Sixth International Conference on Document Analysis and Recognition (ICDAR), Seattle (USA), pp.496-500 (2001.9).

Collection and analysis of on-line handwritten Japanese character patterns

Kaoru Matsumoto, Takahiro Fukushima, Masaki Nakagawa

Dept. of Computer, Information and Communication Sciences
Tokyo University of Agriculture & Technology
Naka-cho 2-24-16, Koganei, Tokyo, 184-8588, Japan
phone: +81-423-88-7144, fax: +81-423-87-4604, e-mail: matsu@hands.ei.tuat.ac.jp

Abstract

This paper describes our second collection of on-line handwritten character patterns and their analysis. 163 writers presented about 10,000 character patterns, covering 4,438 categories mainly in the context of sentences. Together with our first collection, the Kuchibue database containing 12,000 patterns from 120 writers, we have now collected about 3 million patterns. For this second collection of on-line patterns, named Nakayosi, we analyzed stroke number and order variations.

Keywords: database, on-line patterns, character patterns, Japanese characters, data analysis

1. Introduction

A large amount of character patterns is essential to study character recognition on a common basis in an objective and reproducible form[1][2][3].

To collect on-line handwritten patterns, tablets with pens are needed. We developed pattern collection tools for the MS-Windows operating system and collected on-line handwritten patterns on multiple pen computers or PCs with attached LCD tablets. Thereby, two databases have been collected: TUAT Nakagawa Lab. HANDS-kuchibue_d-97-06 (referred to as kuchibue_d below) containing patterns from 120 writers, each presenting about 12,000 character patterns, and TUAT Nakagawa Lab. HANDS-nakayosi_t-98-09 (referred to as nakayosi_t below) storing patterns from 163 writers, each presenting about 10,000 character patterns. In Paper [4], we reported policies to collect on-line handwritten character patterns, described the tools to collect them, and presented a rough sketch of the kuchibue_d database.

This paper summarizes some important features of the databases, reports recent development of access libraries and presents analyses on pattern deformations and variations stored in the nakayosi_t database.

2. Features of kuchibue_d and nakayosi_t

2.1 Policies of pattern collection

In order to collect patterns useful for developing powerful on-line handwriting recognizers in real applications, the following policies were decided:

(1) Each writer writes frequently used characters by copying sentences so that we can collect more casual patterns and can also evaluate the effect of context processing as well as the user customization capability of recognizers. Characters used less frequently but included in some standards are written after the sentences without any meaningful context. This is because writers would have to write millions of characters if we had tried to collect all the script patterns of 3,000 or 4,000 categories in sequences of sentences.

(2) Each writer writes characters one by one in a sequence of writing boxes over which the printed characters are displayed for the writer. This is not a restriction to Japanese people, because Japanese are accustomed to this kind of manuscript paper. This environment is useful to collect a large amount of character patterns with ground truth codes.

(3) No restriction to the quality of character patterns.

(4) We don't show the result of recognition so that writers don't try to write intentionally character patterns that are easy or difficult to recognize.

(5) We record the writers profiles and the specifications of collection environments such as the sampling rates etc.

(6) We first test the correctness of patterns automatically by a software tool and then check them by visual inspection. Human inspection is generally superior to machine tools but humans often overlook wrong characters in meaningful text. Therefore, we first test patterns by a machine tool to pick up erroneous patterns and inspect them visually without any context. Each writer confirms picked-up patterns and rewrites them only when he/she identifies them as wrong patterns. More detailed arguments are given in [4]. Fig. 1 shows a typical writing screen.

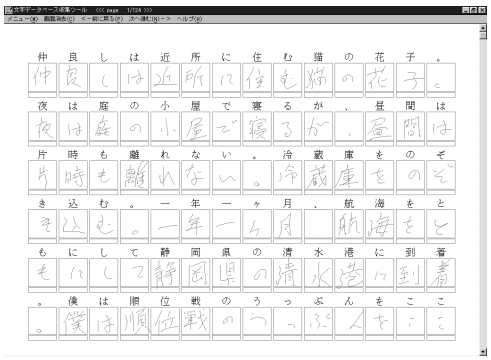


Fig. 1 Database collection tool screen.

2.2 Categories for script pattern collection

In the kuchibue_d database, we collected script patterns for the JIS (Japanese Industrial Standard) 1st level set of Japanese characters except Cyrillic characters, symbols with very low degree of occurrence, filled-in symbols and some others. The sentences according to which character patterns are collected were picked up from the 1993 year's CD-ROM edition of the Asahi newspaper (Japanese language newspaper).

In the nakayosi_t database, we extended the script collection and included about 1,000 JIS 2nd level set characters that are used for naming. The sentences to write patterns were again selected from the same CD-ROM but shorter sentences were used in order to decrease the amount of text while enlarging the character categories for the script collection. Table 1 summarizes the organization of the two databases.

Table 1 The organization of the databases.

	kuchibue_d	nakayosi_t
total characters	11,962	10,403
total char. categories	3,356	4,438
characters in textual part	10,154	7,376
categories in textual part	1,548	1,411
characters in non-textual part	1,808	3,027

2.3 Database distribution

Now, 10 peoples' patterns in kuchibue_d are freely available for research purpose[5].

This sample distribution is only published on CD-ROM under the agreement that the newspaper company can show the copyright for its text used for sampling character patterns. Moreover, the full sets of kuchibue_d and nakayosi_t are available from the book store of Tokyo University of Agriculture and Technology.

2.4 Access methods to the databases

The databases were initially coded in a binary format and accessed through libraries. By using the binary format, each writer's patterns can be stored on a 3.5 inch FD. Moreover, this seemed to be effective when we need to change the internal format since it is covered by the access libraries.

It has turned out, however, that the binary format and access libraries are difficult to use on other platforms than MS-Windows. Therefore, we have now made the databases available in a simple text format and in the UNIPEN format [6]. Recent compaction methods reduce their sizes even smaller than the original binary versions.

2.5 Ideal script set

We compiled a set of script patterns of correct stroke-number and stroke-order for each of the collected categories in the JIS 1st level and all the categories in the JIS 2nd level, thus 7,723 categories in all. This set was collected with the same tool, but using larger writing boxes than those for pattern collection. The analyses on character pattern variations use this set.

3. Script pattern analyses

We analyze stroke number and order variations for 163 writers in nakayosi_t. There are 1,695,689 patterns in all.

3.1 Analysis of stroke number variations

Fig. 2 shows how the number of strokes of character patterns deviates from the standard. For each correct number of strokes shown on the horizontal axis, the actual range of stroke numbers is displayed on the vertical axis. The number of character categories for each correct stroke count is also shown in Fig. 2.

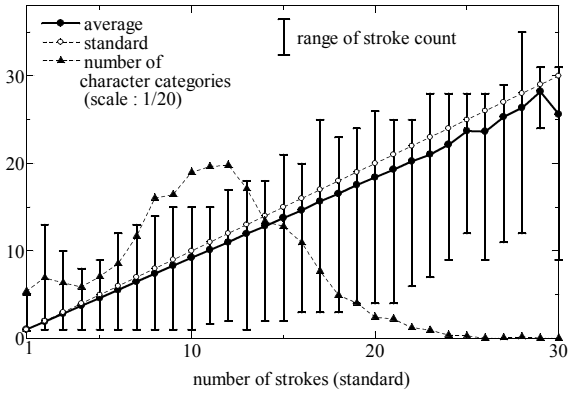


Fig. 2 Range of stroke numbers.

When people write characters casually, strokes are more likely connected and sometimes broken into two or more parts. Thus, the actual range of stroke number strongly varies. The ranges are not so small as supposed until recently for on-line handwritten character recognition.

Fig. 3 (a) and (b) show more detailed views to the range of stroke numbers for character patterns of each correct stroke count from 1 to 30. In these figures, "N strokes" means character patterns that should be written by N strokes in neat correct handwriting.

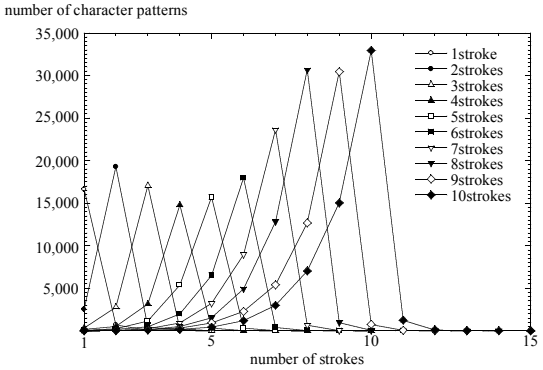


Fig. 3-(a) Distributions of the stroke count (1-10 strokes).

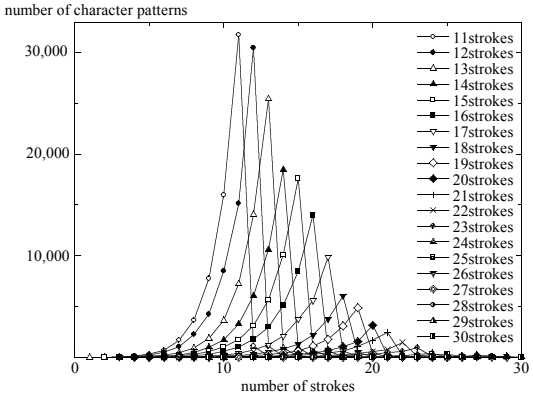


Fig. 3-(b) Distributions of the stroke count (11-30 strokes).

These figures show that as the correct number of strokes becomes larger, the range of the actual stroke number spreads wider. The actual stroke number tends to be lower than the correct number of strokes because character patterns are often written with connected and/or abbreviated strokes.

54.12% of the character patterns are written with the correct stroke number. The percentage of patterns written with a stroke number smaller than the correct stroke number is 43.71%, while the share of patterns written with a larger number of strokes is 2.17%.

The smallest and biggest number of strokes were searched for every category in the database to find the character categories written with large stroke number variations. Fig. 4 shows categories with 14 or more stroke number variations.

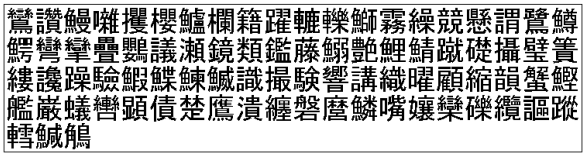


Fig. 4 Character categories with large stroke number variations (14 or more).

Most of them include the radicals (subpatterns) shown in Fig. 5. Those radicals are often written cursively so that the stroke number variations between neatly written and cursively written characters become large.

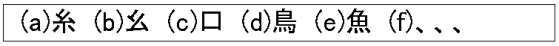


Fig. 5 Radicals with large stroke number variations.

Fig. 6 shows a pattern with the least stroke count for the category with the largest stroke number variation. When it is written correctly, it is composed of 30 strokes. In Fig. 6, it is written with 9 strokes.

Fig. 6 Pattern written with 9 strokes.

Fig. 7 shows character categories written with a different stroke number, deviating from the standard by 3 or more strokes.

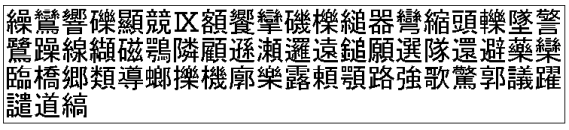


Fig. 7 Character categories with large stroke number deviations from the standard.

Many of them include the radicals in Fig. 5 (a) and (b), some include (c) (d). Fig. 8 shows additional radicals with stroke number deviations. People often write them without pen-up thus producing cursive patterns.

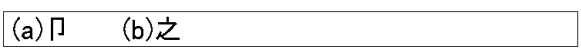


Fig. 8 Other radicals with large stroke number variations.

On average, the character category with the largest stroke deviation has 4.79 strokes less than the standard. Fig. 9 shows a pattern of that category written by 4 strokes. When written correctly, it is composed of 19 strokes.

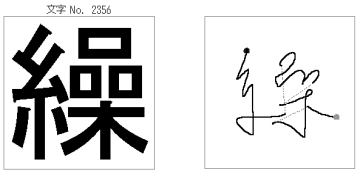


Fig. 9 Pattern written with 4 strokes.

Fig. 10 shows character categories written without any stroke number variation. They are mostly 1 stroke patterns or symbols.

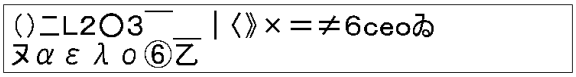


Fig. 10 Character categories without any stroke number variation.

Fig. 11 shows character categories written with an average stroke number that is at least 0.2 higher than the standard.

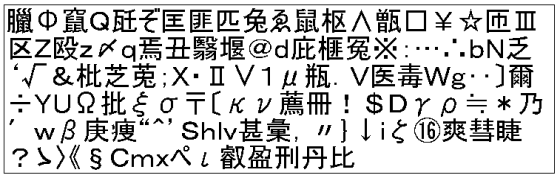


Fig. 11 Categories with a greater number of strokes than the standard.

These categories include radicals that appear less frequently. Most of them are not learned in school. People do not normally write or cannot write these categories so that they have to mimic the font patterns when they are requested to write. Consequently, the number of strokes becomes larger than the standard because people may divide a complex stoke into two or more pieces and produce unstable orders.

3.2 Analysis of stroke order variations

In order to investigate stroke order variations, we have utilized our recognition engine[7] with our strict stroke order dictionary and our general dictionary where several stroke order variations are registered for each subpattern and shared among character patterns.

If the recognition engine with the general dictionary produces a high recognition score of the input pattern for the correct category while that with the strict dictionary rejects the input pattern, this will imply that the input pattern likely has stroke order variations. We picked up those patterns and inspected their writing order visually. We applied this inspection to five data sets and show the result in Table 2.

Table 2 Picked-up patterns.

	Picked-up patterns	Symbols	Correct	Incorrect
NKY0001	126	49(39%)	47(37%)	30(24%)
NKY0002	157	47(30%)	76(48%)	34(21%)
NKY0003	150	73(49%)	53(35%)	24(16%)
NKY0004	161	50(31%)	70(44%)	41(25%)
NKY0005	163	71(44%)	49(30%)	43(26%)
(TOTAL)	757	290(38%)	295(39%)	172(23%)

The writing order for symbols are not clearly defined so that we exclude them from investigation. Then, 23% of the picked-up patterns have an incorrect stroke order. Incorrect stroke order patterns would also exist in unpicked-up patterns. We have applied the pick-up process to the full sets of nakayosi_t, where picked-up patterns hold 1.5% on average. From these numbers, we may estimate that at least 0.015 x 0.23 x 100 = 0.345% of all patterns are written by non-standard stroke order. Since 36% of collected character patterns are hiragana and katakana characters and there are only few wrong stroke order patterns among them, we may estimate that 0.345 x 100 / (100 - 36) = 0.54% of Kanji patterns in nakayosi_t are written with incorrect stroke order.

Fig. 12 and 13 show categories written with incorrect stroke order and radicals that often appear among the incorrect stroke order patterns, respectively. Those radicals seem to cause stroke order variations.

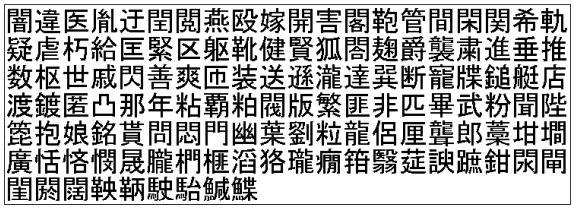


Fig. 12 Character categories written with incorrect stroke order.

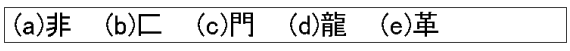
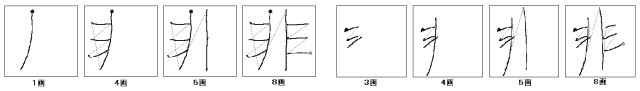


Fig. 13 Radicals causing stroke order variations.

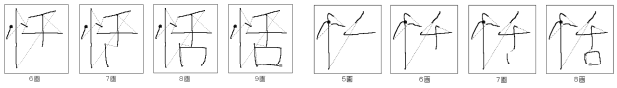
Fig. 14 and 15 show typical wrong stroke order patterns with their ideal script patterns. In Fig. 14, the left vertical stroke must be written first, but many people prefer the common left to right order.



ideal script pattern NKY0003.IPD

Fig. 14 Character " 非"(a model, NKY0003.IPD).

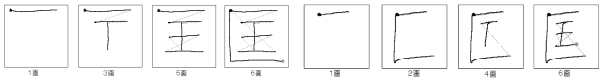
As shown in Fig. 15, many people simplify the "mouse" radical of the bottom right corner by a single-stroke square thus stroke order is changed. Very often, pattern simplification accompany stroke disordering.



ideal script pattern NKY0005.IPD

Fig. 15 Character " 恬 "(a model, NKY0005.IPD).

As shown in Fig. 16, some radicals should not be written by consecutive strokes, but people prefer to write them as a unit.



ideal script pattern NKY0001.IPD

Fig. 16 Character "匡 " (a model, NKY0001.IPD).

3.3 Variations within each writer

Stroke number variations between different handwritings is pretty common, but stroke order variations within an individual handwriting is quite rare. A few exceptions are observed when a missed stroke is added at the last stage of writing a character pattern or when simplified patterns are used instead of normal patterns.

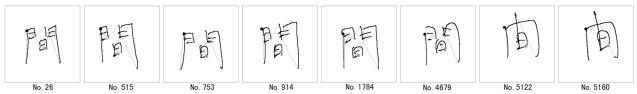


Fig. 16 Normal patterns and simplified patterns for "間 ".

4. Results

We described our second collection of on-line handwritten character patterns and their analysis: 163 writers presented about 10,000 character patterns each, including 4,438 categories mainly in the context of sentences. For this second collection of on-line patterns, we analyzed stroke number and order variations.

* The distributions of stroke number variations are not so small as supposed until recently for on-line handwritten character recognition. The correct stroke number patterns share 54 % of all patterns.
* As the correct number of strokes becomes larger, the actual range of the stroke number spreads wider and the number of stroke tends to decrease rather than increase.
* Character categories whose patterns include some specific radicals are written with larger variations and deviations from the standard number of strokes.
* Unfamiliar character categories are written with a greater number of strokes when people are requested to write.
* Stroke number variations and deviations from the standard is high, even for the handwriting of a single person.
* Stroke order variations are estimated to occur in more than 0.54 % of the Kanji patterns.
* The writing orders for symbols are unstable since they are not clearly defined.
* Some delayed strokes may cause stroke order variations.
* Stoke order variations within individual handwritings are negligible.
* Style variations or simplifications are observed even for a single writer.

References

[1] Joathan J. Hull: A Database for Handwritten Text Recognition Research, IEEE Transactions on pattern analysis and machine intelligence. vol.16, NO.5, pp.550-554 (1994.5).

[2] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman and S Janet: UNIPEN project of on-line data exchange and recognizer benchmarks, Proc. 12th ICPR, Vol. II, pp.29-33 (1994.10).

[3] Christian VIARD-GAUDIN, Pierre Michel LALLICAN, Stefan KNERR, Philippe BINTER: The IRESTE On/Off (IRONOFF) Dual Handwriting Database, IEEE, pp.455-458 (1999).

[4] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, K. Akiyama: On-line Handwritten Character Pattern Database Sampled in a Sequence of Sentences without Any Writing Instructions, Proc. 4th ICDAR, pp.376-381 (1997.8).

[5] http://www.tuat.ac.jp/~nakagawa/ipdb/

[6] Stefan Jaeger, Masaki Nakagawa: Two On-Line Japanese Character Databases in Unipen Format, ICDAR'01, (2001.9).

[7] M. Nakagawa, K. Akiyama: A Linear Time-Elastic Matching for Stroke Number Free Recognition of On-Line Handwritten Characters. 4th IWFHR, pp.48-56 (1994).