文字間の遷移確率による文字認識知識処理

4H－11

郡司 圭子　葛貫 壮四郎　桂 晃洋　横田 登志美　三浦 雅樹
（株）日立製作所　日立研究所

1．背景と目的

ペン・コンピュータでは、操作性向上のために、文字認識率を向上させることが重要である。これまで、1文字毎のパターン認識をもとに文字認識していたが、［タ（カタカナ）／夕（漢字）］のような同形文字は、1文字では識別が困難なため、文字の前後関係から識別する後処理を開発することにした。その際、ペン・コンピュータの遅いCPU・少ないメモリでも快適な速度を維持するために、2文字間の遷移確率を利用した。

2．遷移確率による後処理の課題

2文字間の遷移確率による後処理は、単語照合など他の一般的な後処理に比べ、辞書容量・処理量が少ない特徴がある。しかし、次のような問題がある。

(1) 文字列が長いと、処理量が文字列長の累乗で爆発する

(2) 2文字間の関係がベースのため、後処理対象の候補が多いと、悪影響を及ぼすことがある

3．候補絞り逐次最尤パス選択方式

3．1 処理フロー

前記問題を解決するために、候補絞り逐次最尤パス選択方式を立案した。図1を用いて、以下、動作を説明する。

[1] 手書き文字を入力する

[2] 入力パターンと辞書パターンをマッチング(パターン認識)し、候補文字と距離値を得る(距離値は、入力パターンと辞書パターンの差分値で、小さいほど似ている)

[3] 後処理との融合のため、画数非依存になるよう、距離値を正規化する

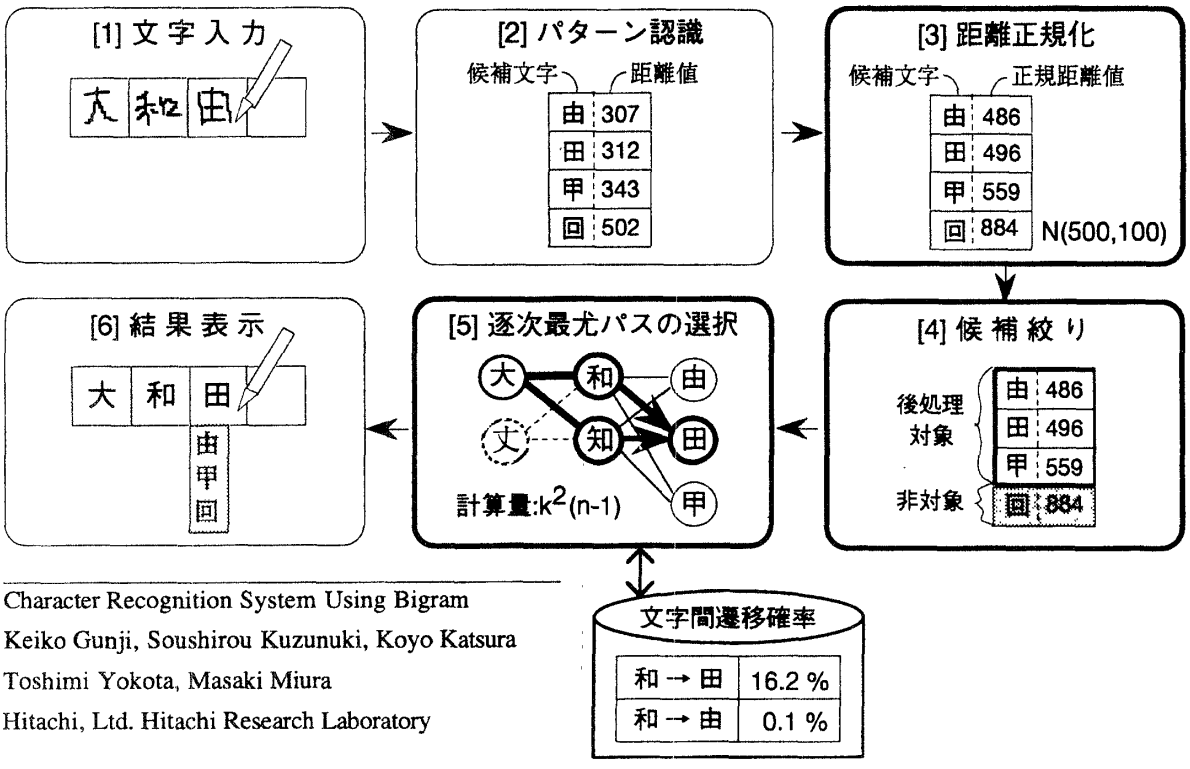
[4] 正規化距離値を元に、後処理対象を正解の可能性が高い候補文字のみに絞り込む

[5] 候補文字を組み合わせ、出現確率が高い文字列(最尤パス)を、候補文字毎に逐次求める

出現確率(大和田)＝出現確率(大)
×遷移確率(大→和)×遷移確率(和→田)

[6] 結果を表示する

図1 候補絞り逐次最尤パス選択方式



Character Recognition System Using Bigram
Keiko Gunji, Soushirou Kuzunuki, Koyo Katsura
Toshimi Yokota, Masaki Miura
Hitachi, Ltd. Hitachi Research Laboratory

3.2 逐次最尤パスの選択

文字認識結果のうち、最初に第一候補を表示することを考えると、後処理では最尤候補パスのみ正確に求めれば十分である。また、2文字間の遷移確率をベースにする場合、文字列の最後の文字の出現確率は、直前の文字のみに依存して決まる。以上2点の性質を利用して、候補文字毎に逐次最尤パスを求め、後処理を高速化する手法を立案した。本方式によれば、計算量が$k^2(n-1)$［k：候補文字数, n：文字列数］と文字列数の倍数で済むため、一般文章のような長い文字列にも適用が可能である（一般の全数組み合わせ方式ではk^nの計算量が必要）。

3.3 距離値の正規化

後処理との融合のために、画数により異なるパターン認識の距離値を、画数非依存に正規化した。その方法を述べる。まず、正規化前の距離値を調べるために、正解文字の距離分布を画数毎に測定したところ、各画数とも正規分布に従っていることが判明した。図2は、画数と前記距離分布（正規分布）の平均・標準偏差の関係を示す。1画～5画、6画、7画～20画ではパターン認識方式を変えているため、距離値の平均値が異なるが、標準偏差はほぼ一定である。距離値分布が正規分布に従っているので、各画数で正規分布の平均と標準偏差を揃えることにより、距離値を正規化した。ここでは、距離値の精度を保つため、平均=500、標準偏差=100に正規化した。

3.4 候補絞り

パターン認識の結果では正解文字である可能性が余り高くないが、たまたま前後の文字同士の繋がりがよいために、後処理で誤認識の候補文字を正解候補と入れ替えてしまうことがある。この悪影響を軽減するため、パターン認識の結果をもとに、後処理対象の候補を絞る方式を検討した。候補絞り方式は、正解候補の取りこぼしが少なく、正解以外の候補をなるべく排除するのが理想である。各種の方式を評価したが、第一候補との距離差による方式が、前記条件を最も満たした。図3は、第一候補との距離差と、後処理対象に含めたい正解文字と、含めたくない文字の割合を示す。図から分かるように、誤認識の場合でも、後処理対象に含める必要のある正解文字は、距離差が小さい範囲に集中している。そこで、第一候補との距離差を基準に、後処理対象に含めるか、否かを判定することにした。判定しきい値を最適化するため、しきい値を変化させて、氏名データの正認識数を調べた(図4)。その結果、しきい値140で正認識数が多いため、この値を判定基準にした。

4. 評 価

図5に、50人分の氏名（漢字･カタカナ）の認識率を、(a)パターン認識のみ、(b)後処理のみ、(c)本方式（候補絞り＋後処理）に分けて示す。また、参考のため、(d)後処理後の理想値（パターン認識で10候補内に入っているものを正解と見なした認識率）も示す。後処理のみ(b)では、パターン認識のみ(a)に比べて認識率が向上しているものの、後処理の理想値(d)には達していない。しかし、本方式(c)では、候補絞りによって後処理の悪影響を押さえたため、ほぼ、後処理の理想値(d)に近い認識率を達成した。以上により、本方式の効果が確認できた。