4D－3

オンライン手書き文字認識のための標本パタンデータベースの作成

秋山勝彦，中川正樹
東京農工大学工学部電子情報工学科

１．はじめに

パタン認識システムの評価には、標本パタン数に対して何パーセント正認識が得られたかという「認識率」がよく使われるが、研究機関ごとに別々の標本パタンを用いたのでは、認識率による評価はあまり客観的意味を持たない。

現在、オフライン文字認識においては ETL という事実上の標準となっているデータベースが存在するが、オンライン文字認識においてはこれといった標準がない状況である。

したがって、標準となるような大規模のオンライン手書き文字パタンデータベースが必要となっている。

しかし、ただ単に大量の文字パタンを集めただけでは、十分な認識辞書の整備や認識エンジンの評価を行なうことはできない。文字パタンデータベースは、できるだけ種々雑多な文字パタンを含む必要がある。ペン入力の実用化を考え、認識方式の研究に役立てるためには、現実的な字体変形が加わった文字パタンを一人当たりで相当量収集する必要がある。そのためには、文章形式の文字セットでパタンを集めたり、方々で筆記者を募ったりといった工夫をする必要がある。

本稿では、文字だけでなく様々なオンラインパタンに対応したデータベースの仕様を提案し、このオンラインパタンデータベースでのデータの収集，配布などについて報告する。

A Sample Pattern Database for On-line Recognition of Handwritten Characters
Katsuhiko Akiyama, Masaki Nakagawa
Tokyo University of Agriculture and Technology

２．パタンデータベースの設計

オンラインパタンデータベースの設計は、配布や保守，汎用性を考えて、シンプルさを重視した。

もっともシンプルなモデルとしては、単に文字パタンが羅列しているようなシーケンシャルなモデルが考えられるが、これではデータのランダムアクセスや編集が困難である。

そこで、ページ番号の付いたノートのようなモデルを採用した。ページ毎にインデックスが付いているので、Indexed Page Data Base （以降 IPDB と呼ぶ）と名付けることにする。

ノートの各ページには自由にデータが書き込めるようになっている。今回の目的はオンラインパタンデータの収集であるから、主にオンラインパタンデータが書き込まれることになるが、ビットマップやテキストデータを書き込んでも構わない。

ページデータの追加，削除，変更などは、インデックステーブルを変更して適当な位置にデータを書き込むことで行なえるので、データの編集も容易である。

３．データ形式

IPDB のファイル上でのデータ形式は次に挙げる項目から成り立つ。

（１）ヘッダストリング
（２）ページ数
（３）インデックステーブルへのインデックス
（４）インデックステーブル
（５）各ページのデータ

項目（１）～（３）は、この順序で決まった位置に決まったサイズで書き込まれているが、

項目（４）と項目（５）は位置とサイズが不定である。項目（４）は項目（２），（３）によってサイズと位置が示され、項目（５）は項目（４）によってサイズと位置が示される。

項目（１）はファイルが IPDB のファイルであることを表す識別情報である。これによって、他のデータファイルとの混同を避けると共に、バージョンを確認して上位互換性を保つことができる。

最初のバージョンではページ 0 への書き込みは行わない。このページは、どのページに何が書き込まれているのかを表すページ割当て情報のために使う予定であるが、しばらくはオンラインパタンとその付加情報にしか IPDB を利用しないので、ページ割当ては固定である。

４．IPDB ライブラリ

IPDB ライブラリとは、IPDB 仕様のデータファイルを取り扱うための C 言語によるライブラリである。最初のバージョンでは、オンラインパタンデータだけを取り扱う。

オンラインパタンデータには、ページセット，ページ，ストローク，ポイント毎に付加情報が記録できる。IPDB の最初のバージョンでは、「ページセット＝ファイル」である。

付加情報のフォーマットはデータベース管理者が定義する。したがって、筆点ごとに筆圧データを記録したり、ストロークごとにタグナンバーを付けたりすることは自由である。

一つの付加情報フォーマットは付加情報項目の集合である。付加情報項目は属性として項目名とデータタイプ，配列数を持つ。

第三者から IPDB ファイルを受け取った場合、この付加情報項目を調べることで、筆点データ以外にどのような情報が記録されているのかを知ることができる。

筆点データはストロークの始点を除いて、前点との座標の差分を取って、もっとも少ないバイト数に収まるように記録される。

５．配布について

我々は、現在までに学内，学外から筆記者を募り、文章形式で約 11000 文字× 27 人分のオンライン手書き文字パタンデータを収集した。

当面、この収集したデータはデータベースの拡大に協力する機関だけに配布する予定である。

しかし将来、IPDB 仕様のオンラインパタンデータベースが標準となり得る程の量と質を持ってきたときには無償，用途無制限の公開を考えている。

オンライン手書き文字パタン収集ソフトと IPDB ライブラリは協力機関に配布する。データ収集ソフトは MS-WINDOWS for Pen 用である。

配布するオンライン手書き文字パタン収集ソフトは、任意の文字列ファイルを指定すると、その文字列と升目を表示して文字パタンを収集するように作成した。

我々が収集に用いた文字列ファイルは、朝日新聞の記事から抜き出した 1500 文字種を含む約一万文字分の文章列と JIS 第一水準の残り 1700 文字種の羅列で構成されている。

６．おわりに

本稿では、オンライン文字パタンデータベースを構築するための IPDB 仕様の設計とオンライン文字パタンデータの収集，配布について述べた。

現在、我々のオンライン手書き文字パタンデータベース作成プロジェクトに 10 機関の協力が表明されており、新たな参加も歓迎している。