

1M-6 枠無しオンライン手書きパターンデータベース

木原伸次郎 中川正樹 福島貴弘

東京農工大学工学部

1. はじめに

手書きパターンに対する認識手法を高度化し、かつ公正な評価をするためには、大量の標本パターンが不可欠である[1]。オンライン手書き文字認識において、筆記枠ありのパターンのデータベースに関しては既に利用されている[2]。我々は、これを基に筆記枠無しオンライン手書きパターンのデータベースを設計した。この枠無し手書きパターンデータベースの概要について述べる。

2. パターン認識におけるデータベース

近年タブレット等ペンベースの入力装置の普及により、自由にペンで入力した文字を認識する“枠無し文字認識”のニーズが高まっている。また、そうした枠の無い環境下においては、文字だけでなく絵や数式なども同時に書かれる場合も想定することができる。そこで、枠を取り払い、文字はもちろん可能な限り自由な環境下で筆記したパターンのデータベース—枠無し手書きパターンデータベース—が必要となる。

3. データベースの設計

枠無し手書きパターンデータベースの設計は基本的には枠ありのそれに準拠しており、拡張性や汎用性のある基本的なデータモデルとして、ページ番号の付いたノートのようなデータモデル—Indexed Page Data Base (IPDB) を利用している。IPDBではデータをページ単位で管理し、各ページのデータは特に規定しない限り任意のバイト列データとなっている。各ページのオンラインパターンデータの形式は、ストロークデータが順に格納されており、このオンラインパターンデータには、ページセット、ページ、ストロークのそれぞれに対して別々のフォーマットの付加情報を設定することができる。

4. ストロークのタグ付け

枠ありパターンデータベースでは、1つの枠に対して一つの文字しか書くことはないため、枠内に筆記された文字パターンとその文字コードを対応させればよい。しかし、枠無しパターンの場合は枠がないため、どのストロークにどの文字コードが対応しているかをシステム側から自動登録することは難しい。しかし、ストロークが何の文字を形成しているかをストローク付加情報に与える必要がある。このとき与える情報をタグ情報と名づけ、これをストローク付加情報として保存する。このタグ情報はタグ番号とタグコードからなり、文字を構成するストロークに対し、一つのタグ番号とその文字（タグコード）をそれぞれのストロークに付加情報として与える（図1）。ただし、続け文字など二つ以上の文字間のストロークが繋げて書いてある場合は、これを一つのタグコードとして扱う（図2）。

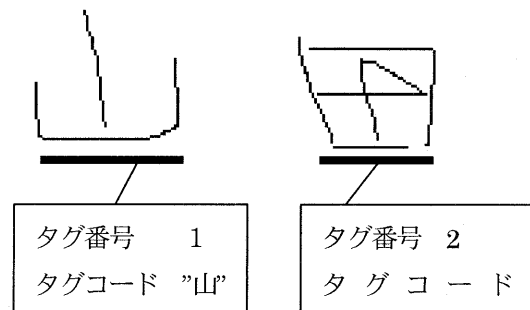


図1 タグ番号とタグコード

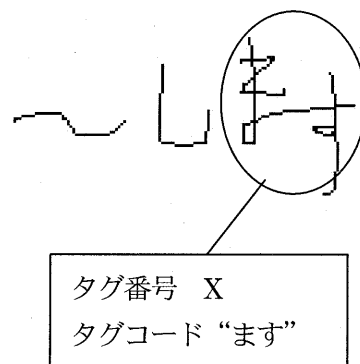


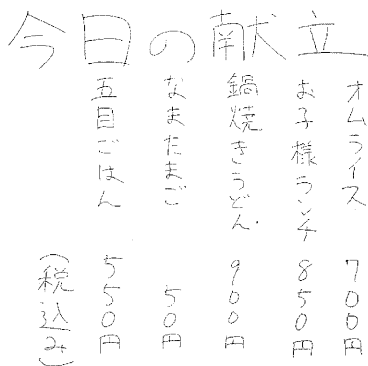
図2 続け文字の場合のタグコード

A database of on-line handwritten patterns without any writing box, grid or guideline.
Shinjiro Kihara, Masaki Nakagawa, Takahiro Fukushima
Tokyo Univ. of Agriculture and Technology, 2-24-16
Naka-cho, Koganei, Tokyo, 184-8588, Japan

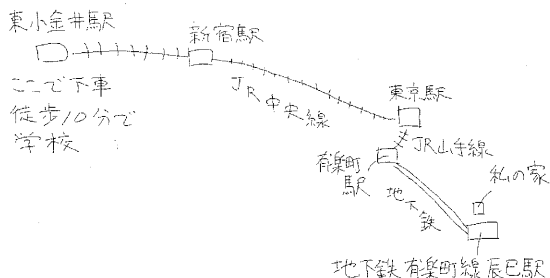
これらのタグ情報は将来的に認識システムの評価など行うときに必要になる。

5. 枠無し手書きパターンデータベースツール

ツールは大きく分けて、パタン収集ツールと、付加情報入力ツールとがある。パタン収集ツールはデータベースに標本パタンを登録するためのツールである。画面サイズは表示一体型タブレットの画面サイズといったサイズの固定している。また、収集する文字パターンは、例文を事前に印刷などして用意しておく。パタン収集ツールの典型的な画面を図3に示す。



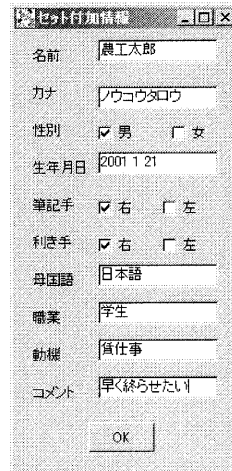
(a)文字入力中の画面の例



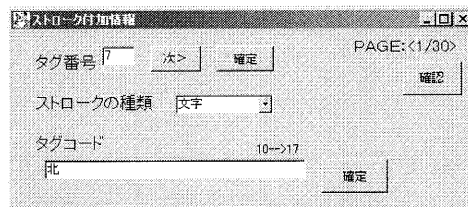
(b)文字及び図入力中の画面の例

図3 パタン収集ツール

付加情報入力ツールは、セット付加情報・タグ情報などのストローク付加情報を与えるためのツールである。ダイアログを表示し、これらの情報を入力・編集できる。これらの付加情報ツールの典型的な画面を図4に示す。



(a)セット付加情報入力ダイアログ



(b)ストローク付加情報入力ダイアログ

図4 付加情報入力ツール

6. 枠無し手書きパターンデータの収集

枠無しオンライン手書きパターンデータは、単純に大量の手書きパターンを集めるだけでなく、丁寧な手書きパターン、雑な手書きパターンなど多種多様な文字パターンが含まれている方が望ましい。そのため文章形式の文字(あるいは図形)のセットで収集を行っている。

なお、現時点で枠無しオンライン手書きパターン収集は、筆者の所属研究室である東京農工大学中川研究室において行っている。

7. おわりに

枠無しオンライン手書きパターンデータベースの設計・開発を行った。今後は使用書や環境を整え、研究室内だけでなく外部からも標本となるパタンの収集にあたり、データベースの規模を大きくしていく予定である。

【参考文献】

[1] 斎藤泰一, 山田博三, 山本和彦, JIS 第1水準手書漢字データベース ETL9 とその解析, 電子通信学会論文誌'85/4 Vol.J68-D No.4, pp.757-763 (1985)
 [2] <http://www.tuat.ac.jp/~nakagawa/ipdb/>