手書き住所読取りにおける街区住所知識処理方式

4R－8

下村秀樹
NEC情報メディア研究所

1. はじめに

手書き住所の読取りでは，一般に画像から文字領域を切り出して認識し，候補文字の組合せと住所に関連する知識・規則とを住所知識処理（後処理）で照合して，結果を決定する．しかし，手書きの住所には，フリーピッチかつ字形が多様である，隣接文字の入組みや接触が多い，ノイズや汚れが多い等，文字の切出しや認識にとって困難な問題が多い．このため，正しい読取りのために切出し・認識候補中の正解含有率を高めようとすると，出力候補数の増加が避けらない．その結果，住所知識処理には，多くの候補の組合せの中から，効率よく適切な読取り結果を判定することが求められる．特に，丁目以下の街区部分は，同形や類似形の文字が多いことなどもともと候補が多くなりやすい要因があるので，この問題は重大である．

本稿では，知識照合の重複処理を減少させた効率のよい街区住所知識処理として，(1)文字領域のカテゴライズ，(2)木構造への候補展開，(3)値範囲の検証，の3つのステップから成る方式を提案する．また，ノイズかどうかわからない文字領域に対する扱いが容易であるといった，提案方式の柔軟性にも言及する．

2. 街区住所知識処理の一般的問題

2.1 候補増加の理由

街区の読取りでは，文字の接触や入組み以外に次のような理由もあり，多くの切出し・認識候補を扱わないと住所を正しく読むことが難しい．

(1)縦書き漢数字の切出し位置が複数ある

例えば，縦書きで横棒が3本並んだ場合，「一一一」「二一」「一二」「三」の4通りの切り方が考えられる．

(2)使用文字に同形や類似形の文字が多い

例えば，横書きの「一」「－」「～」，縦書きの「１」「｜」など，単独では区別の困難な文字群がいくつかある．

(3)住所知識が町名ほど強力でない

地名部分では，例えば「川崎市宮？区」のように一部で認識不良が起きた場合，地名の階層構造の知識から「川崎市宮前区」を推測することができる．

A street number reading algorithm for handwritten address reading
Hideki Shimomura
NEC Corporation

これに対して街区では，認識不良が起きたてもその数値を前後関係から一意に推定することはできない（数値の範囲なら可能）．したがって，切出しや認識候補に正解が含まれていないと，読取り不能となる．

この他，ノイズらしい領域の採否も，複数の候補が出る事例の1つと考えることができる．

2.2 従来方式による処理の重複

街区で使われる主な住所知識を次に示す．

・街区文字列の生成規則
　－ 数字列，区切り記号の繰り返しである
　－ 使われる数字の字種は同じである
　－ 使われる区切り記号の種類は同じである
・町名ごとの街区の値範囲

従来の街区知識処理の一般的方式では，切出し・認識の候補を組み合わせて読取り結果の候補を作り，それを順次，上記の知識と照合する方式が採られていた．しかし，この方式には，次のような処理の重複があり効率が悪い．例えば，次の候補は最後の1文字だけが異なる．

「１－２－８」
「１－２－３」

これを独立に検証すると，「１－２－」の部分についての検証処理が重複してしまう．また，別の観点からの処理の重複として次の例がある．

「二の三の８」
「二の三の３」

この2つの候補は，号の部分だけがアラビア数字なので規則に合致しない．この場合，問題は「８」「３」といった数値ではなく，最後がアラビア数字だということである．したがって，両方が同じ文字領域から認識されたのなら，どちらか一方を検証すれば十分である．

関連文献[2]では，文字切出し領域の関係を木構造に組み立てる方式が提案されているが，処理の重複を減少させるという観点からの議論はない．

3. 効率のよい街区住所知識処理方式

2節の議論から，次のようなステップでの街区知識処理を提案する（図1）．

ステップ1：文字領域に対してその認識結果の文字のグループを示す記号（カテゴリ記号）を付与する．

ステップ2：数字と区切り記号の並び方規則を接続規則で表現し，規則を満たすカテゴリ記号列の候補を木構造に生成する．

ステップ3：生成した候補を文字列に戻し街区の値範囲を検証する．

ステップ1のカテゴリ記号付与は，2節で議論した同一字種への重複検証を避けるために行う．カテゴリは街区での文字の役割に応じて定義し，同じ文字領域に同一カテゴリの文字が2つ以上あっても，カテゴリ記号は1つしか付けない．

ステップ2は，同一位置への検証の重複を避けるためである．文字列を先に作り独立に検証してゆくのではなく，木構造を部分共有しながら候補を生成してゆくので，共有部分の検証処理が1回で済む．区切り記号が一貫性を持って使われているかどうかなど，単に連続する記号間の接続関係では表現できない規則は，いったん候補を作ってから残ったものについてチェックする．

ステップ3では，カテゴリ記号の木構造として表現されている候補を，再度文字列に戻して，町名が限定する値の範囲を検証する．

4. ノイズへの柔軟な対処

2節では詳細にはふれなかったが，街区の読取りにおいてノイズへの対処は重要である．街区で使われる文字は単純図形なので，ノイズが数字や区切り記号に読まれて，思わぬ誤読を引き起こすことがある．

住所の一部かノイズかわからない文字領域があった場合，住所知識処理は両方の解釈で読取り候補を作成し，どちらがより住所らしいかを判定することが望ましい．これに対して，本稿で提案する木構造の候補表現は，ある領域を使用する場合の候補と使用しない場合の候補を，木の枝分かれとして，他の候補と全く同様に表現できる（図2）．もちろん，その後の街区数値範囲の検証も，他の候補と区別なく行える．すなわち，木の生成規則を若干拡張するだけで，提案方式の特長を損なうことなく，ノイズに対処できる．

5. おわりに

本稿では，処理効率のよい街区住所知識処理方式を提案した．本アルゴリズムを実現し実験を行った結果，処理の効率化は確認できた．

その一方で，提案方式では多くの候補を検証できるが故に，偶然知識に合致する候補による誤読が発生するという問題が目立つようになった．これに対しては，文字領域の形状や配置情報を利用した尤度判定，知識処理から文字認識へのトップダウン再検証処理などのアイデアも提案し，検討中である[1]．この点も含めた精度の評価が今後の課題である．

謝辞

本研究に関連して議論していただいたNEC産業オートメーション事業部の山内氏，NEC情報メディア研究所の福島氏に感謝する．

参考文献

[1]下村他:手書き住所読取りにおけるパタン処理と連携した住所知識処理方式,第50情処全大,4D-1

[2]大井:住所読取りにおける丁目・街区認識方式,信学技報,NLC92-26

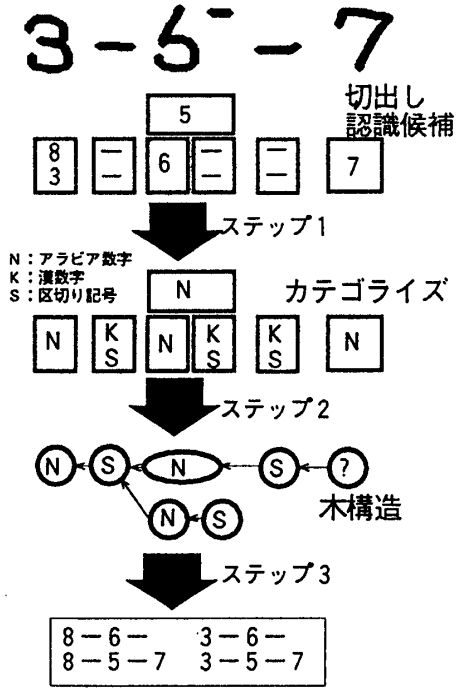

図1　提案する街区住所知識処理方式

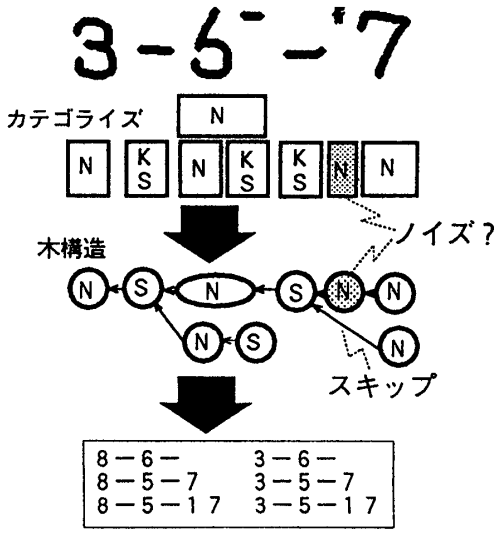

図2　ノイズらしい文字領域への対処