

手書き文字時系列筆跡パターン解析 2

4 G-8

山中由紀子

東山孝生

中川正樹

東京農工大学

1. はじめに

ペン入力の認識技術の高度化のためには、実際に筆記される変形字体を含む大量の手書き筆跡パターンが不可欠である。しかし、今までこのような形でのオンラインデータベースは存在しなかった。そのことをふまえ、我々は筆記者の書体には何の制限も与えない環境で、文章列を例文としてオンライン筆跡パターンの収集を行った。この方針での収集により、より自然な形の筆跡パターンが収集できたと考えられる。

筆跡パターンは、最初当研究室で 30 人分収集した。その後、共同利用を呼びかけた他の研究機関からもパターン収集の協力を得て、現在約 80 人分のデータが収集済みである。

将来的には共同利用可能な大規模データベースへと発展していく予定である[1]。

本稿では、初期に収集した 30 人分のデータ（各データには nk01～nk30 のデータセット名が付けられている）を用いて、筆画数変動の観点から筆跡パターンの変動解析を行い、結果を報告する[2][3]。

2. 筆画数変動

2.1. 収集対象例文に含まれる文字

収集対象字種は JIS 第一水準の頻出頻度が高いものを中心とした 3,356 字種（第二水準 11 字種を含む）である。それらを含む収集対象例

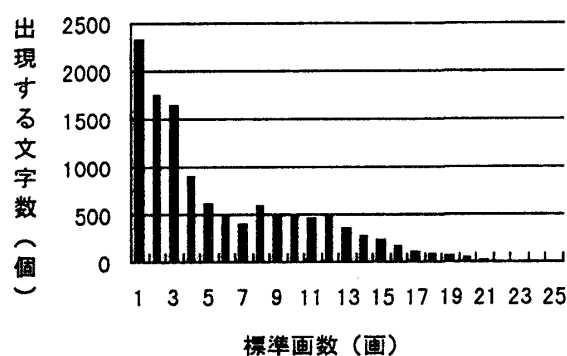


図 1 例文に出現する文字の標準画数別分布
(全 11,962 文字、最大 25 画)

表 1 例文中に 10 回以上出現する漢字 (画数順)

一	力	二	人	十	子	大	上	小	三	月	日	公	分	文	中	化	民	生	本
立	用	出	年	自	同	合	全	地	会	見	社	物	者	国	的	实	事	金	思
後	政	書	高	時	動	現	部	場	間	開	新	業	関	機					

文は 10,154 文字の文章部と 1,808 文字の文字の羅列部分の二つからなる。

図 1 に例文に出現する文字の標準画数別の分布を示した。これを見ると、1 画から 3 画が特に多く、全体的に低画数文字の出現数がかなりの部分を占めることがわかる。この収集例文は大部分が文章部からなるため、同じ字種が繰り返し出現することが多く、それらの文字では平仮名、句読点、括弧などの画数が少ない字種が多い。また一部の漢字も重複して出現するものがあるが、表 1 に示した例文中で出現頻度が高い漢字には、20 画以上の高画数文字は含まれなかった。

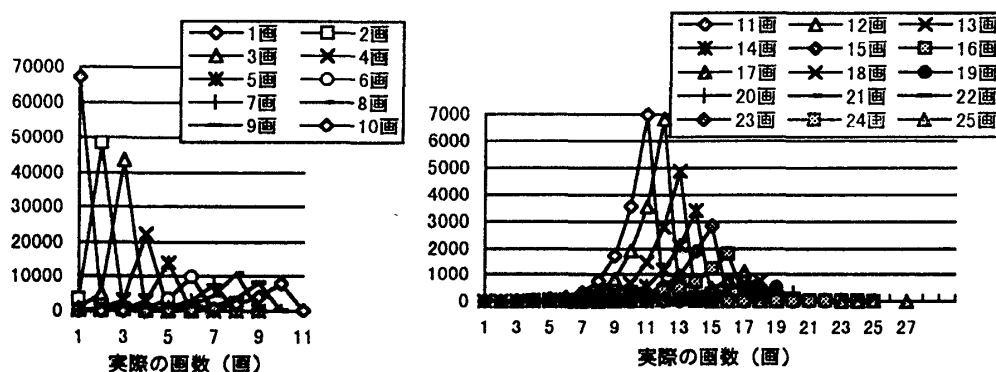


図2 標準画数別の実際の筆画数分布(全 358,860 パタン)

2.2. 筆画数分布

全 358,860 パタン(11,962 パタン×30)には、1 画から 27 画までのパタンが存在した。

図 2に示した標準画数別の実際の筆画数分布を見ると、標準画数で書かれたパタンを最大とし、少ない画数の方へ減っていく分布の形をしている。標準画数より少ない画数のパタンには、ストローク間に起きる続けの影響がある。また、標準画数より多い場合もわずかではあるが存在し、原因としてゴミやストロークの切れなどがある。

ストローク間の続けが起きる部分は共通していることも多い。例えば、部首『幺』が続けて 1 画で書かれたパタンは共通して見られた。

データセット別でも変動の形は異なる。図 3 に標準画数と実際の筆画数の差の変動の 3 例を挙げた。nk14 を変動が大きい例、nk21 を変動が小さい例、nk27 を平均的な例として示した。

2.3. 個人内変動

データセット内の同字種での画数変動を見るため、例文に 10 回以上出現する漢字(表 1)の変動を調べた。その結果、各セット共通して変動が大きいものは比較的画数の多い字種(『機高部』など)が多かった。変動が起きている例として、図 4に nk7 の『機』のパタンを示した。この場合、12 画から 16 画までの筆画数変動が起きている。

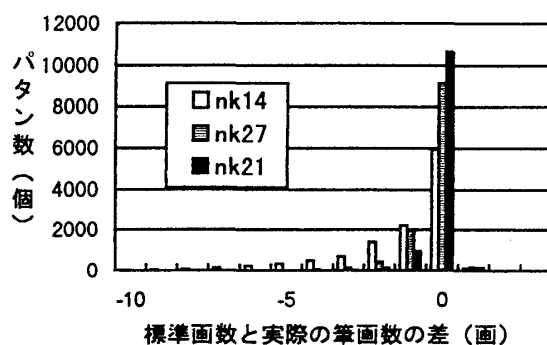


図3 データセット別筆画数変動



図4 nk7 の『機』

3. おわりに

今回用いたデータは、筆画数変動がかなり大きかった。やはり書体に制限を与えなかったことから、続け字を含む自然な筆跡パタンが多く含まれているのではないと思われる。

今後、データベースの拡張に伴い、対象を拡大して解析を行っていきたい。

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ー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] 東山、山中、澤田、中川：手書き文字時系列筆跡パターンの一解析と今後の計画、公開シンポジウム「人文科学とデータベース」公演予稿集、37-42(1995,12)