オンライン続け字認識における類似文字同定方式

3C-3

鎌田　洋　　石垣　一司　　植田　郁子
(株) 富士通研究所

1．はじめに

我々は，続け字も認識できるオンライン文字認識方式として，「特徴点逐次対応法」(1)(2)(3)を先に提案した。特徴点逐次対応法は文字の続き方に関する知識を文字毎に記述し，この知識を基に入力文字と辞書文字の特徴点を対応づけて距離計算を行い，認識結果を求めるものである。

画数の多い漢字については実用レベルの認識率を達成した。しかし，非漢字などの画数の少ない文字については，類似文字間で特徴点が対応づくことが多いため，それほど高い認識率を達成していなかった。そこで，今回，類似文字間の誤読を減少させて認識率を上げるための類似文字同定方式を開発した。本稿では，本方式の内容と認識性能について報告する。

2．認識処理のながれ

次の手順により，入力文字を認識する。

① 特徴点逐次対応法により，入力文字と辞書文字の特徴点を対応づけて，認識候補を求める。

② 類似文字同定方式により，認識候補文字で順位の高いものから同定条件を適用し，最初に合格した認識候補文字を認識結果とする。

3．同定条件の構成

本方式では類似文字の誤読ペア毎に条件を設けている。すなわち，文字2が文字1に誤読するとき，これを防ぐため，条件【文字1，文字2】＝｛文字1が満たし，文字2が満たさない条件｝を設ける。そして，文字1の同定条件を，条件【文字1】＝AND ｛条件【文字1，文字2】｜文字2は認識対象カテゴリに属する｝と構成した。

条件【文字1】を直接に設けず，条件【文字1，文字2】により構成したのは次の理由による。

① 誤読：文字2⇨文字1と条件【文字1，文字2】が1対1に対応するので，誤読の原因究明が容易であり，同定条件の作成が容易にできる。

例えば，「マ」や「了」は「ア」に誤読しやすいが，条件【ア】を設けるには，条件【ア，マ】と条件【ア，了】を別々に考えればよい。

② 認識対象を，数字や英字などの特定の字種カテゴリに限定できるとき，全ての字種を認識対象とする場合よりも，条件【文字】を小さく構成できる。このため，性能と速度を最適化できる。

先の例において，カタカナのみを認識対象とする時は，条件【ア】から条件【ア，了】を外すことができる。

③ 辞書の拡張や変更も，条件【文字1，文字2】の単位で追加・削除することで，容易にできる。

4．続け字の同定方法

続け字も同定できるように，特徴点逐次対応法で得られた入力文字と辞書文字の特徴点の対応づけを利用して同定条件を構成した。次の手順による。

① 辞書文字の特徴点に対応づいた入力文字の特徴点を求める。

② 上記の入力文字の特徴点を基準として，辞書文字について記述した同定条件を入力文字に適用する。

例えば，「オ」（カタカナ）の1画目と2画目を続けて入力すると，図1のように，辞書文字「才」（漢字）と特徴点が対応づく。

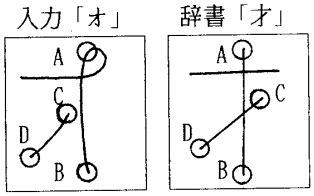

図1．入力と辞書の特徴点の対応例

このとき，「才」（漢字）の同定条件で，「オ」（カタカナ）と識別するための，条件【才，オ】を

① 線分AB，CDの交差が一定以上のとき，合格

② そうでないとき，不合格

と構成できる。この同定条件は入力文字が続け字かどうかに関係なく有効に作用する。

Identification of Similar Characters for Online Cursive Character Recognition
Hiroshi KAMADA, Kazushi ISHIGAKI, Ikuko UEDA
FUJITSU LABORATORIES Ltd.

5．手書文字の曖昧性への対処

手書文字の変動のために，どの文字を書いたのか，人間が見ても不明な文字が非常に多くある。例えば，図2の手書文字は「ナ」（カタカナ）であるか「十」（漢字）であるか判別できない。

そのような文字は同順位の認識候補として積極的に残し，文法情報等により後で判別すべきである。

図2．曖昧な文字

そこで同定条件の論理を一般によく考えられる2値から3値に拡張し，手書文字の曖昧性に対処するようにした。

即ち同定条件の結果として，①合格．②不合格．の他に③判別不能．を設けて，同定条件を適用した結果が③である時，判別不能の文字を同順位の認識候補とするようにした。

この他に，「へ」（ひらがな・カタカナ）のように外見が全く同じ字種は最初から同等に扱っている。

6．開発方法

入力文字と辞書の特徴点の対応の安定性や特徴量の分布を図形画面で対話的に調査できる実験ツールを開発して用いた。特徴点逐次対応法により学習用文字を認識させて生じた誤読ペアのうち，誤読数の多いものに対して重点的に同定条件を作成した。

学習用文字は頻度の高い2188字種とし，各字種40字である。常用漢字レベル2010字種と非漢字（数字・英大文字・ひらがな・カタカナ）178 字種からなる。

7．開発結果

合計４１３の同定条件を約２４Ｋステップの Ｃ言語プログラムとして開発した。表1に使用特徴と同定条件における利用率を示す。1条件あたり使用した特徴は3種類までであり，2種類までの特徴で作成できた条件は全体の９８．８％であった。

表1．使用特徴と利用率

特徴	利用率	特徴	利用率
画の交わり	３５．８	位置	７．０
曲がり	２０．３	長さ	６．５
画の続き方	１３．６	辞書の制約	１．２
距離	１２．６	ループ	１．２
画の有無	９．４	同じ形	１．０
画の向き	８．５	ハネ	０．５

8．認識実験

学習用文字と未学習文字について実験を行った。未学習文字は学習用と同じ字種で，各字種60字である。

認識辞書は，上記の字種とJIS 第1水準の漢字を含む3498字種とした。

画数の多い漢字は，もともと特徴点逐次対応法による誤読が少ない。ここでは，非漢字と5画以下の漢字（207 字種）と6画以上の漢字で特徴点逐次対応法で誤読があった字種の認識結果を表2，3に示す。

表2．学習用文字の認識結果

	非漢字	～5画	6画～(235種)
同定前	８４．６	９１．２	９２．１
同定後	９６．８	９７．２	９５．５
累積5位	９８．８	９８．６	９７．７

表3．未学習文字の認識結果

	非漢字	～5画	6画～(356種)
同定前	８３．８	８９．８	９０．５
同定後	９４．８	９４．７	９２．１
累積5位	９８．０	９７．３	９５．０

平均同定時間はミニコンS-3500で0.01秒であった。同定条件を設けた誤読ペアについて，解消できた誤読の百分率は，学習用文字で９７．０％，未学習文字で９１．６％であった。

9．考察

本同定方式が学習用文字だけでなく，未学習文字に対しても効果があることが表2，3から分かる。

同定条件を設けた誤読ペアに対しては，その誤読をほぼ解消できている。同定条件を追加すれば認識率を累積認識率にさらに近づけることができると考える。同定条件を追加するには，表1の特徴を1～2個組み合わせるだけでよく，容易にできる。

１０．まとめ

続け字も認識できる特徴点逐次対応法における類似文字同定方式を開発した。入力文字と辞書文字の特徴点の対応づけを基準として同定条件を設けることにより，続け字をも同定可能にした。類似文字ペア毎に同定条件を設けることにより，着実な認識率向上を計った。さらに，同定条件に3値論理を導入し，手書きの曖昧性に対処した。この結果，非漢字などの低画数の文字の認識率を大きく向上できた。

今回の開発により，オンライン文字認識装置の実用化に大きく近づけたと考える。

【参考文献】
(1)石垣他：昭和61年後期，情処全大，予稿　1633
(2)石垣他：昭和62年後期，情処全大，予稿　1953
(3)大橋他：昭和63年後期，情処全大，予稿　1640