5P-6

文字相互の接続関係を用いた文字認識高速化の検討

佐藤 哲司，津田 伸生，松尾比呂志
ＮＴＴ電気通信研究所

1．はじめに

手書き文字等の認識処理では、多次元の特徴ベクトルを用いたパタン照合によって、手書き特有の変形を許容して多数の文字種を識別する為に、パタン照合処理量が認識速度を制限する要因となっている。

本稿では、連続した文字の認識過程において、文字相互の接続関係に基づいて照合対象文字種を限定し、パタン照合処理量を低減するアルゴリズムを提案し、特許等の分野を限定した文章での有効性について述べる。

2．照合対象文字種選択による高速化法

図１に、照合対象文字種を直前に認識した文字から選択し、パタン照合によって文字認識する方法を示す。本方法では、あらかじめ日本語文章における連続した文字間の接続関係を抽出し、ある文字に対してその次に出現する確率が高い文字のコードを接続文字種テーブルに用意しておく。

図１の例では、「文字相互の接続・・・」を入力文章として、第２番目の「字」を認識する過程を示している。

(1) 照合文字種選択 ： 直前に認識した「文」から接続文字種テーブルを検索し、次に認識する文字として出現する確率の高い文字を選択して対応する標準特徴ベクトルＹを、パタン照合回路に入力する。（図１①－③）

(2) 入力処理：入力文字パタンから、ノイズ除去や正規化等の前処理と特徴抽出を行ない、入力特徴ベクトルＸを求めてパタン照合回路に入力する。（図１④）

(3) パタン照合 ： (1)(2)で求めた特徴ベクトルＸ、Ｙの間でベクトル間距離を求める。パタン照合で求めた距離値を判定し、認識結果「字」を出力する。　（図１⑤－⑥）

以上、提案したアルゴリズムでは、認識済の文字から接続文字種テーブルを検索し、次に認識する文字の候補を選択してパタン照合を行う。

入力特徴ベクトルとの距離値が所定のレベル以下となる標準特徴ベクトルが得られなかった場合には、照合対象文字種の範囲を広げてパタン照合を継続する。

パタン照合によって得られた連続文字の認識結果から、接続文字種テーブルを更新することによって、次に認識する文字の照合対象文字種を限定し、認識処理における平均的なパタン照合処理量を削減する。

3．有効性の検証と適用

3.1 日本語文章の評価

提案したアルゴリズムの有効性を検証するために、分野を限定した特許出願文書を対象に、文字間の接続関係を評価した。対象とした文章は、表１に品種ごとの出現率を示したＲ１－Ｒ４とそれらを合成した文章「全文」の５種類である。

この表から、品種毎の出現率は文章によらずほぼ一定であり、漢字とカナを合わせた文字種の出現率が全体の約５０％であるといえる。

表１　文字の出現率(%)

文例 ：	Ｒ１	Ｒ２	Ｒ３	Ｒ４	全文
記号	10.9	10.5	10.8	8.4	10.0
数字	3.9	2.7	3.2	3.1	3.3
英字	1.6	0.9	0.9	1.4	1.3
かな	35.1	35.9	33.9	31.3	33.8
カナ	16.7	14.5	12.7	17.1	15.6
漢字	31.8	35.5	38.5	38.7	36.0

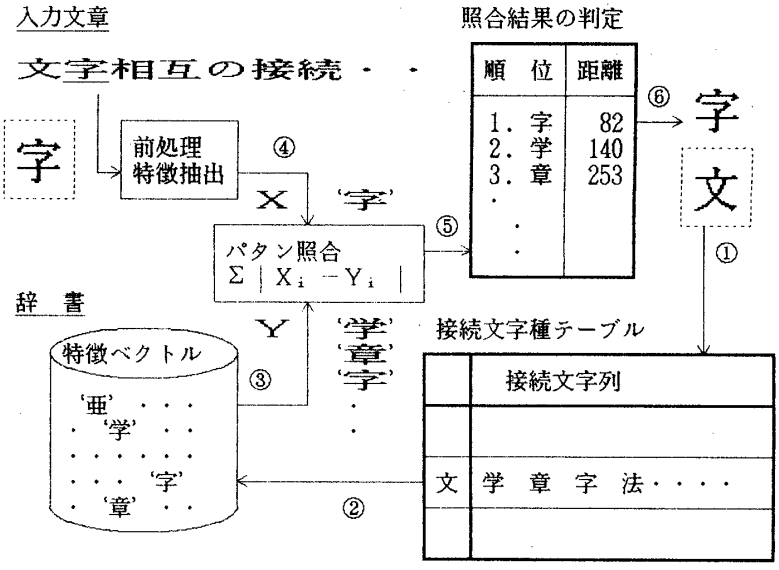

図１．照合文字種選択による文字認識高速化法

Pattern Matching with Reference-category Expectation
Tetsuji SATOH, Nobuo TSUDA, Hiroshi MATSUO
NTT Electrical Communications Laboratories

(1) 接続文字種数の分布： 接続文字種数とは、文章中において、ある文字に対してその次に接続する文字の種類数である。図２に接続文字種数の分布を示す。この図では、接続文字種数で分類した文字種の出現率を、文章中に出現する文字種数で規格化して表している。この結果、接続文字種数の分布は文章によらずほぼ一定であり、出現率の累積値から全文字種の９０％は接続文字種数が２０以下となることが判る。

(2) 接続関係によるグループ化： 図３に、文字を６品種に分類した場合の品種間の接続関係を示す。文字の品種を以下のグループA、Bに分類することによって、表２に示す関係が成立する。グループBでは、１字種あたりの出現率が小さく、平均接続文字種数も小さいことから、文字認識における照合対象文字種の選択が、照合計算量削減に有効である。

グループA＝記号,数字,かな

グループB＝英字,カナ,漢字

表２ 文字の出現率と接続関係

	グループA	グループB
出現率	約 0.5	約 0.5
字種数	約 200	約2000
出現率 ／ １字種	大	小
平均接続文字種数	大	小

3.2 文字認識装置への適用

図４は、接続文字種を接続頻度順に並び換えた場合に、上位k種までに正解となる文字が含まれている確率（含有率）を表している。この結果、グループBでは上位２０位迄にほぼ正解文字が含まれると期待できるが、グループAでは正解文字の含有率が９０％程度であり、照合対象文字種を選択して実用的な認識精度を得るには、正解文字の含有率が低く問題である。以上のことから、認識精度を低下させずに照合対象文字種を選択する方法として、グループBのみ接続文字種テーブルを検索して特定の文字と照合を行ない、グループA

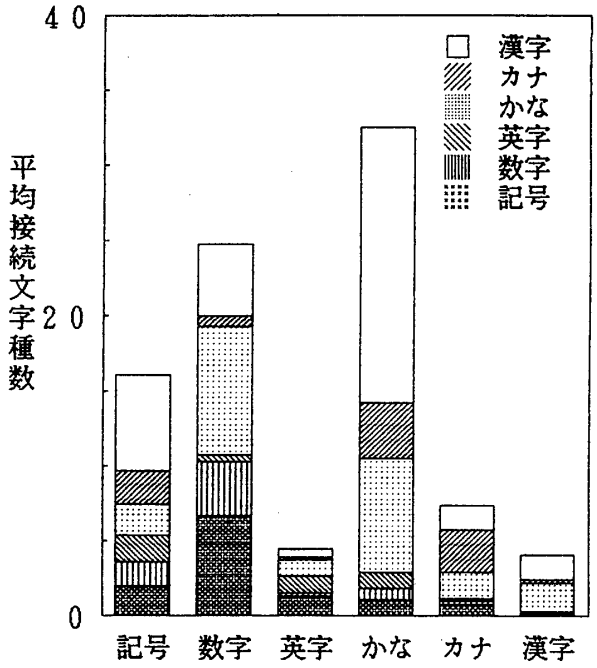

図３．文字相互の接続品種

の文字については常時照合処理を行なうのが適当といえる。

この場合の平均的な照合対象文字種数は、

グループA（２００種）＋B（２０種）

となり、全文字と照合する場合と比較してパタン照合演算量を１／１０に低下できる。

４．まとめ

文字認識等における照合対象文字種を、直前に認識した文字から絞り込むことによって、パタン照合演算量を低減する方法を示した。本方法を特許出願文章に適用した結果、認識率をほとんど低下させることなく、照合演算量を約１桁低減できることを明らかにした。本方法は、認識対象文字を常用文字種以上に拡大する場合、あるいは汎用システムで特定分野の文字認識を行う場合に有効である。

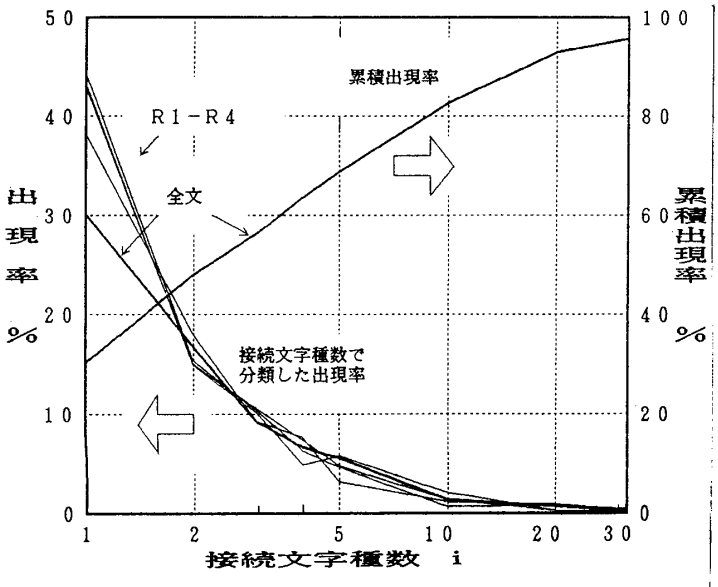

図２．接続文字種数の分布

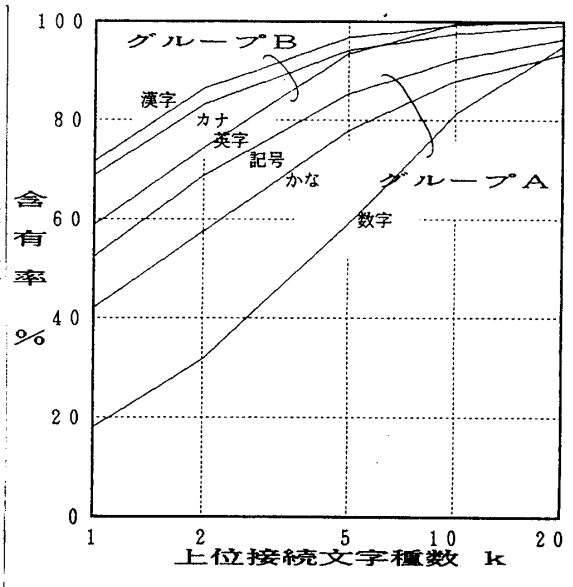

図４．正解文字含有率