完全一致法を用いた手書き住所文字列の認識

梅田 三千雄† 本庄 大介†

自由書式の手書き住所文字列の認識では，文字切り出しと個別文字認識の高精度化が重要な課題である．特に自由書式では，文字の大きさや形状にばらつきがあることから，正確な文字切り出しは非常に困難である．本稿では，切り出し候補ラティスと完全一致法を用いた手書き住所文字列の認識手法を提案する．まず，文字切り出しでは，ラベリング処理によって小さな基本矩形に分割し，これを組み合わせることで，切り出し候補ラティスを作成し，切り出しに多くの可能性を与える．次に，切り出し候補をそれぞれ個別文字認識し，後処理的役割として，完全一致法を用いる．完全一致法では，切り出し候補の矩形数と同じ文字数の住所のみを認識対象とし，住所文字列へと導く．提案手法では，キー文字を抽出しないため，キー文字が正しく認識できない場合でも，住所文字列として認識することができる．また，完全一致法を用いることで認識対象を限定することが可能となり，個別文字認識の精度が向上する．本手法の有効性を検討するために，筆者らが独自に採取した 11,586 件のサンプルパターンを使って認識実験した結果 97.57%という高い正解率を得た．

Recognition of Handwritten Characters String of Japanese Address Using a Complete Correspondence Method

Michio Umeda† and Daisuke Honjo†

This paper proposes a recognition method of handwritten characters string of Japanese address using a segmentation candidate lattice and a complete correspondence method. Segmentation processing divides a characters string of handwritten address pattern into small basic segments by labeling processing, and this processing creates a segmentation candidate lattice by combining basic segments. The proposed technique is using a complete correspondence method as a role of post-processing. This method limits a recognition object to the address of the same number of characters as the number of rectangles of segmentation candidate. The complete correspondence method can derive a correct answer, even if it cannot recognize a key character, since a key character is not used. Since the proposed method limits a recognition object, the accuracy of individual character recognition increases. We got 97.57% of correct recognition rate, as a result of having recognized the handwritten address characters string using 11,586 sample patterns which were collected for this study.

1. はじめに

昨今の文字認識の研究分野では，手書きされた文字パターンを高精度に認識する種々の手法が提案されている[1),2)]．しかし，手書き文字パターンは大きさや太さ，形状などが必ずしも一定ではなく，つぶれ字やかすれ字，略字など様々なものも存在する．このような文字パターンを対象とした高精度認識の手法も実現されている[3)]．一方，単独の文字ではなく文字列を認識の対象とすると，行書や草書のように，複数の文字が連結して書かれたり，1つ1つの文字の大きさにばらつきがあるなどの原因により，与えられた文字列パターンから個々の文字を正確に分割して抽出することが非常に困難となる．また，個別文字認識の精度が文字列全体の認識精度に大きく影響を及ぼし，認識率が著しく低下する原因となるなどの問題がある．

文字列認識における文字切り出しには，多くの手法が提案されている．たとえば，ある種の方法で切り出した文字パターンを認識した結果から切り出しの誤りを見つけ出し，1度目とは違った手段で切り出しをやり直す方法がある[4)]．しかしこの場合，誤って切り出されたパターンに対する整合処理での距離値が必ずしも大きくなるとは限らず，正確に誤りを発見することが困難である．これに対して，現在主流となっているものに，候補ラティス法[5),6)]がある．これは，切り出し部で，あらかじめ文字列パターンを小さな切り出し矩形である基本矩形として求めておき，基本矩形とそ

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

の組合せに対する認識結果を候補ラティスとして複数個作成し，正しく切り出される可能性を広げておいて，言語知識を利用した後処理によって正解へ導く方法である．しかしこの方法では，全文字種を対象とした認識を行うために，誤認識の原因となる不要な候補が多く含まれる．したがって，正しく認識した候補だけを正確に選出する手法が必要となる．また，様々な認識結果の組合せを考慮する必要があり，計算量も膨大になるなどの問題がある．

住所文字列のみを対象とした認識では，キー文字と呼ばれる文字を抽出し住所文字列へ導く手法[7),8)]などが提案されている．しかし，このときキー文字が必ずしも正しく抽出され，認識できるとは限らない．また，キー文字がキーとなる場所以外にも存在すると，混乱を招く原因にもなる．一方，計算量を削減するために，候補宛名に優先度を付け大分類する方法もある[9)]．しかし，切り出しに失敗すると後の知識処理に大きく影響し誤認識を生じることになる．

これに対して，本稿ではキー文字抽出をせずに，切り出しに失敗した部分があっても正解へ導くことのできる，完全一致法を用いた住所文字列認識システムを提案する．まず，正確な処理が困難な文字切り出しには，ラティス法を使用し，切り出しに多くの可能性を持たせる．さらに，個別文字認識の特徴抽出法には加重方向指数ヒストグラム特徴[10)]を使用する．また，住所文字列として認識するための後処理法に完全一致法を用いる．

この完全一致法では，まず，切り出し候補ラティスにより，入力された文字列が何文字であるかを推定し，その文字数で構成される住所のみを認識対象として住所辞書から選択する．そして，切り出されたパターンを 1 つ目から順に個別文字認識する．このとき，1 つ目のパターンを認識するときの対象カテゴリには，選択した住所の中で 1 文字目にある文字種のみを採用する．同様に 2 つ目のパターンは，2 文字目の文字種のみと整合処理する．これによって，キー文字を抽出せずに，住所文字列を構成する文字数が一致し，かつ各文字が完全に一致すると仮定して，妥当な住所だけを住所辞書から選出することが可能となる．さらに，対象カテゴリを動的に限定することで，個別文字認識の精度が向上し，切り出しに失敗した部分があっても正解へと導くことができる．

ここでは，日本の住所において，行政区域内に存在する全住所文字列を認識の対象とした．しかし，日本では住所を書くにあたって，必ずしも都道府県名から書き始められるとは限らない．そこで，1 つの住所でも様々な書き方を考慮した住所辞書を作成し，この問題に対処する．

また，提案したシステムを評価するためにサンプルパターンを採取し，認識実験によって，本システムの有効性について検討する．

2. システム概要

手書き住所文字列認識システムの処理の流れを**図 1** に示す．まず，システムの概要について述べる．本システムは，前処理，切り出し，個別文字認識，完全一致法の 4 つの処理で構成される．イメージスキャナによって取り込まれた手書き住所文字列パターンには，雑音が含まれているため，孤立点や極小領域の黒画素は前処理部で雑音として除去する．次に，切り出し部で住所文字列パターンを個々の文字パターンに分割する．ここで得られた文字パターンから特徴抽出し，個別文字認識部で標準パターンと整合処理して，各カテゴリとの距離値を算出する．この距離値を基に，完全一致法によって住所文字列認識結果を導出する．完全一致法は，住所文字列の文字数を基本として，文字列全体に着目して認識する手法であり，住所辞書はその住所を構成する文字数別に用意しておく．そして，与えられた文字列の切り出し結果によって，その文字列パターンの文字数を推定する．このとき得られる文字

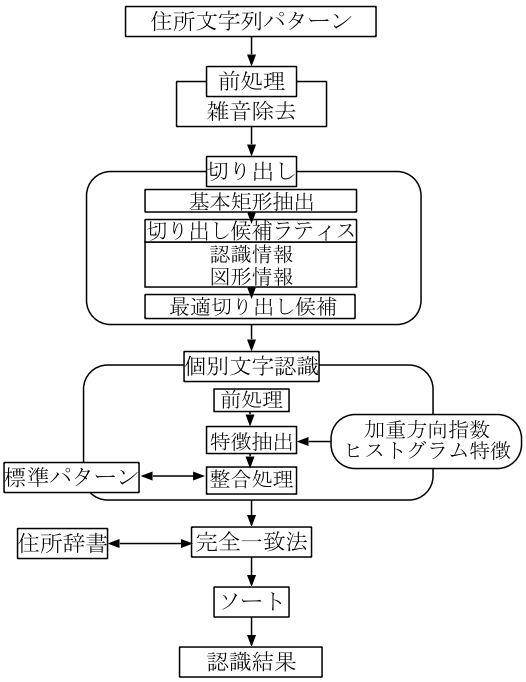

図 1 処理の流れ
Fig. 1 Stream of processing.

数と同じ文字数で構成される住所のみを住所辞書から選出して使用し，文字列パターンとの距離値を算出する．この距離値算出においては，推定された特定の住所文字列のみを対象とすることにより，個別文字認識における対象カテゴリが大幅に削減される．つまり，個別文字認識の精度が多少低くても，認識対象が限定できることにより，文字列としての認識精度を向上させることが可能となる．以下では，これらの個々の処理について述べる．

3. 文字パターンデータ

本システムにおいて使用する標準パターンの作成には，電総研で作成された手書き文字データベースETL9B[11]を使用した．一方，システム評価には，筆者らが独自に収集した手書き住所文字列パターンを使用した．住所文字列パターンについては，本来なら様々な年齢層の筆者によって書かれたサンプルであることが望ましいが，今回はサンプル採取の都合から，一部の年齢層（大学生）の筆者によって書かれたサンプルのみとなった．標準パターンについては，住所文字列サンプルを記入した筆者によって書かれた文字パターンから標準パターンを作成すると，より高い認識精度が期待できると考えられるが，システム評価に汎用性を持たせるために既存のデータベースであるETL9Bを用いることとした．

3.1 標準パターン

個別文字認識に使用する標準パターンの作成には，ETL9Bを用いた．このデータベースには3,036文字種について，それぞれ200個の手書き文字パターンが存在する．ここでは，このうち行政区域内の住所に使用される936文字種のみを使用した．また，標準パターンの作成には，奇数番目の100パターンを使用し，偶数番目のパターンについては，個別文字認識における認識精度の評価用の未知パターンとして使用する．標準パターンは，まず1つ1つの文字パターンから特徴抽出を行い，得られた特徴量をカテゴリごとに平均することで標準パターンとした．

3.2 住所文字列パターン

本住所文字列認識システムの評価用データとして，ドロップアウトカラーで印刷された縦10 mm×横90 mmの枠内に，水性ボールペンによって筆記された住所文字列サンプルを採取した．このサンプルには，日本の行政区域内の住所として書かれる可能性がある，住所の書き方をすべて含めた11,586件の住所文字列が1度ずつ出現する．この住所文字列を1人につき52件，合計223人によって書かれたものを採取した．また，

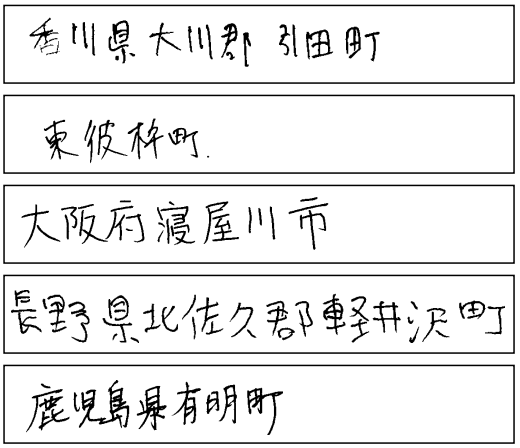

図2 手書き住所文字列パターンの例
Fig. 2 Examples of handwritten address characters string pattern.

これらのサンプルをイメージスキャナによって300 dpiで光電変換した．つまり，1つの住所文字列パターン領域は縦144×横1,064 pixelとなる．これを適当なしきい値で2値化したものを評価用パターンデータとして使用した．収集した手書き住所文字列パターンの例を**図2**に示す．

4. 文字切り出し

入力された文字列パターンを住所文字列として認識するために，まず個々の文字領域ごとに切り出す．自由に手書きされた文字列では，他の文字に喰い込んでいる文字があったり，文字と文字が接触しているパターンなどが存在するため，個々の文字に切り出すことは非常に困難である．この文字切り出しには，候補ラティスを使用する手法や，一度切り出された領域を認識して，誤って切り出されている矩形であると出力した部分を再度切り出し直す手法などがあるが，ここでは，基本矩形から切り出し候補ラティスを生成して，文字らしさの評価値を与える認識情報と形状情報の2つを重みとして使用し，動的計画法によって，最適な切り出し候補を求め，その上位30個の矩形の組合せを切り出し候補とする．

4.1 基本矩形

切り出し部では，まず文字パターンとなる可能性がある，できるだけ小さな矩形を求める．これを基本矩形と呼ぶ．基本矩形を大きくとりすぎると，2つの文字パターンが1つの矩形に入ってしまう場合がある．1つの矩形に複数の文字パターンが存在すると，誤認識となり，再度切り出しのやり直しが必要となる．ここでは，このようなことが起こらないように，基本矩

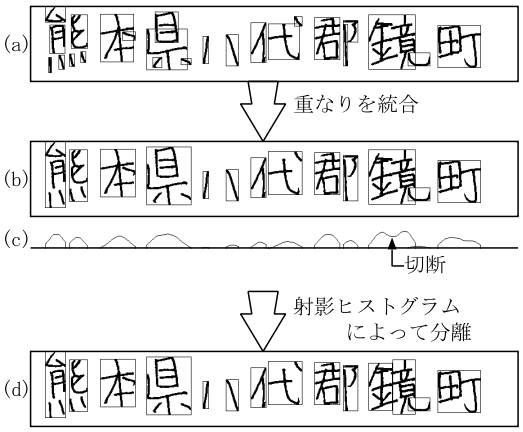

図 3 基本矩形の導出
Fig. 3 Extraction of basic rectangles.

形はできるだけ小さく切り出されるように設定しておく．そして，後述する切り出し候補ラティスによって，この基本矩形をどのように組み合わせて解釈するかを検討する．このときの基本矩形導出の流れを**図 3** に示す．

基本矩形の導出には，まず入力された文字列パターンに対してラベリングを施し，連結成分ごとに外接矩形を求めることで，図 3 (a) のようにパターンを小さな矩形領域に分割して切り出す．次に，得られた矩形に対して統合処理をすることで，ある程度矩形をまとめる．ラベリングにより分割された領域は，ある文字パターンの一部であることが多く，この矩形は互いに重なり合って 1 つの文字パターンを形成していることが多い．そこで，以下のような条件を満たす場合のみ，矩形を統合する．

- 着目している矩形 i と重なっている矩形 j があり，その重なっている部分の横幅が，矩形 i または矩形 j の横幅の 60%以上であるとき．

図 3 (a) で得られた小領域に統合処理を施すと，図 3 (b) のような矩形が得られる．しかし，この統合処理だけでは誤った矩形どうしを統合してしまい，文字パターンとならない矩形が得られる場合がある．

これを防ぐため，統合して得られたすべての矩形に対して，ヒストグラムを利用した分離処理をすることで，再度小領域に分割する．つまり，矩形の黒画素を縦方向に射影してヒストグラムを作成する．次に，このヒストグラムに対して，すべての矩形の平均横縦幅の 1/4 のサイズで移動平均をとる平滑化を 2 回施すことにより，図 3 (c) のようなヒストグラムを得る．このとき，図 3 (c) の “鏡” の部分のように，ヒストグラムに極小値が存在した場合には，その極小値の部分で縦

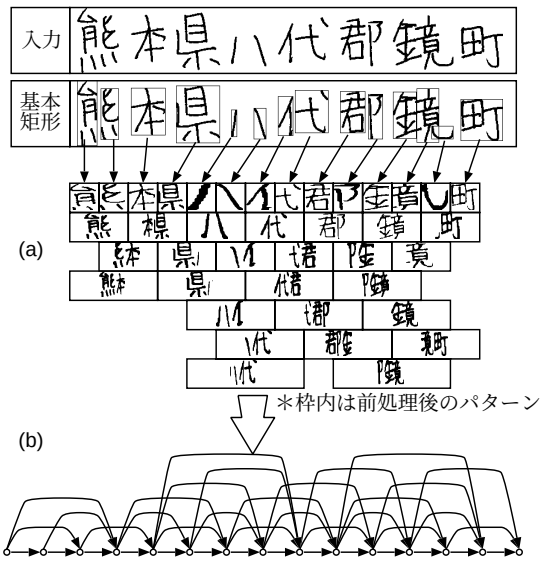

図 4 切り出し候補ラティス
Fig. 4 Segmentation candidate pattern lattice.

方向に切断し，矩形を分離することにすると，図 3 (d) のような矩形が得られる．このような統合，分離の 2 つの処理によって得られた矩形を切り出し基本矩形とする．

4.2 切り出し候補ラティス

得られた基本矩形を基に，切り出し候補ラティスを生成する．まず，基本矩形を複数個組み合わせてできる矩形に着目する．予備実験によれば，この矩形の横幅が，基本矩形の平均縦幅の 2 倍以下であれば，その矩形は文字パターンである可能性がある．そこで基本矩形を組み合わせてできる矩形が基本矩形の平均縦幅の 2 倍以下であることを切り出し候補の条件とする．この条件を満たす基本矩形を組み合わせてできる矩形によって切り出すと，たとえば**図 4** のような文字列パターンに対して，矩形を組み合わせた候補ラティスは図 4 (a) のように求められる．また，これは図 4 (b) のような 2 端子グラフで表現することができる．

次に，この切り出し候補ラティスの各ブランチに対して，重み付けをする．この重みには，文字らしさの評価値を与える．文字らしさの評価には，認識結果からの情報と，文字パターンの形状からの情報の 2 つをパラメータとした．たとえば，i 番目の矩形に j 個の基本矩形を統合した切り出し候補を対象とするとき，認識結果からの情報 R_{ij} は，矩形内のパターンから特徴抽出し，整合処理で標準パターン中の l 番目のカテゴリとの類似度

$$r_l = \frac{\sum_{m=1}^{n} q_m \cdot p_{ml}}{\sqrt{\sum_{m=1}^{n} q_m^2 \cdot \sum_{m=1}^{n} p_{ml}^2}} \qquad (1)$$

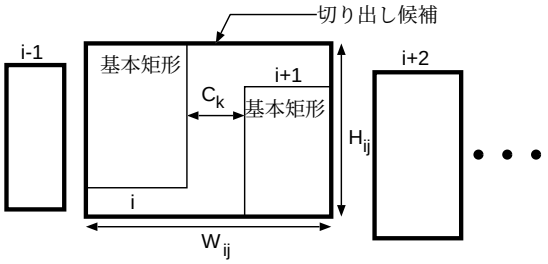

図 5 形状情報
Fig. 5 Pattern figure information.

q_m：矩形の特徴パターン　n：次元数（256）
p_{ml}：標準パターン　l：文字種
を求めておき，

$$R_{ij} = max(r_l) \tag{2}$$

で定義する．つまり，R_{ij} には個別文字認識で 1 位候補となったカテゴリとの類似度を与える．

一方，形状からの情報は，矩形の正方形らしさ Q_{ij} と，矩形内の内部余白の割合 L_{ij} を用いる．Q_{ij} は，**図 5** のように矩形 i に j 個の基本矩形を統合した矩形の横幅 W_{ij} と，縦幅 H_{ij} から正方形らしさを

$$Q_{ij} = \frac{min(H_{ij}, W_{ij})}{max(H_{ij}, W_{ij})} \tag{3}$$

H_{ij}：矩形の縦幅　W_{ij}：矩形の横幅
i：矩形番号　j：組み合わせた基本矩形の数
のように定める．次に，L_{ij} は，まず矩形を統合してできる矩形間の間隔を算出する．たとえば，図 5 のように，矩形 i に対して対象とする複数の基本矩形を統合したときにできる矩形間の幅 C_k の合計を求めて，そこに存在する基本矩形の平均横幅 A をこれで除算し，

$$L_{ij} = min(1.0, \frac{A}{\sum_{k=1}^{j-1} C_k}) \tag{4}$$

C_k：矩形間の幅　A：基本矩形の平均横幅
によって定める．そして，Q_{ij} と L_{ij} の相加平均をもって形状からの情報 P_{ij} と定義する．認識情報 R_{ij} と形状情報 P_{ij} はともに 0.0〜1.0 の値で得られ，1.0 に近いほど 1 つの文字パターンらしいことになる．したがって，これら 2 つの情報を用いて，各ブランチの評価値 F_{ij} を

$$F_{ij} = 1 - \frac{(R_{ij} + P_{ij})}{2} \tag{5}$$

で与える．

4.3 切り出し候補の決定

次に，この切り出し候補ラティスの中で，1 つのパス p に着目する．このとき**図 6** (a) のように，ブランチの評価値 F_{ij} と基本矩形の数 j の積をブランチの重みとし，つまり

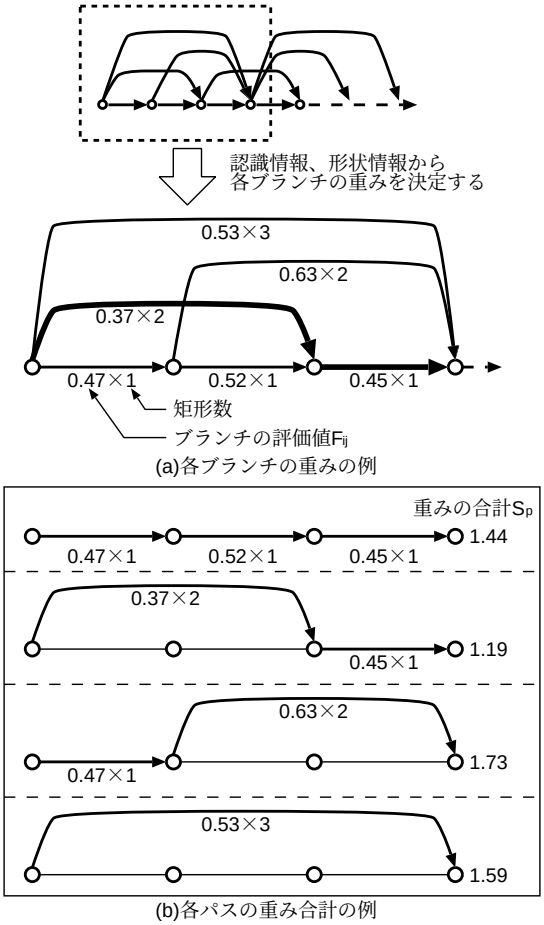

図 6 ブランチの重み算出の例
Fig. 6 An example of branch weight calculations.

$$S_p = \sum_{q=1}^{m} F_{ij} \cdot j \tag{6}$$

i：基本矩形番号 j：矩形を構成する基本矩形の数
p：パス番号　q：ブランチ数
によって，図 6 (b) のようにパス p での重み合計 S_p を求める．切り出し候補ラティスによって生成される基本矩形の組合せは，基本矩形の数によって著しく増大する．そこで，生成される矩形の組合せに絞り込み処理を施す．ここでは，予備実験により，出力する切り出し候補の数は第 30 位までとし，この上位 30 位までのパスを動的計画法を用いて算出し，これを切り出し候補とする．

5. 個別文字認識

切り出し候補であるとされた基本矩形の組合せについて，個別文字認識を行う．個別文字認識は，前処理，

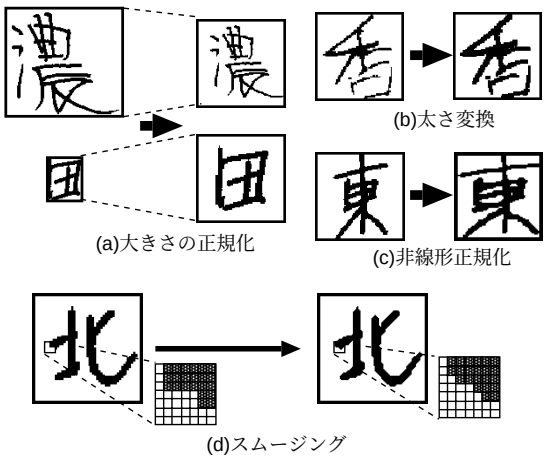

図 7 前処理
Fig. 7 Pre-processing.

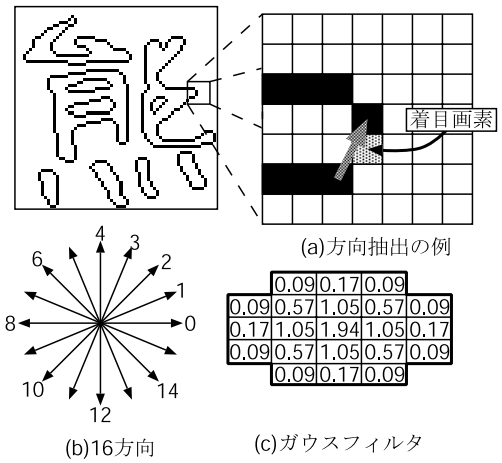

図 8 加重方向指数ヒストグラム特徴
Fig. 8 Weighted direction index histogram feature.

特徴抽出，整合処理の 3 つの部分で構成され，最後の整合処理部で標準パターンと比較して距離値を算出する．与えられた文字パターンには，かすれた文字や大きさの異なる文字，縦長横長の文字など様々なものが存在する．このため，前処理で文字パターンの画一化を図る．この前処理は，大きさを均一化する**図 7** (a) のような大きさの正規化，図 7 (b) のようにかすれ字やつぶれ字を補正するための太さ変換，図 7 (d) のように文字の輪郭部分を平滑化するスムージング，そして文字線の間隔を一定の形状に整える図 7 (c) のような非線形正規化[12]，さらにもう一度スムージングを施すことでパターンを均一化する．次に，加重方向指数ヒストグラム特徴によって特徴量を抽出する．これは，文字パターンの輪郭部分に着目した特徴抽出法で，まず，輪郭線追跡によって**図 8** (b) のような 16 の方向指数を求める．たとえば，図 8 (a) の着目画素において，その 1 つ前の画素から 1 つ後の画素への方向 3 を方向指数として得る．次に，その頻度を 16×16 の領域において計数し，図 8 (c) のようなガウスフィルタを 1 画素おきにかけて 8×8 領域にするとともに，方向を 8 方向に圧縮し，さらに反対方向を同一視することで $8 \times 8 \times 4 = 256$ 次元の特徴量を得る．整合処理には，

$$D_i = \sum_{j=1}^{n} \{max(0, \mid q_j - p_{ij} \mid - a_{ij})\} \qquad (7)$$

q_j：入力パターン　n：特徴の次元数
p_{ij}：標準パターン　i：文字種
a_{ij}：標準パターンの標準偏差

で定義される重みつき絶対値距離を用いた．

6. 完全一致法による住所文字列認識

得られた切り出し候補矩形を基に，個別文字認識のみを用いて住所文字列を認識した場合，切り出し候補矩形が正解の場合であっても，個別文字認識の精度によって，また住所文字列の文字数によって，住所文字列中に誤認識となる文字が出現する可能性が高くなり，住所文字列全体では誤認識となる場合がある．このため，個別文字認識で誤認識となる文字が存在する場合でも正解へと導く何らかの後処理が必要となる．そこで，この後処理的役割をする処理として，完全一致法を用いる．これは，住所文字列に含まれるキー文字に着目するのではなく，住所文字列の文字数と，文字列全体に着目する方法で，キー文字などが誤認識となり，うまく抽出できなかった場合でも正しい住所文字列へと導くことができる．また，個別文字認識で認識できない文字が存在した場合でも，住所文字列として認識することができる．以降に，この完全一致法で使用する住所辞書と，完全一致法の処理内容について述べる．

6.1 住所辞書

日本の住所を書く場合，同じ住所でも，たとえば “大阪府寝屋川市” と書くことや “寝屋川市” と書くことがあるように，都道府県から書き始める場合や，市郡区などから書き始める場合がある．また，郡名などが省略されて書かれることもある．そこで，このような不規則な書き方にも対応するため，**図 9** のように住所として書かれうるすべての書き方を考慮した住所辞書を作成する．また，完全一致法は住所文字列の文字数に着目しているため，これらの住所群を文字数でソートしておき，住所辞書とする．

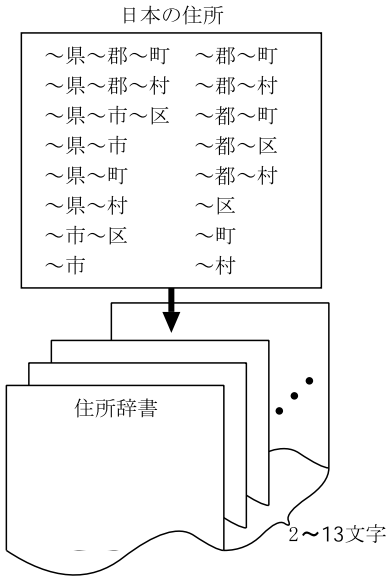

図 9 住所辞書
Fig. 9 Japanese address dictionary.

6.2 完全一致法

完全一致法では，与えられた文字列パターンより得られた切り出し候補矩形の数によって使用する住所辞書を限定する．つまり，切り出し候補矩形を構成する矩形数と同じ文字数の住所のみを住所辞書から選択し，これらの住所とのみ比較する．このとき，与えられた文字パターンと住所辞書中の住所 DIC_i との全文字での平均距離を算出する．そして，得られた平均距離値でソートし，最小となる住所を第 1 位認識候補として出力する．たとえば，**図 10** の場合，まず第 1 位切り出し候補となった矩形に着目する．この矩形は，7 個の矩形で構成されているため，“青森県十和田市”，“岩手県大船渡市”，などの 7 文字の住所辞書を用いる．まず，矩形 [1-1] と標準パターンの “青” との距離値を算出する．次に，[2-3] と “森” についても同様に距離値を算出していき，すべての矩形との距離値を算出した後，その合計を矩形数で除算することで平均距離を算出する．また，“岩手県大船渡市”，“大阪府寝屋川市” についても同様に平均距離を算出する．次に，第 2 位の切り出し候補矩形についても同様の処理を行う．そして，すべての切り出し候補矩形に対して住所辞書との平均距離を算出した後，さらにこれをソーティングし，最小のものとなった “大阪府寝屋川市” を住所文字列認識結果として出力する．

7. 結果と考察

本認識システムを評価するため，個別文字の認識実験と住所文字列サンプルデータ 11,586 件の手書き住所文字列パターンを用いた個別文字認識のみでの認識実験，および本システムでの認識実験の 3 つの実験を行った．また，ここでは日本の住所として，標準パターン作成に ETL9B を使用したため，JIS 第 1 水準漢字とひらがなのみで構成されている行政区域内の住所を対象とした．

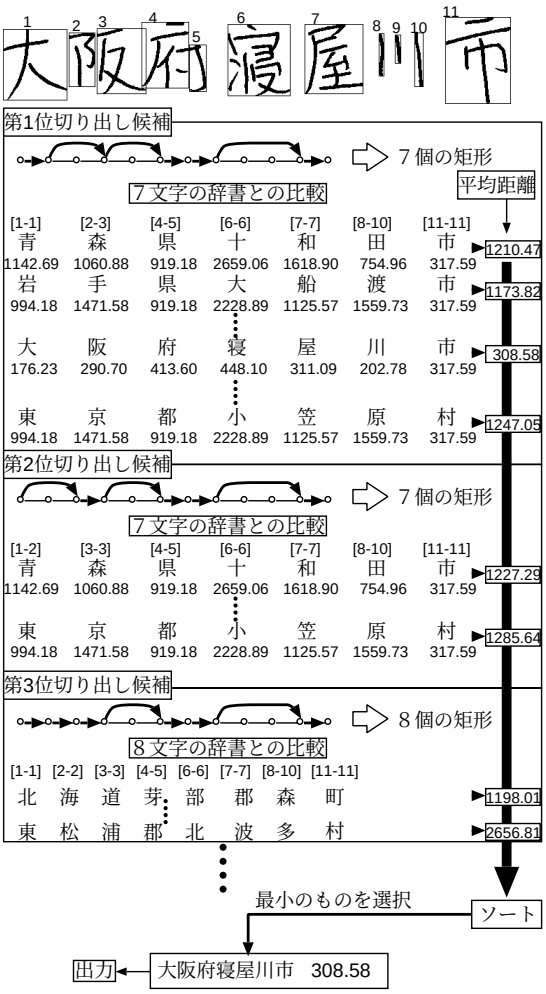

図 10 完全一致法
Fig. 10 A complete correspondence method.

7.1 個別文字認識結果

本システムで使用している個別文字認識の性能評価を目的として，標準パターンの作成に使用したのと同じ ETL9B を用いて，個別文字認識実験を行った．個別文字認識実験で使用した文字パターンは ETL9B に登録されている 3,036 文字種のうち，行政区域内の住所に使用されている漢字とひらがなの合計 936 文字種を対象カテゴリとした．また，標準パターンの作成に

表 1 累積個別文字認識率
Table 1 Accumulative accuracy rate of individual character recognition.

候補順位	1	2	3	4
学習パターン [%]	95.92	97.54	97.97	98.23
未知パターン [%]	95.09	97.13	97.69	97.99
候補順位	5	10	20	30
学習パターン [%]	98.42	99.09	99.90	99.95
未知パターン [%]	98.23	98.98	99.83	99.89

表 2 個別文字認識のみでの住所文字列認識結果
Table 2 Recognition result of address characters string only using an individual character recognition method.

文字数	正解件数 [件]	合計件数 [件]	認識率
2 文字	47	65	72.31%
3 文字	1,338	2,453	54.55%
4 文字	182	495	36.77%
5 文字	81	198	40.91%
6 文字	1,201	4,107	29.24%
7 文字	345	1,389	24.84%
8 文字	64	283	22.61%
9 文字	259	1,555	16.66%
10 文字	131	866	15.13%
11 文字	15	156	9.62%
12 文字	2	17	11.76%
13 文字	0	2	0.00%
合計	3,665	11,586	31.63%

使用した奇数番目の 100 パターンを学習パターンと定義し，残りの 100 パターンを未知パターンと定義して評価実験を行った．個別文字認識実験の結果を**表 1** に示す．

この結果より第 1 位認識率として，未知パターンに対しても 95.09%と高い認識率を得ることができた．また，認識候補として第 30 位までをとると，99.89%ときわめて高い認識率が得られた．

7.2 個別文字認識での住所文字列の認識実験

本システムで後処理的役割として使用している完全一致法の有効性を確認するため，個別文字認識のみを用いた住所文字列の認識実験を行った．この実験では，切り出し候補ラティスが生成された段階で，第 1 位切り出し候補となった矩形の組合せについて個別文字認識を実行し，1 位認識候補となった文字をそのまま連結して文字列とし，住所文字列の認識結果とした．このときの文字数別の認識結果を**表 2** に示す．

この場合，文字数が増えるに従って誤認識となる文字パターンが含まれる可能性が高くなるため，住所を構成する文字数の増加によって認識率に大きな影響が出ることが予想できる．事実，表 2 からも分かるように，個別文字認識のみを用いて住所文字列を認識すると，文字数が多くなるほど認識率の低下が大きくなることが確認できる．全体の認識率では 31.63%であった．

表 3 完全一致法を用いた住所文字列認識結果
Table 3 Recognition result of address characters string using a complete correspondence method.

文字数	正解件数 [件]	合計件数 [件]	認識率
2 文字	62	65	95.38%
3 文字	2,350	2,453	95.80%
4 文字	491	495	99.19%
5 文字	196	198	98.99%
6 文字	4,054	4,017	98.71%
7 文字	1,365	1,389	98.27%
8 文字	282	283	99.65%
9 文字	1,515	1,555	97.43%
10 文字	826	866	95.38%
11 文字	145	156	92.95%
12 文字	16	17	94.12%
13 文字	2	2	100.00%
合計	11,304	11,586	97.57%

7.3 完全一致法を用いた住所文字列の認識

本システムを用いた住所文字列サンプルパターン 11,586 件に対する認識実験結果を**表 3** に示す．この結果からも確認できるように，本システムでは，文字数が増加しても認識率は低下せず，高い文字列認識率が得られることが分かる．ちなみに，全体の認識率は 97.57%が得られた．なお，住所文字列の文字数別に見ると，件数の少ない 13 文字を省くと，4 文字のときに認識率は 99.19%と最も高くなっている．文字数が少ない場合は，完全一致法の効果が少なく，文字数が多い場合は個別文字認識での距離値が大きいパターンが含まれる割合が多くなる．このことから認識率もわずかに低下している．個別文字認識では正解とならないが，完全一致法を用いた住所文字列認識システムでは正しく認識できたサンプルの例を**図 11** に示す．

なお，1 つの文字列の認識に要する処理時間は，住所文字列の長さ，基本矩形への分割個数，検索する住所の件数などによって大きく異なるが，全住所データで平均すると，Pentium3（866 MHz）のパソコンを使用して約 15 秒であった．

また，本システムで誤認識となった原因を調べると，最適切り出し候補の選出のときに誤った切り出し候補を選んでおり，正解である矩形が上位 30 位以内に入っていない場合が多かった．この誤認識となったサンプルパターンとその出力結果を**図 12** に示す．図 12 (a) の例では，切り出し矩形 [16-19] の “仙” が誤認識の原因となっている．これは，“仙” の偏と旁が大きく離れて書かれているため，形状情報の余白の割合が大きくなり，偏と旁を別々の文字パターンであると判断し誤

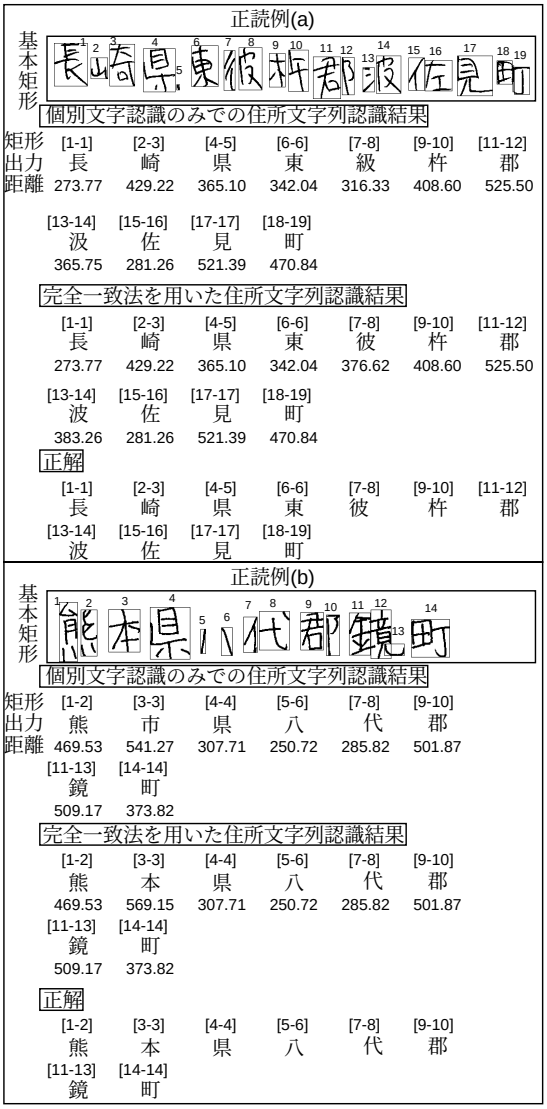

図 11 正読パターン例
Fig. 11 Examples of correctly recognized pattern.

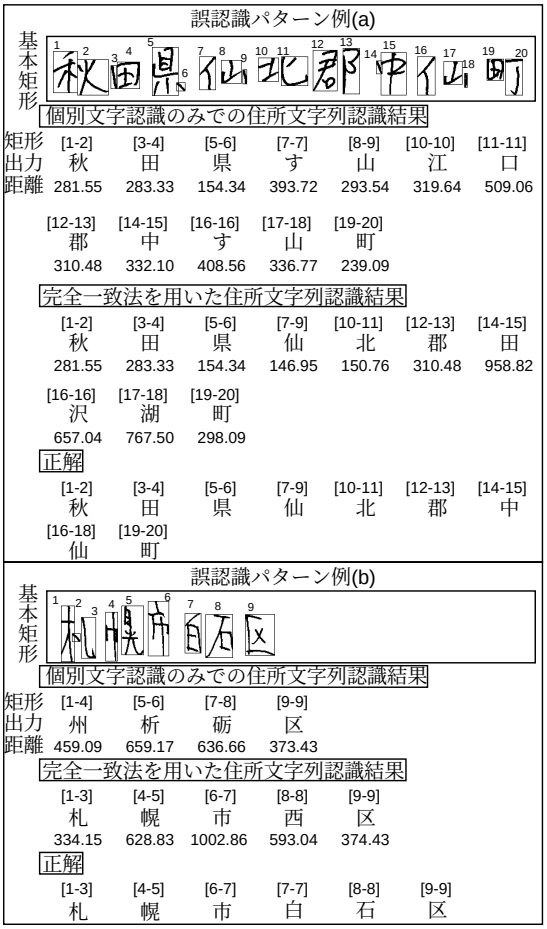

図 12 誤認識パターン例
Fig. 12 Examples of misrecognized pattern.

認識となっている．また (b) では，切り出し矩形 [6-6] の “市” と [7-7] の “白” を 1 つの矩形としてしまい誤認識となっている．この例では，矩形 [7-7] の “市” のパターンが極端に縦長に書かれているため，形状情報での正方形らしさを判断する際に，正方形とはほど遠い形をしているので文字らしくないと判断してしまい，切り出し候補として正解の矩形が選択されなかったことが原因である．

8. おわりに

本稿では，住所文字列の切り出しから認識までを一連の処理として実行する，手書き住所文字列認識システムを提案した．切り出しにラティス法を使用することによって多くの可能性を持たせ，かつ切り出し候補の矩形数によって対象カテゴリを動的に変更する完全一致法を用いることによって，97.57%という高い認識率を得ることができた．また，都道府県名以外で書き始められた住所も住所辞書に加えることによって，これらにも対応することができた．本システムでは，キー文字の抽出を行っていないため，キー文字が正しく抽出できない場合でも，正解を得ることができる．しかし，切り出した基本矩形の組み合わせ方を決定する際に誤った組合せを選択している場合が多く見られた．このため，今後は文字切り出し部分の高精度化が必要であると考えられる．また，ここでは住所文字列として，行政区域内の住所のみを対象としたが，住所の行政区域以外の地名部分についても検討していく必要がある．さらに，完全一致法を用いた場合，文字数と住所の件数によって認識率に影響が出ることや，住

所の検索にかかる処理時間についても検討が必要である．

謝辞 本研究を行うにあたり，手書き住所文字列データの採取に協力していただいた皆様に深く感謝いたします．また，手書き文字データベース ETL9B を提供してくださった電子技術総合研究所に感謝いたします．

参 考 文 献

1) 郭 軍，孫 寧，根元義章，佐藤利三郎：整形変換を用いた手書き文字データベース ETL9B の高精度認識，信学論，Vol.J76–DII, No.5, pp.1015–1022 (1993).
2) 加藤 寧，安倍正人，根元義章：改良型マハラノビス距離を用いた高精度な手書き文字認識，信学論，Vol.J79–DII, No.1, pp.45–52 (1996).
3) 矢田勝啓，鶴岡信治，木村文隆，三宅康二：加重方向指数ヒストグラム法のつぶれ文字への対応，信学技報，PRU90–128, pp.21–26 (1990).
4) 仲林 清，北村 正，河岡 司：あいまい用語検索を用いた高速枠なし手書き文字列読み取り方式，信学論，Vol.J74–DII, No.11, pp.1528–1537 (1991).
5) 村瀬 洋，若原 徹，梅田三千雄：候補文字ラティス法による枠無し筆記文字列のオンライン認識，信学論，Vol.J68–D, No.4, pp.765–772 (1985).
6) 村瀬 洋，新谷幹夫，若原 徹，小高和己：言語情報を利用した手書き文字列からの文字切り出しと認識，信学論，Vol.J69–D, No.9, pp.1292–1301 (1986).
7) 鈴木雅人，孫 寧，阿曽弘具：キー文字駆動型地名推論による手書き宛名認識アルゴリズム，信学技報，PRU95–5, pp.33–40 (1995).
8) 濱裕治郎，梅田三千雄：2 段階個別文字認識を用いた手書き住所文字列の認識，信学技報，PRMU96–150, pp.71–79 (1995).
9) 徳本一崇，鈴木雅人，加藤 寧，根元義章：候補あて名の優先度付けによる高速大分類法を用いた手書きあて名認識システム，信学論，Vol.J84–DII, No.1, pp.83–92 (2001).
10) 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，信学論，Vol.J70–D, No.7, pp.1390–1397 (1987).
11) 斎藤泰一，山田博三，山本和彦：JIS 第一水準手書き漢字データベース ETL9 とその解析，信学論，Vol.J68–D, No.4, pp.757–764 (1985).
12) 吉田収志，本郷保夫：非線形正規化を用いた背景特徴パターンマッチング法による手書き漢字認識，信学技報，PRMU92–34 (1992).

(平成 13 年 3 月 14 日受付)
(平成 14 年 11 月 5 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，映像情報メディア学会，画像電子学会各会員．

本庄 大介

昭和 52 年生．平成 11 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課程前期情報工学専攻入学，現在在学中．文字認識，特に手書き住所文字列の認識に関する研究に従事．電子情報通信学会会員．