

デジタルペンによるスタイルフリーな手書きパターン標準化 マネジメントシステムの提案とデータベースの構築

Proposal of a Standardization and Management System for Unconstrained On-line Japanese Handwriting Patterns Written by Digital Pen and Collection of Their Databases

織田 英人†
Hideto Oda

寺田達也‡
Tatsuya Terada

徳野 淳子††
Junko Tokuno

伊藤 禎宣‡
Sadanori Ito

中川 正樹‡
Masaki Nakagawa

1. はじめに

近年 Anoto ペン[1]を代表に, DigiMemo や Technote, PC Notes Taker など, 紙に書き込んだ内容を読み取ることができる様々なデジタルペンが開発されている。これらは, アナログとデジタルを結ぶデバイスとして注目されているが, 本格的な普及には至っていない。その一因として, 次のようなことが考えられる。

- (1) 各デジタルペンが独自のデータフォーマットを持っているため, それを利用したアプリケーションのデバイス依存性が高い。
- (2) デジタルペンで収集した手書きデータを効率的に管理, 再利用する手段が少ない。
- (3) デジタルペンの利点は, ノートのように広い筆記面に書かれた文字列や絵図などが混在する複雑かつ自由な手書き文書(スタイルフリーな手書きパターン)を電子化できるところである。しかし, その価値を高める文字認識・検索の研究分野において, このような大規模なデータベースは存在しないため, 技術の発展に支障をきたしている。

我々はこれまでに(2)の問題に対して, 「手書き体験メディアへのメタデータ付与システム」[2]などを提案してきた。しかし, (1)~(3)の問題を踏まえると, まずは, デジタルペンによる標準的な手書き文字パターンの収集・管理環境を整備する必要があると考えた。

そこで, 本研究では, (1), (2)に対して, デバイス独自のデータフォーマットを吸収し, 収集された手書きパターンを Web 上で効率的に管理する手書きパターン標準化マネジメントシステムを構築した。また, (3)に対して, 関連分野の研究を促進するために, 新たにスタイルフリーな手書きデータを収集した。本稿では, それぞれの詳細について報告する。

2. 手書きパターン標準化マネジメントシステム

我々が提案する手書きパターン標準化マネジメントシステムは, 様々なデジタルペンのデータフォーマットの違いを吸収する標準化モジュールと, 収集した手書きデータを効率的に管理するマネジメントシステムから成る。

2.1 データフォーマットの標準化モジュール

ペンデバイスの共通フォーマットとして, InkML 及び Scalable Vector Graphics(SVG) を採用した(図 1)。InkML は

† 富士ゼロックス株式会社, Fuji Xerox Co., Ltd.

‡ 東京農工大学, Tokyo University of Agri. and Tech.

†† 福井県立大学, Fukui Prefectural University.

ペンの筆点座標を XML で記述するための仕様で, 現在 W3C から草案が公開されている。InkML は手書きデータの筆点座標だけでなく, 筆圧やペンの傾きなども記録できる。SVG は XML によって記述されたベクターグラフィックス言語で, W3C でオープン標準として勧告されている。SVG は Web ブラウザ上で閲覧できるため, 収集した手書きデータを容易に Web アプリケーションに組み込むことができる。また, 本モジュールは, Windows アプリケーションの開発に広く用いられている .net Framework 用クラスライブラリとして提供されるため, 提案する手書きパターン標準化マネジメントシステム以外にも容易に組み込むことが可能である。

現在, Anoto ペン, DigiMemo, Technote の 3 種類のデジタルペン以外にも, 通常のペンタブレットやホワイトボード型ペンデバイスなど, ペン入力イベントを発生するデバイスにも対応しており, この他のペンデバイスについては今後対応していく予定である。しかし, 既存のペンデバイスは, 何れも手書きパターンを筆点列座標として収集するという共通の機能を持っており, 今後もこの機能が劇的に変化するとは考えにくい。そのため, 新たに開発されたデジタルペンが登場しても, 比較的容易に対応できると考える。

<pre><ink> <trace> 10 0 9 14 8 28 7 42 6 56 6 70 8 84 8 98 8 112 9 126 10 140 13 154 14 </trace> <trace> 130 155 144 159 158 160 170 154 179 143 179 129 166 125 152 128 140 136 131 149 126 163 124 177 </trace> </ink></pre>
(a) InkML の例
<pre><svg > <polyline points="10 0 9 14 8 28 7 42 6 56 6 70 8 84 8 98 8 112 9 126 10 140" fill="none" stroke="blue" /> <polyline points="130 155 144 159 158 160 170 154 179 143 179 129 166 125 152 128 140 136 131 149 126 177" fill="none" stroke="blue" /> </svg></pre>
(b) SVG の例

図 1 データフォーマットの共通化

2.2 手書きパターンマネジメントシステム

デジタルペンで収集した手書きパターンを効率的に管理するために, 我々が文献[3]で提案した手書き文字列検索技術を用いた Web ベースの手書きパターンマネジメントシステムを開発した。

本システムの手書き文字列検索技術は, 与えられた手書き文字列パターンを文字認識技術によって一意の文字コード

列に変換したうえで検索を実施するのではなく、認識候補と分割候補を複数個生成し、各候補に対して検索を実施する。例えば、図2のような手書きパタンを検索する場合、文字認識結果として可能性の高い「公太郎」と「ハム太郎」の双方に対して検索を行うことができる。(ここで、各ノードは、各文字切出し結果に対する文字認識候補を表す。)このような方式により、検索誤りや検索漏れを軽減することができる。

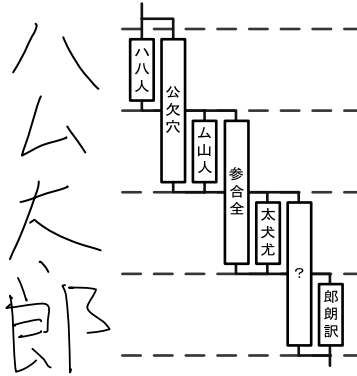


図2 手書き文字列検索

3. スタイルフリーな手書きパタンデータベース

3.1 収集環境

スタイルフリーな手書きパタンデータの収集は、東京農工大学の消費生活協同組合(以下、生協)の協力のもと、同組合が組合員からの意見を聞くために設置されている『ひとことカード』を用いて行った。『ひとことカード』は、用紙を上段と下段に2分割し、上段枠には、組合員が自由に意見を書けるように、下段枠には、寄せられた意見に対して生協の職員が回答を付けられるようになっている。

データ収集には Anoto ペン、及び Anoto ペーパーを利用した。Anoto ペンは、ペン先に組み込まれた特殊なカメラで Anoto ペーパーに印刷されたドットパタンを読み取り、筆記された内容を電子的に取得するデバイスである。我々は、データ収集を行うに当たり、通常設置されている『ひとことカード』の用紙を Anoto ペーパーに変更した。また、本研究はスタイルフリーなデータを収集することが目的であるため、用紙の上段枠もしくは下段枠の中を書くという以外は、特に制約を設けなかった。

3.2 手書きパタン標準化マネジメントシステムによるデータの管理

データ収集は約1ヶ月間実施し、この間に146件のデータを収集した。収集したデータを手書きパタン標準化マネジメントシステムで管理している様子を図3に示す。

手書きパタン標準化マネジメントシステムでは、画面上部に手書きパタンを含む『ひとことカード』がサムネイルで表示され、それを選択すると詳細な内容を確認することができる。また、管理し易いように、2.2で述べた手書き文字列検索のキーワード入力欄を画面左上部に設けてある。

収集したデータは『ひとことカード』ということもあり、上段枠は日付や氏名の行を除いて、2~3行程度の文章が、下段枠は、5行前後の比較的短い文章が書かれていた。また、今回のデータ収集では特に制約を設けなかったことが

ら、予定していた通り、スタイルフリーな手書きパタンを収集することができた。特に、ボールペンなどの筆記用具で紙に書いた時に生じ易い、行外からの文字列の挿入や、矢印による注釈の挿入が多く見られた。また、×印や二重線による文字の訂正や2段構え、下線などによる文字の強調、絵文字や記号を含むデータなども見られた。

その反面、制約のない環境でデータ収集を行ったため、ボールペンや Anoto ペンの電池が切れている状態で記入したことが原因で、データが正しく収集されていないという問題が生じた。今回146件収集したデータのうち、上段・下段いずれかのデータしか保存できていないものが、それぞれ27件、26件あった。今後、収集方法についてさらに検討していく予定である。

なお、収集したデータベースは、日本語手書き文字列認識の標準的な評価用データベースとなることを目指して、手書きに関わる研究機関に無償で公開する予定である。

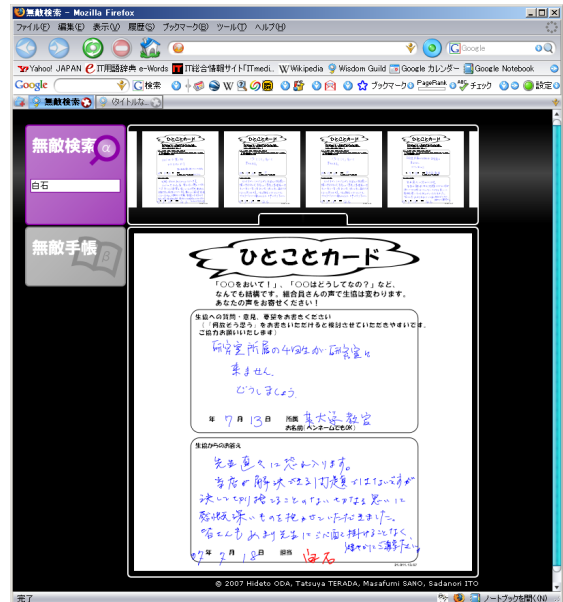


図3 手書きパタン標準化マネジメントシステムによる収集した『ひとことカード』の表示

4. おわりに

本研究では、デジタルペンの更なる普及を目的として、スタイルフリーな手書きパタンの収集環境の整備、及び、データベースの構築を行った。今後は、これらを応用した手書きアプリケーションの開発、及び、スタイルフリーな手書き文字認識精度の向上に取り組む予定である。

謝辞

本研究の一部は、情報処理推進機構、2006年度下期未踏ソフトウェア創造事業の支援による。

参考文献

- [1] <http://www.anoto.co.jp/>
- [2] 織田, “手帳による体験メディアへのメタデータの付与,” 第21回人工知能学会全国大会, Jun. 2007.
- [3] H. Oda, et al. “A search method for on-line handwritten text employing writing box-free-handwriting recognition,” Proc. 9th IWFHR, pp.545-550, Oct. 2004.