手書き文字時系列筆跡パタンの一解析と今後の計画

An analysis into pen-trace patterns of handwritten characters and a future plan

東山孝生、山中由紀子、澤田伸一、中川正樹

Takao Higashiyama, Yukiko Yamanaka,
Shin-ichi Sawada, Masaki Nakagawa
東京農工大学電子情報工学科
〒184　小金井市中町2－24－16

Dept. of Computer Science, Tokyo Univ. of Agriculture and Technology
2-24-16 Naka-cho, Koganei, Tokyo, 184
phone: 0423-88-7144, fax: 0423-87-4604, e-mail: hig@cc.tuat.ac.jp

あらまし: 我々は文章形式、字体制限なし、などを特徴とするオンライン手書き文字時系列筆跡パタンデータベースを作成し、それを利用した字体変動解析を行っている。オンライン筆跡パタン採集の対象とする文章は新聞から抜き出し、頻出の字種を中心にJIS第一水準1,537字種からなる約1万文字の文章列を作成した。そこに含まれなかったJIS第一水準文字は最後に個々に筆記してもらい、合計3,345字種を収集対象とした。現在までに80人分の収集が終了し、最終的に110人分がデータベース化される予定である。また、データベースの次期バージョンの準備も始めている。本稿は初期に収集した30人分のデータを中心に、手書き文字の筆画数変動、筆順変動の解析について報告する。また、次回の筆跡パタンデータベース作成計画と今後の字体変動解析の方針について述べる。

キーワード：オンライン入力、時系列筆跡パタン、データベース、字体変動、筆画数変動、筆順変動

Summary: A database of on-line handwritten character patterns sampled in a sequence of sentences without any instructions has been made. The sentences for which character patterns are sampled have been picked up from newspapers with the result that they are composed of about 10,000 characters and include 1537 kinds of JIS 1st set character categories. The rest of the JIS 1st set categories are written one by one at the end of the above text. Our laboratory collected 30 people's patterns. We proposed common usage of this database with each offering patterns from 5 people. 15 collaborators have joined this project. Recently, we added 5 sets. Stroke number and order variations have been analyzed from the initially collected 30 people patterns.

Key words: on-line input, pen-trace pattern, database, stroke number and order variation.

1. はじめに

ペン入力の実用化の流れの中で、オンライン文字認識の高度化を望むには、現実的な字体変形が含まれる大量の筆跡パタンのデータベースが必要となる。しかし、これまでそのような形のデータベースは存在しなかった。そのことをふまえ、我々は大量の手書き文字時系列筆跡パタンを収集し、共同利用可能な大規模データベースを作成した[1]。

現在当研究室で収集した30人分のパタンに加えて、各機関5人分のパタン提供を呼びかけた結果、約80人分のパタン収集が終了している。さらに6機関の参加希望があり、それを含めると合計110人分のパタンが収集される予定である。

本データベースでは字体変形を含む自然な筆跡パタンを収集するという方針から、筆記者には字体に注文を付けず、指定した文章列を筆記してもらった。同時に筆記者の個人情報を記録してある。また、オンライン入力パタンは筆点座標が時系列で採集されるので筆跡過程を追うこともできる。このようなことからさまざまな面からの解析が可能である。

本稿は手書き文字の筆画数変動の解析[2]、筆順変動、字体変動解析についての報告とデータベース収集の今後の計画について述べる。

2. 筆画数変動

標準画数別の実際の筆画数分布を図 1に示した。分布を見ると、どのグラフでも標準画数で書かれたものをピークとし、画数が減少する方へ傾斜した形になっている。また、図 2に標準画数別に文字パタンの最小画数、最大画数を示した。ここで、斜めの直線が交わる点がそれぞれの標準画数である。標準画数が増加するにつれて分布の幅が広がることがわかる。例えば、標準画数20画の場合、最大22画、最小3画で筆記されたパタンが存在する。実際に筆記された画数が標準画数より減少している原因はストローク間の続けが起きていることがあげられる。標準画数が増加するに従いストローク間の続けがより多く起きていることがわかる。しかし、21画以上の高画数のパタンの分布の幅は狭くなっている。この

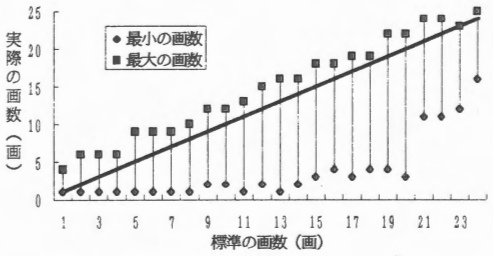


図 2　筆画数分布の範囲

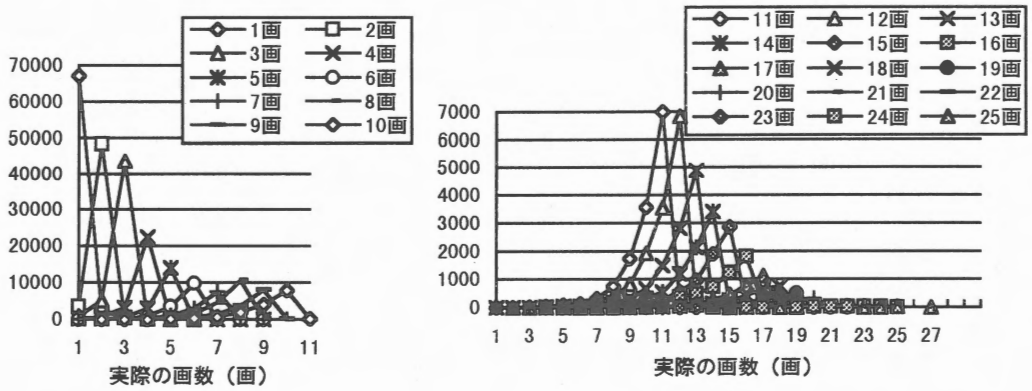


図 1　標準画数別の実際の筆画数分布(全 358,860 パタン)

原因はパタン数の絶対数が少ないことがあげられる（標準画数15〜20画：670×30パタン、21画以上：31×30パタン）。また、標準画数21画以上の文字は筆記しなれていない文字が多いために続けが起こりにくいことが考えられる（表 1）。

表 1　標準画数21画以上の文字

鰯艦顧轟鐸鶴纏灘鰭魔躍鑓露鰻驚轡讃襲鱈聾鰹鑑讐鱒鷲鷺鷹鱗麟欝

画数別で全体的に見ると前述のような分布をしているが、各字種別で見ると、表 2に示したように標準画数で筆記されたパタンが必ずしも最も多いわけではない。全3,345字種中731字種では標準画数で筆記されたパタンよりも、標準ではない画数で筆記されたパタンの方が多い。例えば、部首『幺』は標準画数は3画だが続けて筆記され1画または2画で筆記されることが多い。部首『糸』も同様である。表 3に示したように部首『幺』を含む文字『磁』は標準の14画よりも7画で筆記されたパタンのほうが多い。他にも部首『幺』『糸』を含む字種はその影響からか標準画数より少ない画数のパタンが多く見られる(繰磯幾機轡縮など)。部首『口』やしんにょうなどを含む筆跡パタンについても同様なことが起きている。このことから、ストローク間の続けが起きやすい部首が存在していることがわかる。

また、実際に筆記された画数が標準画数よりも多いパタンもわずかながら存在した。この原因としてストロークの切れやゴミなどの影響によるものがあげられる。この他の例として、図 4 に2つのパタン例を示した。『比』は本来4画だが、5画で書かれているものが多く見られた。『比』の左側の部首は本来2画であるが、ほとんどの場合3画で筆記されている。このことは筆記者が普段から筆記していることも考えられるが、表示されているフォントを見て、その形をそのまま筆記していることも考えられる。『篦毘枇庇』や『鼠』などはフォントの影響が大きいと思われる。このような影響を取り除くためにフォントを見せずに音声による指示が考えられる。しかし、実際には見ないとすぐに筆記できない文字が余りにも多い。また、音声からでは仮名で筆記すべきか、漢字で筆記すべきか判断できない。このような点から、フォントを表示する方法を採用している。

表 2　標準画数以外の画数で筆記されたパタンがもっとも多かった字種

731字種（3,345字種中）

画数の差	-7	-5	-4	-3	-2	-1	1
文字種数	1	1	8	19	111	575	16

（ここで画数の差とは、ある字種において標準画数と実際に筆記されたパタンの画数の中で最も多い画数との差である）

表 3　筆跡パタン『磁』の画数

総パタン数 30個　標準画数 14画

実際の画数	7	12	11	14	13	9	8	10	4,5,6
パタン数	6	5	3	3	3	3	2	2	各1

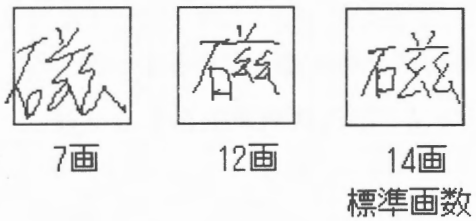


図 3　『磁』の筆跡パタン例

表 4　筆跡パタン『鼠』 の画数

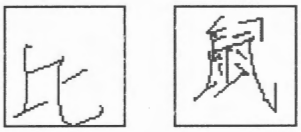
総パタン数 30個　標準画数 13画

実際の画数	14	13	15	12	16
パタン数	16	9	2	2	1

表 5 筆跡パタン『枇』の画数

総パタン数 30 個 標準画数 8 画

実際の画数	8	7
パタン数	16	14



比 鼠

図 4 標準画数より画数が多い例

3. 同一筆記者の筆順変動と字体変動

3. 1. 検査方法

データベースの文章部 10154 文字は 1537 文字の JIS 第一水準文字と 11 文字の JIS 第二水準文字により構成されていることからわかるように、同じ文字が何度も繰り返し出現している。出現数別の文字カテゴリ数を表 6に示す。ここに示されている 2 回以上出現する 857 文字カテゴリを用いて同一筆記者の筆順変動と字体変動を調べる。

表 6 出現数別文字カテゴリ数

出現数	1	～5	～10	～20	～100	101～
カテゴリ数	680	554	181	56	49	17

筆順変動、字体変動を調べるために本研究室で作成されたオンライン手書き文字認識システム[3][4][5]を用いる。この認識システムの特徴としてストロークのつながりに寛容であるために筆画数の変動は認識結果に大きな影響を及ぼさないことがあげられる。

次のような方針で変動を調べる。同一筆記者において、ある文字カテゴリ内で最初の筆跡パタンを辞書パタンとして登録する。登録の終了後、2 番目以降の筆跡パタンに対して前述の認識システムを用いて認識を行う（図 5）。これによりあるカテゴリ内の最初の筆跡パタンとそれ以外の筆跡パタンとの類似度が得られる。この類似度は 0～1000 で表され数字が大きいほどパタン間の差が少なく、類似度が 1000 の場合は全く同じパタンである。よって類似度が大きいパタン間では筆順変動、字体変動がほとんど起きていないと考えられる。類似度の小さいパタン間では筆順、字体の変動が大きく起きていると考えられる。この類似度に一定のしきい値を設定して、しきい値以下のパタンの筆順、字体を実際に目で見ることにより筆順、字体の変動を調べることができる。

また、認識を行うことにより、あらかじめ辞書に登録されているパタンとの類似度を得ることができる。

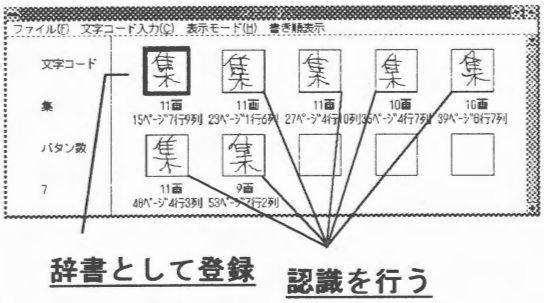


図 5 筆順変動、字体変動の検査

3. 2. 同一筆記者の筆順変動と字体変動の例

筆記者が異なると文字の筆順や字体が異なることは多く見られたが、同一筆記者内ではこれらの変動はほとんど起きていない。

しかし、そうした例がないわけではない。次に同一筆記者内で筆順変動の起きている例をあげる。図 6に示すように『馬』は同一筆記者によって 3 通りの筆順で筆記されていることがわかる。1 つ目の例は正しい筆順の 6 画目を筆記している途中で 5 画目を重ね書きしている。その後に、6 画目に継ぎ足しが行われている様子がわかる。このパタン画数は標準の 10 画から、2 画増えて、12 画になっている。2 つ目の例は正し

い筆順の1画目と2画目の順序を逆に筆記していることがわかる。3つ目の例は正しい筆順で筆記されている。このように同一筆記者においても筆順の変動が起きていることがわかる。

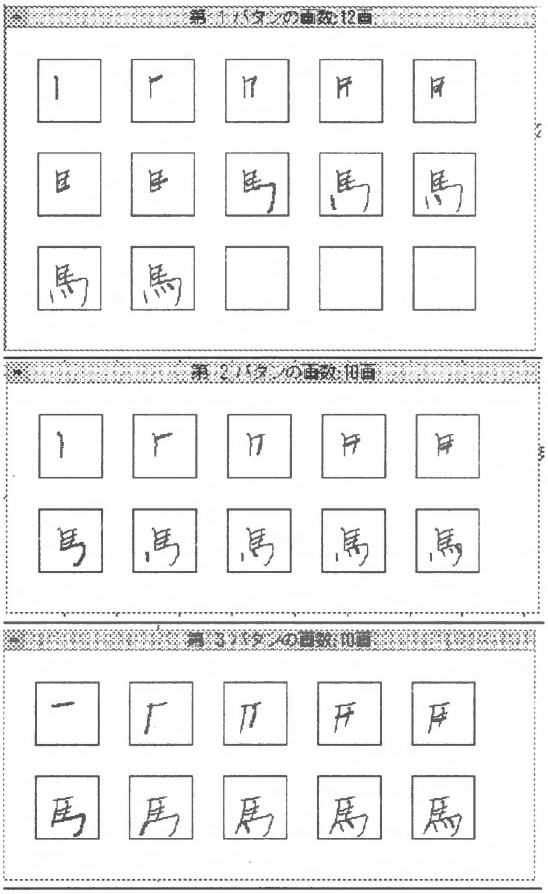


図 6　同一筆記者の筆順変動の例

図 7は同一筆記者における字体変動を示している。『第』を標準の字体とその略字体で筆記している。表示しているフォントは標準の字体の『第』であることから、この筆記者は普段は略字で筆記しているが、例として示されているフォントを参照しながら文字を筆記しているために、標準の字体で筆記したと考えられる。なお、本データベースでは自然な筆跡パタンの収集を目的としているために、略字体の制限も行っていない。

この例のような筆順変動、字体変動の起こる割合などを出すことが今後の課題としてあげられる。

図 7　字体変動の例

4. 今後の計画

今回のデータベースの版でいくつかの問題点があがった。それらを考慮した新しい筆跡パタンデータベース作成の計画について述べる。

4. 1. 例文表示フォント

特にデータベース最後の文字の羅列の部分においては、一般的に目にすることの少ない文字が含まれている。そのために筆記者によっては知らない文字、筆記したことのない文字が含まれている。特に画数の多い文字は表示しているフォントが見難いことが多く、このような場合今回の版ではあらかじめ印刷した例文を見ながら筆記してもらった。しかし、このことは視線の移動や印刷された例文から筆記すべき文字を探すために、大きな筆記の中断が起き、自然な筆跡パタンを収集することの妨げとなる。次の版では表示フォントをペンでタップすることにより拡大されたフォントが表示され（図 8）、筆記の中断を最小限にすることができ、筆記が容易になる。

この他にも上記2でも触れたが、表示フォントの形の問題がある。表 7の『備』『心』は明朝体では通常筆記する字体と異なる字体であるが、正楷書体では通常筆記する字体である。ところが、『継』の場合は明朝体が通常筆記する字体で正楷書体が異なる字体である。このような問題が起こるためにすべての文字が通常筆記する字体で正しく表示されるフォントが必要となる。

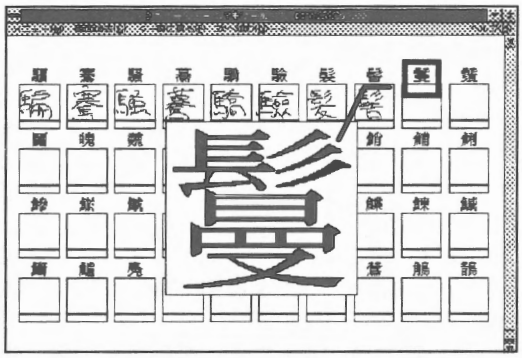


図 8　拡大された例文フォント

表 7　表示フォントの問題

MS 明朝	備・心・継
HG 正解書体	備・心・継

4. 2. 収集パタン座標

今回の版ではマウスの割り込みによりマウス座標で筆点を収集していた。このためにある点でペンを止めたとき、止めている時間によらず、ある点のデータは 1 つしか得ることができなかった。次の版では Windows95 で標準装備されるであろう penwin.dll を利用することによって、一定間隔の割り込みによる筆点の収集が可能となった。よって筆記者がある点でペンを止めたときは、止めている時間だけ、同じ座標を記録することができる。これにより筆記の過程がより正確に追跡できることになる。また、タブレット座標で採種することが可能となるためにさらに細かい分解能で筆跡パタンの収集が可能になる。

4. 3. 第二水準文字の追加

新しいデータベース収集システムの作成に伴い、筆記する例文も新しく作成した。主な特徴として、JIS 第二水準文字を文字の羅列として約 1,500 文字追加した。また、文章部の一文あたりの字数に制限を設けた。現在のペン入力は単語や短いメモ程度の文章の入力に用いられている。このことから現実にあわせて、できるだけ字数の少ない文章により例文を構成することを目的としている。

4. 4. 誤字・脱字検査ツール

収集されたパタンに対して、誤字・脱字の検査を行っている。検査の後により正確なデータベースを作成するために筆記者に訂正を依頼している。今回検査ツールは誤字・脱字の記録は筆記者に訂正を依頼するための簡単なメモ程度の記録しかしていなかった。また、明文化された検査基準がなかったために、検査を行った者の主観に頼る面があった。次の版では今回の誤字・脱字検査の記録を元に可能な限り、検査者の主観の入らない検査基準を設け、細かな記録を残す検査ツールの作成を行う。

謝辞

本研究は試験研究 05558027、および、重点領域研究 07207207 の一部補助による。

参考文献

[1] 中川、東山、山中、澤田、レー、秋山：文章形式字体制限なしオンライン手書き文字パターンの収集と利用、信学技報PRU95-110、43-48(1995.9).

[2] 山中、東山、澤田、中川：オンライン手書き筆跡パタンの収集とその一解析、情処人文科学とコンピュータ研資28-1、1-6(1995.11).

[3] 秋山、中川：ストロークのつながりに寛容なオンライン手書き文字認識、画像の認識・理解シンポジウム（MIRU'94) I、67-74(1994.7).

[4] M.Nakagawa and K.Akiyama:A Linear-time Elastic Matching for Stroke Number Free Recognition of On-line Handwritten Characters, Proc. 4th IWFHR, 48-56(1994.12).

[5] レー、秋山、中川：ストローク数非依存の高速オンライン手書き文字認識手法、情処第50回全大、2-61/62(1995.3).