

# 等分割可変点近似による続け書き文字認識

6Q-1

郡司 圭子、横田 登志美、三浦 雅樹、葛貫 壮四郎、

(株)日立製作所 日立研究所

安江 司

(株)日立製作所 オフィス事業部 旭工場

## 1. 目的

簡単に操作できる手書き入力システムが注目されている。発表者らも、1980年代より、ペン入力インタフェースに着目し、操作性向上技術の開発に取り組んでいる[1]。スムーズな入力を行うには、応答性がよく、認識率が高く、続け書き文字などの変形文字も認識できることが必要である[2]。本報告では、続け書きに対応した文字認識について述べる。

## 2. 文字認識の開発課題

図1に、従来の文字認識の問題点を示す。画数の少ないひらがなと多い漢字では別の対策を取る必要があるが、ここでは前者について述べる。ひらがなでは、誤認識原因のうち続け書きによるものが半数近くを占めている。これは、従来の近似方式は文字を1画あたり6点で近似しており、図の「を」・「な」等の続け書き文字では、手書きパターンでは違いがあるのに、近似精度が不十分で近似パターンは区別がつかなかったためである。そこで、高精度近似方式の開発を行うことにした。

## 3. 高精度近似方式の開発

高精度近似方式として各種の方式を検討したが、等分割可変点近似が最適であると判断した。図2に等分割可変点近似処理の概要を示す。近似点数Nを増加しながら等分割で入力パターンの近似を行い、十分な近似精度が得られた時点で近似パターンを出力する。図の例では、「を」に対して9点で必要な近似精度が得られた様子を示している。等分割可変点近似を実現するには、十分な近似精度が得られたことを判定する方式と、辞書毎に近似点数が異なるために起こる距離値の差を正規化する方式を確立する必要がある。

## 4. 近似精度判定方式

形状の複雑さに応じた近似点数にするために、近似判定に用いる近似精度として、近似パターン長 $L'$ のストローク長 $L$ に対する割合 $(L'/L)$ を採用した。仮に、近似判定パラメータを $a$ とすると

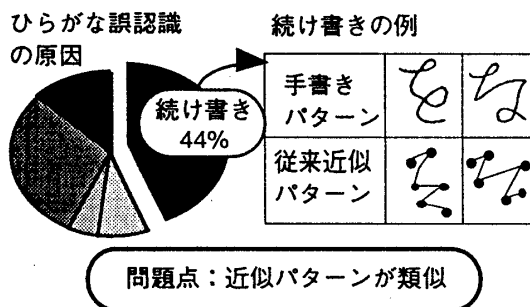


図1 従来の文字認識の問題点

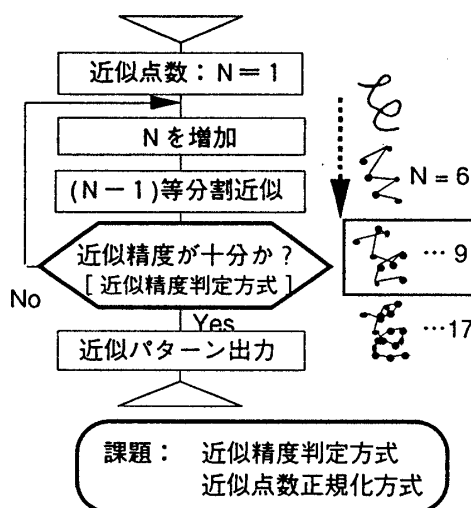


図2 等分割可変点近似方式

$1 \geq (L'/L) \geq a$ ならば近似精度が十分であることを表し $(L'/L) \geq a$ となるまで近似点を変える。

この近似判定パラメータでは、相似形のパターンならば大きさにかかわらず同じ近似点数に近似してしまう。しかし、同じギザギザした形状であっても、短いストロークの場合は、手ブレによるノイズなので、ノイズを除去するために少ない点数で近似しなければならず、逆に長いストロークでは、パターンの特徴であるため、特徴を残すように多い点数で近似する必要がある。この問題は、ストロークの長さに応じて近似判定パラメータを最適化するすることにより解決できる。

図3のグラフは、識別が難しいサンプルで適正に

近似されたものについて、近似パターン長  $L'$  とストローク長  $L$  の関係を示す。ストローク長  $L$  が長いものと短いものでは  $L'$  と  $L$  の関係で傾きが違っているので、次のように近似判定パラメータを修正した。

$$0 < L \leq 150 \text{ のとき } (L' + 5) / L \geq 1.0$$

$$L > 150 \text{ のとき } (L' - 5) / L \geq 0.9$$

すなわち、近似した線よりも上側の領域内ならば、近似精度が十分であることを表している。この近似判定パラメータを用いると、短いストロークでは手ブレによるノイズを除去し、長いストロークではパターンの特徴的な形状をうまく保持できた。

### 5. 近似点数正規化方式

前記等分割近似を用いると、文字毎に近似点数が異なる。一般に近似点数が増加すると、パターン間の距離値も増加する。このため、候補文字のなかで近似点数が異なる辞書パターンがあると、候補文字の間で距離値に不公平が生じ、誤認識の原因となる。このような誤認識をなくすために、距離値を正規化する必要がある。

図4は6点近似と17点近似の距離値の差、および9点近似と17点近似の距離値の差の分布を示したものである。この図から、6点近似と17点近似の距離値の差は平均値20.3付近に、9点近似と17点近似の距離値の差は平均値9.6付近に分布していることが分かった。

以上の分析結果をもとに、マッチング時の近似点数により、平均値に近い値（6点近似のとき20、9点近似のとき10）を、求めた距離値に加算する最高近似点数正規化方式を立案した。これにより、6点近似や9点近似で得られた距離値は17点近似でマッチングを行ったとき得られる距離値に正規化されるため、近似点数の違いによる距離値の不公平を改善できる。

### 6. 評価

前述した方式の妥当性を、1、2画の文字に適用して評価した。図5に、続け書き文字データおよび、ラフに書かれた数字データの認識率を示す。等分割可変点近似方式の適用で、従来識別が難しかった続け書き文字の認識率が9%アップした。近似点数正規化方式の適用では、少ない近似点数の数字に効果が現れ、認識率が2%アップした。また、図示しないが、辞書パターンのチューニングにより数字の認識率はさらに9%アップしている。

### 7. 今後の課題

ひらがなの「へ」とカタカナの「へ」などの同形

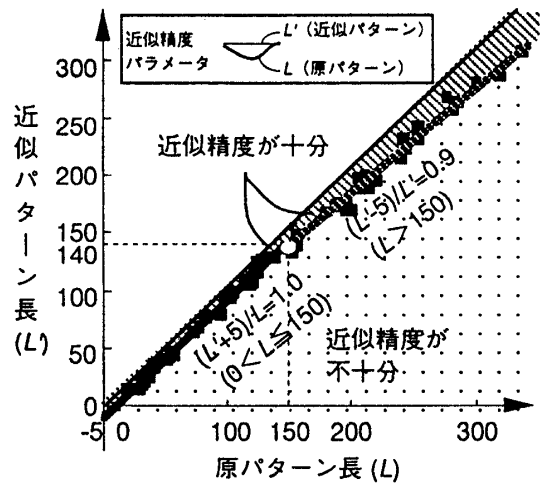


図3 最適化した近似精度判定方式

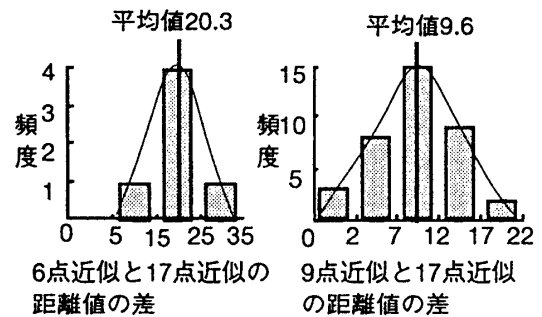


図4 近似点数の違いによる距離値の差の分布

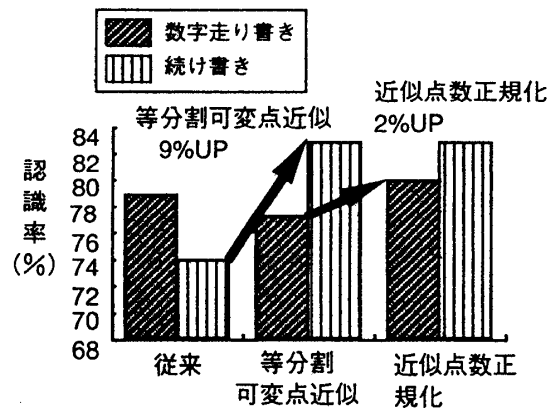


図5 開発技術の適用による認識率向上

文字は1文字だけでは識別できないので、前後関係よりこれらを識別できるような後処理を開発したい。

### 8. 参考文献

[1] 福永：「紙のような計算機を目指して」, 電子情報通信学会, EID91-20, 1991.6.27  
 [2] Ying-Jian Liu, Li-Qin Zhang, Juwei Tai : A New Approach to On-line Handwritten Chinese Character Recognition : IEEE/2th ICDAR pp192-195 , '93.10.20