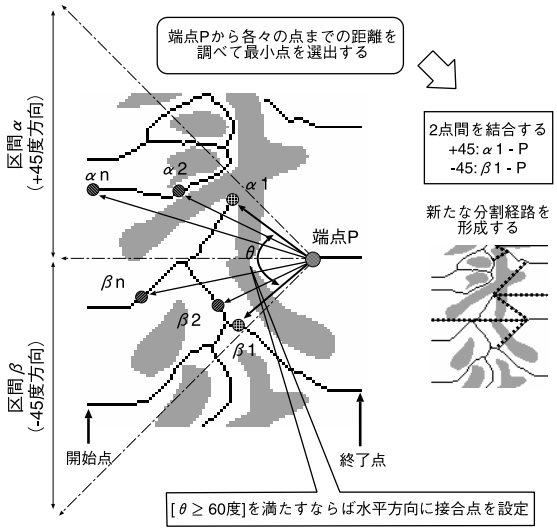

図 8 2 次経路結合処理
Fig. 8 The second processing of route connection.

囲を拡大させることができるが，ここでは結合処理を最小限にするため，平均文字線幅内に存在する近隣の端点のみと結合することを目的として設定した．そして，領域統合処理を適用し，各々の領域を判断する．

図 7 に 1 次経路結合処理例を示す．ここでは，検出した端点 A から指定した半径 r の円内に存在する端点 B を結合することで切断されていた経路がつながり，新たな分割経路を形成することが可能となる．しかし，半径を r1 とした場合には，端点 C が該当するため余分な経路を形成することになる．このことは，期待しない領域を形成したり，領域統合処理にともなう処理時間の増加を招いたりすることになる．

5.3 2 次経路結合処理

2 次経路結合処理は，1 次経路結合処理で該当しなかった端点に対して適用する．まず，対象とする端点が分割経路に対して設定した開始点と終了点のどちらから導出したものであるかを調べる．このとき仮に，**図 8** のように終了点から出発したと判断されれば，その端点から開始点に向かって新たな分割経路を追加していく．

まず，端点 P を原点として，経路を追加しようとする方向に対して+45 度方向と −45 度方向の区間 α，β を求める．そして，それぞれの区間内に存在する分割経路上の点 (r_i, r_j) と着目端点 (p_i, p_j) との距離 D_r を

$$D_r = \sqrt{(r_i - p_i)^2 + (r_j - p_j)^2} \qquad (3)$$

により求め，その距離が最小となる点を選出する．これを各方向 1 点ずつ，合計 2 点求めて結合点とし，端点 P と結合する．これより，1 次経路結合処理のように端点と端点だけを結合するのではなく，経路途中にも結合点を設定することが可能となる．しかし，端点と選出した 2 点のなす角度 θ が大きい場合には，期待する経路が形成されないことがある．そこで，角度判別しきい値処理を導入して，

$$\theta \geq 60$$

を満たすときは，水平方向に結合点を加えることとした．つまり，端点 P から 3 点につながる分割経路が設定される．これより，不自然に分割しようとしたときには，端点から線形に分割する可能性が残されることになる．

5.4 未確定領域に対する包含処理

2 段階の経路結合処理を終了しても，なおかつ未確定領域が残る場合には，次の包含処理を適用する．

まず，検出された未確定領域において，領域サイズの高さが 50 pixel 未満であれば，隣接する確定領域と結合して認識する．この値は目視による文字パターンの分布により，高さが平均文字列幅の 2/3 未満であれば，1 つの文字領域でない可能性が高いとして設定した．そして，認識誤差が最小となる結合領域を検出し，どの領域と包含すればよいかを決定する．

6. 従来手法による文字切り出し

認識実験においては，これまでに提案されている切り出し手法と比較検討する．これらの手法は，いずれも文字を構成する黒画素に着目して切り出すものである．各手法による文字切り出し結果を**図 9** に示す．

手法 1[6]は，黒画素の横方向射影ヒストグラムと，文字パターンを囲む外接矩形の高さや面積などを手がかりとして切り出すものである．手法 2[6]は，手法 1

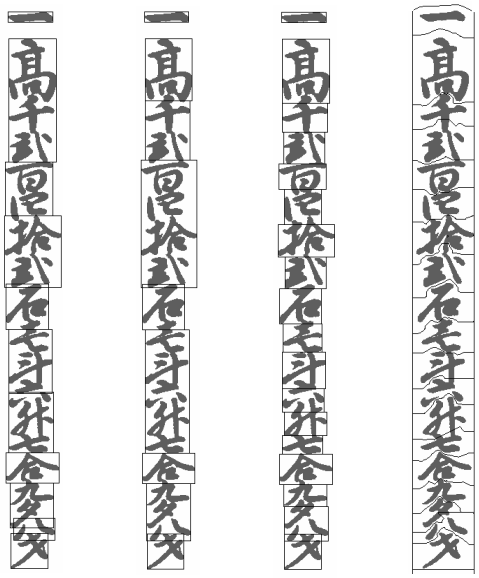

図 9 各手法による文字切り出し結果
Fig. 9 Character segmentation result by each method.

に加えて再切り出しを導入し，切り出し失敗矩形を選出して認識処理を援用しながら最適な切り出し位置を定めるものである．分割位置は，選出された矩形に対して，細線化と黒画素の横方向射影ヒストグラムから分割候補を設定し決定する．このとき，連続して失敗した矩形が選出されると，これらを統合した後に再切り出しを適用するため，文字列によっては図 9 のように分割位置選出に失敗することがある．そこで，より適切な切り出し位置の設定を目的とした手法 3[7]は，文字と文字との接合点を検出して分割候補を増やし，この問題に対処したものである．これにより，手法 2 に比べ飛躍的に精度を向上することが可能となる．しかし，文字の食い込みが激しい部分については，水平方向に直線で切り出すため，正確な切り出しが不可能となる．これらに対して，本手法では非線形な切り出しが実現でき，食い込みにも対処可能となることが分かる．

7. 認識実験

認識実験の対象とする古文書データとして，内閣文庫の書物「天保郷帳」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 500 dpi で採取した．1 文字列あたりの画像サイズは縦 1,140 pixel×横 100 pixel である．**図 10** に対象とした文字列例を示す．この古文書では，石高表記部に

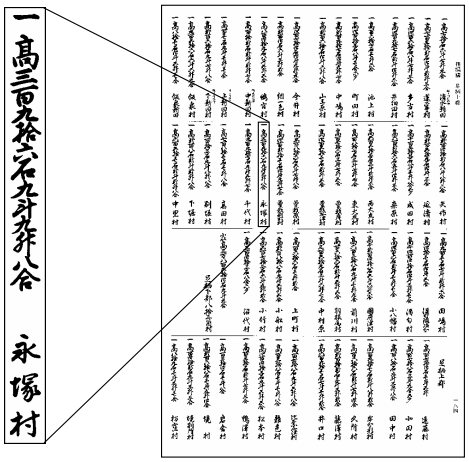

図 10 対象とした古文書
Fig. 10 Ancient document used in this study.

限れば，文字列は間隔をとって筆記されており，文字列間の文字の接触や食い込みは存在しない．また，ある文字から続けて次の文字が筆記されることもない．

自己想起型 NN による学習では，文字列から任意に 100 パターンずつ選出し，学習パターンとして使用した．なお，総文字数が少ない字種については，「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．これらは，文字切り出しの段階で比較的文字パターンが正確に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，文字列パターンにおける石高表記部のみに着目する．石高表記部の文字列パターンの切り出しには，黒画素の射影ヒストグラムを用いた．石高表記部に含まれる全 20 字種を認識対象として，全文字列 615 個に出現する文字総数とその個別文字認識結果を**表 1** に示す．表には，比較の対象とした各手法による認識結果もあわせて示した．

この結果より，ほとんどの字種において，本切り出し手法を導入することによって，高い認識率が得られることが分かる．特に，隣接文字の食い込みや接続が多く見られる「合」や「夕」などの字種において，従来手法では黒画素に着目して切り出していたことから，誤った位置で分割してしまうことが認識に大きく影響したのに対し，背景領域のみに着目して切り出す本手法で高精度な認識が可能となった．このことは，非線形な文字切り出しがもたらした効果であるといえよう．

しかし，「才」においては，認識率が低下した．これには，次の 2 つの理由が考えられる．1 つは，自己想起型 NN の学習に使用したパターン数が 20 個と少ないことから，形成されたネットワークは汎化能力に

表 1 各切り出し手法に基づく個別文字認識結果
Table 1 Character recognition result based on each segmentation method.

単位：(%)

字種	文字数	手法 1	手法 2	手法 3	提案手法
一	615	100.00	100.00	100.00	100.00
高	615	94.63	99.84	100.00	100.00
壱	269	82.53	89.59	94.42	96.65
弐	437	94.05	94.74	97.94	99.54
三	412	75.97	88.83	93.93	96.60
四	368	77.72	94.02	97.01	98.37
五	397	88.66	93.70	96.22	96.98
六	367	84.20	90.46	94.28	98.37
七	347	70.89	92.22	97.69	98.56
八	322	81.68	92.24	97.20	98.45
九	323	89.47	94.12	96.90	99.07
拾	552	78.44	91.30	98.37	98.91
百	545	81.65	94.86	98.72	98.53
千	35	77.14	97.14	94.29	94.29
石	615	95.28	98.21	98.21	99.35
斗	546	96.15	99.08	99.08	99.45
升	531	93.97	97.18	98.31	98.87
合	535	93.83	97.20	96.26	99.44
夕	117	74.36	74.36	73.50	86.32
才	39	92.31	84.62	71.79	79.49
平均	7987	87.99	94.87	97.15	98.52

表 2 各手法による文字列認識率
Table 2 Character string recognition rate by each segmentation method.

対象手法	認識文字列数	文字列認識率 (%)
手法 1	209	33.98
手法 2	449	72.85
手法 3	531	86.34
提案手法	555	90.24

乏しく，認識動作が不安定になった．他の 1 つは，目視による学習用パターンの切り出しにおいて，矩形で切り出したために，領域の一部に他の文字が侵入した文字パターンが存在し，これが認識精度に影響を及ぼした．

次に，個別文字認識ではなく，文字列としてどの程度の認識率が得られるかについて検討する．ここでは，文字列中に存在する文字数だけの切り出しが実現でき，かつすべての文字が正しく認識できて初めて正解とする．その認識結果を**表 2** に示す．

これより，90%を超える文字列認識率が得られ，本手法を用いることで文字列の構造を反映した切り出しが可能となり，高精度な文字列認識が実現できるといえる．しかし，認識できなかった文字列の中には，**図 11** の例のように，細線化によっても期待する位置に端点が出現せず，分割経路が形成されなかったことで切り出しに失敗したものが見受けられた．また，領域統合処理において，領域どうしを組み合わせる順序によっては，適切に切り出すことが困難となる部分も存在した．

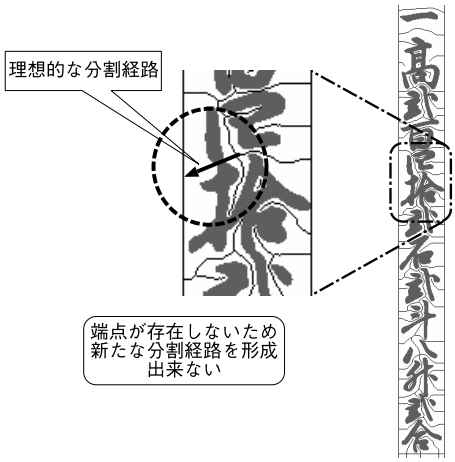

図 11 文字切り出し失敗例
Fig. 11 Failure example of character segmentation.

8. おわりに

本論文では，文字の接触や食い込みなどが頻出する毛筆書体の古文書文字列を対象として，背景領域に着目した細線化に基づく文字切り出し手法を提案し，文字列としての認識の可能性について検討した．

文字列からの文字切り出しは，従来からの文字部を構成する黒画素に着目する方法ではなく，背景領域の細線化処理によって得られた細線化パターンを分割経路とし，認識処理を援用しながら段階的に経路結合と領域統合を進めることで，非線形な切り出しを実現した．なお，細線化処理は，左右に鏡像パターンを結合した合成パターンに適用することにより，分割のための基準点の検出を容易にした．また，認識処理では，対象カテゴリ数の増減に柔軟に対応できる，自己想起型 NN を用いた．

その結果，提案した文字切り出し手法により，個別文字認識率は 98.52%となり，文字列認識率は 90.24%が得られた．これを従来手法での認識結果と比較すると，本手法によって非線形な切り出し領域を設定したことで，切り出し精度が向上し，それが認識精度の向上につながったといえる．逆に，高精度な文字列認識を実現するためには，正確な文字切り出しが重要であるともいえよう．

今後は，認識精度をより向上させるためにも，文字切り出しについてさらに検討を加える必要がある．たとえば，細線化によっても期待する位置に端点が出現しない場合や，逆に複数箇所で文字がつながることによって余分な孔が生じ，結合すべき端点が数多く存在する場合への対処などである．

また，本手法は毛筆書体の文字列だけでなく，ボールペンや他の筆記具による文字列に対しても有効な方法であると考えられることから，他の古文書に対象を拡大するとともに，古文書以外の文字列にも適用して，その有効性を検討する必要がある．

参 考 文 献

1) Umeda, M.: Advances in Recognition Methods for Handwritten Kanji Character, *IEICE Trans. Inf. Syst.*, Vol.E79-D, No.5, pp.401–410 (1996).
2) 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).
3) 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).
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，2000-CH-45-2 (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).
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，2002-CH-53-8 (2002).
7) 橋本智広，梅田三千雄：認識処理とストローク接合部検出を融合した石高表記文字列認識，電気関係学会関西支部連合大会，G11-2 (2002).
8) 馬場口登，塚本正敏，相原恒博：認識処理の導入による手書き文字切出しの一改良，電子情報通信学会論文誌，Vol.J69-D, No.11, pp.1774–1782 (1986).
9) 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り出しに関する一手法，情報処理学会論文誌，Vol.38, No.2, pp.280–289 (1997).
10) 諏訪美佐子：グラフ理論の手法を利用した自由手書き文字切出し，信学技報，PRMU2000-87 (2000).
11) 山口輝幸，吉川大弘，篠木　剛，鶴岡信治：線分の接続状態に基づく手書き接触文字の分割法，信学技報，PRMU2000-178 (2001).
12) 手塚慶一，北橋忠宏，小川秀夫：ディジタル画像処理工学，日刊工業新聞社 (1985).
13) 鶴岡信治，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, No.7, pp.1390–1397 (1987).
14) 井上　聡，若林哲史，鶴岡信治，木村文隆，三宅康二：自己想起回路による手書き数字認識，情報処理学会論文誌，Vol.39, No.8, pp.2476–2484 (1998).
15) 内閣文庫所蔵史籍叢刊 55「天保郷帳（一）」.

(平成 15 年 3 月 3 日受付)
(平成 16 年 2 月 2 日採録)

梅田三千雄（正会員）

昭和 20 年生．昭和 43 年大阪大学卒業．同年日本電信電話公社（現 NTT）入社．平成元年大阪電気通信大学工学部教授．現在，同総合情報学部教授．工学博士．文字認識，画像処理，認知科学等の研究に従事．電子情報通信学会，電気学会，映像情報メディア学会，画像電子学会，計量国語学会各会員．

橋本　智広

昭和 53 年生．平成 13 年大阪電気通信大学情報工学部情報工学科卒業．同年同大学大学院工学研究科博士前期課程情報工学専攻入学．現在在学中．文字認識，特に古文書を対象とした文字認識に関する研究に従事．電子情報通信学会会員．