IJDAR (2004) 7: 69–81
Digital Object Identifier (DOI) 10.1007/s10032-004-0125-4

International Journal on Document Analysis and Recognition
IJDAR

Collection of on-line handwritten Japanese character pattern databases and their analyses

Masaki Nakagawa[1], **Kaoru Matsumoto**[1,2]

[1] Department of Computer, Information and Communication Sciences, Tokyo University of Agriculture and Technology, Naka-cho 2-24-16, Koganei, Tokyo, 184-8588, Japan
[2] Research Institute, The SANNO Institute of Management, Todoroki 6-39-15, Setagaya, Tokyo, 158-8630, Japan

Received: 14 December 2002 / Accepted: 26 October 2003
Published online: 22 April 2004 –

Abstract. We describe the design of on-line handwritten Japanese character pattern databases, software tools for pattern collection and verification, and analyses of collected patterns. Two databases containing over 3 million patterns were compiled: one with 120 people contributing 12,000 patterns each and another with 163 participants contributing 10,000 patterns each. Patterns were collected mostly in their sentential contexts and verified by machine and human inspection. Their analyses reveal greater variations in stroke count for characters having many strokes, with people generally using fewer strokes; they additionally reveal that stroke order variations are generally caused by common habits and added strokes.

Keywords: Database – On-line patterns – Character patterns – Japanese characters – Data analysis

1 Introduction

Although many approaches to on-line recognition of handwritten Japanese characters have been presented until now [5,12], each of them has been evaluated with different data and so it is not clear which method is superior and how. Under these conditions steady progress cannot be made. On the other hand, the technology for off-line recognition has made significant progress in Japan since the ETL (Electrotechnical Laboratory) database (especially ETL-9 [9]) was made available as a common benchmark.

In the field of pattern recognition, large volumes of patterns are as important as recognition methods [10]. In fact, publications on pattern databases are increasing and attracting more and more attention [1,2,14]. For on-line recognition of handwritten Japanese characters, however, no large and reliable database is available.

To fill this void, we spent 4 years compiling two databases of on-line handwritten character patterns called "TUAT Nakagawa Lab. HANDS-kuchibue_d-97-06" (hereafter Kuchibue_d) [8] and "TUAT Nakagawa Lab. HANDS-nakayosi_t-98-09" (hereafter Nakayosi_t) [4].

As for on-line handwritten character pattern databases, the UNIPEN project is well known [1], but it does not include oriental characters. Our database project started independently of the UNIPEN project and now provides the UNIPEN format versions of the databases as well [3].

In this paper we summarize these efforts by presenting the design of the databases in Sect. 2, a description of character categories in Sect. 3, and software tools for script collection and verification in Sect. 4. Section 5 provides actual data collection and distribution, Sects. 6 and 7 give analyses of pattern deformations and variations stored in the databases in terms of stroke-number variations and stroke-order variations, and Sect. 8 offers suggestions for making robust recognition systems against those deformations and variations. Section 9 concludes the paper.

2 Design of the databases

2.1 Computer-supported data collection

In order to collect on-line handwritten patterns, i.e., sequences of coordinate values sampled from pen tip movements, tablets and computers are needed at the time of writing, unlike off-line pattern collection for which collecting pattern samples on ordinary paper is sufficient.

This has been one of the obstacles to collecting a large amount of on-line handwritten patterns from many people. With the advent of Pen PCs or ordinary PCs with common tablet interfaces, however, this problem can be resolved. By employing pattern collection tools, we can gain the cooperation of multiple organizations for script collection and even take advantage of software tools available on such PCs.

2.2 DIT as writing environment

With the advent of LCD-based display integrated tablets (DIT), writing on display surfaces has become as natural

as writing on paper. Although there are still some deficiencies in DITs compared with paper, they may provide additional electronic features of editing and handwriting recognition that are unavailable on paper. We use this environment to collect handwritten patterns rather than plain tablets since writing on PDAs and on Pen or Tablet PCs is becoming increasinlgy popular, and the demand for handwriting recognition on DITs is increasing.

2.3 No data like more data

First, it is generally accepted that collecting a large amount of sample patterns is as important as elaborating recognition methods. Second, we need to collect character patterns written naturally or casually without any writing constraints so that the collected patterns can be used to train and evaluate recognition systems for real applications. Third, in the trend of personalization of information processing devices and systems, there is a need to collect a considerable amount of patterns from each individual so that the user customization and adaptation capabilities [7,15] can be tested reliably.

With these in mind, we have set the following policies for on-line handwritten character pattern collection.

2.4 Policies for pattern collection

2.4.1 Writing in boxes but recording within the page. In order to collect on-line handwritten character patterns, we display sequences of square boxes similar to a Japanese writing pad on a DIT so that the user writes characters one by one into each box. This type of user interface has been used in many commercial products to avoid character segmentation problems. Moreover, it is not as restrictive as for English since the Japanese are accustomed to this style of writing pad.

Character patterns extracted from each box are useful for developing and evaluating character recognition algorithms, although character patterns written without any writing box or grid should be collected as well for developing format-free handwriting recognition.

Although we employ writing boxes, we record pen tip coordinates within the page (whole area of the DIT) rather than within each box. In fact, we have been concurrently preparing the database of format-free on-line handwriting including text, drawing, formulas, etc., where we must record writing trajectories within the whole writing area. Therefore, recording handwriting within the page rather than within the character box provides more consistency among our databases of on-line handwriting. Moreover, this recording format enables us to test writing-box-free recognition to some extent by discarding the box information.

2.4.2 Collecting character patterns for a common text. One way to collect handwritten character patterns is to let people write whatever they want and later provide ground-truth codes to their handwriting. The merit of this method is that we can collect natural patterns; on the other hand, the labor to label ground-truth codes is high and collecting patterns for the whole set of characters is not easy.

One method to overcome these problems is to ask each participant to write characters according to a common prescribed text so that we can collect script of all the participants for the common text without having to label them. On the other hand, patterns thus collected may not be as natural as those collected by the above method. By employing a sequence of sentences as the text, however, we expect people to write naturally as described in the next subsection. This is a compromise to collect samples ranging across all the necessary categories while keeping the patterns natural.

2.4.3 As many character patterns in sentences as possible. By asking people to write characters according to a sequence of sentences, we can expect them to write natural and casual patterns in the course of reading, understanding, and reproducing sentences. It is generally recognized that if people write characters one by one without any meaningful context, their character patterns become unnaturally neat. On the other hand, they tend to write characters casually when writing sentences.

We collect sentences from various articles in Japanese newspapers so that they are not specific to any domain. We also select sentences that cover as many frequently used characters as possible. To collect scripts for all of the commonly used 3,000 to 4,000 categories from a sequence of sentences, however, we must ask people to write hundreds of thousands of characters, which is not very practical. Therefore, we ask them to write sentences that cover the frequently used characters and then to write less frequently used characters one by one.

The character patterns thus sampled according to the common sentences are also useful for evaluating recognition performance where actual character appearance probabilities, the effect of user customization, and the effect of context processing are taken into account.

2.4.4 Displaying characters to be written by font. When collecting script, character patterns should not be shown in order to collect natural patterns. For this, having people transcribe text from an audiotape would seem best. However, there are too many characters that people cannot write if they do not see them (this is called the "reading a thousand characters, writing a hundred" phenomenon). Furthermore, the text we want in Kanji (ideographic characters of Chinese origin) might be written in Kana (phonetic characters made from Kanji), while text we want in Kana might be written in Kanji due to the multitude of ways of writing Japanese expressions.

Therefore, we display font patterns above the character writing boxes and allow the participant to enlarge each font pattern by touching it with the pen since some complex Kanji patterns are difficult to see without enlargement. We understand that the participants may mimic the font patterns in their writing, but if their

writing flows naturally with the text, the patterns gained can be expected to be natural. This is again a compromise between collecting handwritten character patterns for the common text and collecting natural patterns.

Another potential problem may also be the collection of handwritten character patterns for displayed font patterns. When people are asked to write characters that they do not normally write or cannot write, their stroke order is often different from the standard and the need for stroke-order-free recognition will be overemphasized. (In an actual environment using on-line recognition, these characters would not be written. If needed, they could be written in Kana and then transliterated to the actual Kanji characters).

2.4.5 No display of recognition result. If the recognition results are shown for handwritten patterns, some people may try to write patterns intended to be easy or difficult to recognize. This hinders our objective of collecting natural and casual patterns. In order to avoid this problem, no display of recognition result is made during pattern collection.

2.4.6 No restriction on writing styles and quality of patterns. Before a participant starts to write characters, we explain the purpose of collecting natural and casual handwritten character patterns and do not impose any restriction on writing styles or quality of patterns except for asking that they write characters naturally. We pay a small and fixed amount of money (10,000 yen for students) to each participant for the whole writing task so that they try to finish the task as quickly as possible. We expected this to result in handwritten patterns that are more casual rather than neater. The total time required per person for writing all the patterns was generally about 10 h, though the shortest time was about 5 h. These times are comparable to the speed of writing with pen and paper.

2.4.7 Recording the participant profile and collection environment. We record the participant profile including their sex, birthday, native language, dominant hand, writing hand, writing period, etc. and the specifications of the collection environment such as input resolution, display resolution, sampling rate, etc. These data are recorded because the handwritten patterns may later be analyzed with respect to these factors.

2.4.8 Verifying handwriting by both machine and human. As learning patterns or benchmark patterns, patterns inconsistent with ground truth are useless and even harmful. Therefore, we inspect collected patterns by both machine and human. Human inspection is generally superior to machine verification, but humans often overlook wrong characters in a meaningful context. Therefore, we first do a machine verification to pick up omitted and erroneous patterns and then employ human inspection free from the context side effect. After verification, each participant verifies detected patterns and fills omissions or rewrites them if he/she accepts them as wrong patterns. Thus, in a strict sense incorrect patterns may exist, but we believe that they should be accepted as long as other people can read them since they are the patterns actually written in daily life.

3 Categories for script collection

Basically, we need scripts of characters that are in everyday use. They are Kanji, Hiragana, and Katakana, called Kana collectively, alphanumerals, and symbols. Kanji characters have complex shapes and are written by many strokes, while Kana characters have simple shapes and are written by a few strokes (generally from one to four strokes and up to six strokes to make a dull sound).

The first database Kuchibue_d was intended to collect script for the JIS (Japanese Industrial Standard) first level set of Japanese characters. The first level set covers most of the daily used characters. However, Cyrillic characters, symbols with a very low degree of occurrence (those that appeared less than ten times in a corpus of approximately 13 MB collected from newspapers, graduate theses, etc.), symbols normally written with multiple characters (“ TEL ”, “ ㌫ ”, “ kg ”, etc.) and filled-in symbols (“ ★ ”, “ ● ”, etc.) were excluded.

The sentences used for character pattern collection were taken from the 1993 CD-ROM edition of the Asahi newspaper (Japanese language newspaper), which included 1,227 frequently appearing character categories with the result that they were composed of 10,154 characters and included 1,537 categories in the JIS first level set. Here, “frequently appearing” refers to categories that were detected at least 50 times in the abovementioned 13-MB corpus. Eleven characters in the second level of JIS (熙焙漱釉剎咤誦軻儷拗鍼) that appeared in the sentences were not changed to Hiragana but left to be written as such. There were approximately 200 sentences, so that each sentence contained 50 characters on average. The remaining 1,808 categories in the JIS first level (non-Kanji: 102 categories, Kanji: 1,706 categories) were appended at the end. For this text, the average writing time per person is approximately 10 h. If we try to prepare sentences that cover all of the JIS first level categories, it would come to around 100,000 characters, making the script collection an unrealistic task.

The second database Nakayosi_t extended the categories and included 1,093 categories used in names from the JIS second level set. The sentences for collecting 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 script collection. The time required for writing was almost the same as Kuchibue_d. Table 1 summarizes the organization of the two script collection texts. Figure 1 shows the beginning and ending portions of the Nakayosi_t text for script collection.

Table 1. Organization of the texts for collecting character patterns

	Kuchibue_d	Nakayosi_t
Total characters	11,962	10,403
Kanji/Kana/	5,643/5,068/	5,799/3,723/
symbols/alphanumerals	1,085/166	816/65
Total character categories	3,356	4,438
Characters in sentences	10,154	7,376
Categories in sentences	1,548	1,411
Categories after sentences	1,808	3,027

仲良しは近所に住む猫の花子。夜は庭の小屋で寝るが、昼間は片時も離れない。冷蔵庫をのぞき込む。一年一ヶ月、航海をともにして静岡県の清水港に到着。僕は順位戦のうっぷんをここで晴らしている。正月も休まずの操業だ。毎日午後になると、「アー」と鳴いて催促する。

(a) beginning part

, . : ; ^ ¯ _ ヽ ゞ 〃 ／ ＼ ‖ | ‥ □ 〔 〕 ［ ］ ｛ ｝ 〈 〉
《 》 ＋ － ± × ÷ ＝ ≠ ＜ ＞ ≦ ≧ ∞ ∴ ♂ ♀ ° ′ ￥ ＄ ＃ ＆ ＊ ＠
§ ☆ ◎ ◇ □ △ ※ 〒 → ← ↑ ↓ ∧ ∨ 6 7 A B C D F G H I K
N O Q R U W X Y Z a b c d e f g h i j k l m n o p
q r s t u v w x y z ぃ ぅ ぉ ぢ ぴ ぺ ぽ ゅ わ ゐ ゑ ィ ゥ ゾ チ

(b) ending part

Fig. 1. Beginning and ending portions of the Nakayosi_t text

4 Software tools and text preparation

We developed script collection and verification tools for the MS Windows operating system since it is a common platform for pen computing.

4.1 Script collection tool

We made a tool for collecting on-line handwritten character patterns on a DOS/V machine connected with a display integrated tablet (DIT) (including Pen PC) and the MS Windows operating system. By specifying a text file, the corresponding character string is displayed along with the writing boxes and the participant writes characters as shown in Fig. 2. The pen trajectory is recorded within the whole area of a DIT as depicted in Fig. 3. The participant can quit writing and later resume it since it is very hard for people to continue writing for several hours.

This tool is also used when the participant fills missing characters or rewrites characters judged erroneous by the verification tool described next. When this tool reads a message file of erroneous patterns, the participant can navigate to skip from one erroneous pattern to the next.

The size of each character writing box was set to 60×60 dots in the display coordinates, that is, about 1.7×1.7 cm on a 9.5-in. VGA (640×480) DIT and 1.43×1.43 cm on a 12-in. XGA (1024×768) DIT. They almost correspond to a large-boxed manuscript paper and a small-boxed manuscript paper, respectively.

Fig. 2. Screen for collecting character patterns

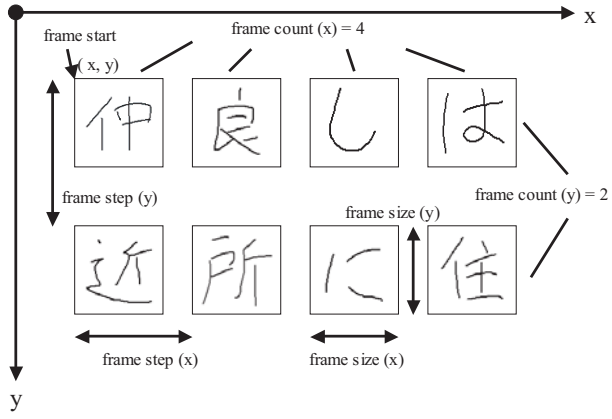


Fig. 3. Handwriting recording coordinates

The initial version recorded the pen trajectory in the display (mouse) coordinates, but later it was enhanced to record it in either display or tablet coordinates. Although both of them produce pen tip coordinates at a fixed time interval depending on the tablet, the display coordinate system produces new coordinates only when the pen tip moves, i.e., it does not produce multiple coordinates when the pen is stationary, while the tablet coordinate system produces coordinates constantly even when the pen is at the same position. Moreover, the tablet coordinates have resolutions roughly ten times finer than those of the display coordinates.

Kuchibue_d was recorded using display coordinates, while Nakayosi_t used tablet coordinates. The postfix of their names denotes the corresponding coordinate system.

When a participant uses this tool, he/she is requested to write his/her name, sex, occupation, mother tongue, writing hand, dominant hand, motivation, birthday, and comments before starting to write text. On the other hand, the data administrator's name, affiliation and e-mail address, starting date and time, closing date and time, the text file name for writing characters, arrangement of writing boxes (starting position, pitch, size, and number of boxes in a line and number of lines in the whole DIT), DIT information (make, model, display res-

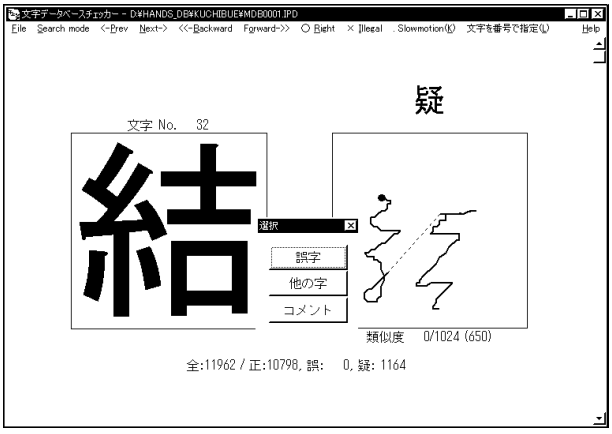


Fig. 4. Screen for verifying character patterns

olution, input resolution, physical size, and sampling rate) are automatically recorded into the data file.

4.2 Verification tool

In order to inspect on-line handwritten character patterns, a verification tool was made that employs a character recognition engine [6] and suggests for human inspection and confirmation those character patterns that are judged erroneous. When a human inspector judges a pattern as erroneous, he or she can type a message as to why it is erroneous, which will be shown to the participant if required, when the participant is rewriting it.

At first, an on-line handwritten character pattern was judged "erroneous" if the correct category was outside the top ten candidates of the recognition result. Later, however, this condition was changed to require the pattern similarity to the correct character category to be smaller than a certain threshold. In the former method the input pattern must be recognized, i.e., matched with all the standard patterns, so that it took more time than the latter method. We confirmed that there is no problem with the latter method if the similarity threshold is set small enough in matching with the correct category. Therefore, the latter method was adopted. The inspection screen is shown in Fig. 4.

The verification method has proved quite useful. If only a recognition engine is employed for pattern inspection, it is not reliable. Too many correct patterns would be judged erroneous by machines. On the other hand, human inspection alone is unreliable since humans do not notice wrong characters when reading meaningful text. A recognition engine sensitive to pattern deformation picks up almost all the erroneous patterns as well as some correct patterns and the human can verify them without context.

Here, we must also consider the case when the verification tool makes false positive errors, i.e., a character that was written incorrectly but detected as correct. In this case, the incorrect writing may go into the database.

When we use a recognition engine, it may make recognition errors with a certain probability. However, the probability that an incorrectly written pattern will be misrecognized in a specified category rather than any other categories is quite small. By setting the similarity threshold high, we can further decrease this risk. Moreover, the purpose of this tool is not to pick up all the incorrect patterns. We wish to keep natural and casual patterns that may be incorrect in a strict sense but that people write every day and others have no problem in reading (say, a short stroke written in the opposite direction in a complex Kanji pattern).

Fortunately, no reports on truly incorrect patterns or missing patterns have been made by our database users so far.

5 Script collection and release

This section describes the details of pattern collection for Kuchibue_d and Nakayosi_t databases and their release.

5.1 Kuchibue_d

We collected on-line handwritten character patterns from 30 participants in our laboratory. Then we offered to share the database with other organizations, with the stipulation that each organization should provide us character patterns from five additional participants. We provided them with the script collection tool. Nine organizations joined the project initially. Meanwhile, we ourselves added data from five additional participants. Our group inspected all the patterns, and the person who wrote the initial pattern was requested to rewrite any missing or wrong pattern. By February 1996 we had collected patterns from 80 people with 11,962 patterns per person, which were made available to all the collaborating organizations. Later, eight more organizations joined the project, with the result that patterns from 120 people, with each person contributing 11,962 character patterns, were collected.

5.2 Nakayosi_t

Nakayosi_t was mainly compiled from students of our university (mostly male) and female students of another women's university in order to achieve a gender balance. We also welcomed four organizations that initially requested to join the Kuchibue_d project into the Nakayosi_t project since the Kuchibue_d was already being used by the collaborating organizations. We followed the same scheme of pattern verification as for the Kuchibue_d project. As a result, patterns from 163 people with 10,403 patterns per person were collected.

5.3 Ideal script set

We compiled a set of correct stroke-number and correct stroke-order patterns for all the collected categories in the JIS first level and all 4,152 categories in the JIS second level, resulting in a total of 7,723 categories. This set

Table 2. Participant occupations

Occupation	Kuchibue_d		Nakayosi_t	
Company employee	46%	(56)	6%	(10)
Student	36%	(43)	92%	(150)
Teacher	15%	(2)	0%	(0)
Other	2%	(18)	1%	(2)
Unknown	1%	(1)	1%	(1)
Total	100%	(120)	100%	(163)

was compiled with the abovementioned script collection tool using the display coordinates but within a larger writing box (i.e., 120×120 pixels) than the one used for pattern collection. For each category, a script pattern was recorded without using any sentence context. The analyses for character pattern variations make use of this set.

Strictly speaking, the standard stroke order to write Kana and Kanji is defined for a slightly smaller set than the JIS first level set, which can be extended to almost all of the JIS second level set of Kanji characters since they share radicals. However, these standards do not cover symbols and foreign letters. Consequently, we consulted various textbooks to determine the stroke order for writing these letters and used them in our ideal script set.

5.4 Statistics on participant profiles

Distribution of the participants' age is shown in Fig. 5. The average age of the participants for Kuchibue_d is 28.08, with a variance of 47.94: the oldest is 57 and the youngest 18. The average age for Nakayosi_t is 23.1, with a variance of 23.02: the oldest is 58 and the youngest 19. Kuchibue_d covers a wider range of participants with respect to age.

User occupations are shown in Table 2. Since Kuchibue_d has been compiled mainly from the collaboration among companies, about half of its total patterns are from employees of those companies, while Nakayosi_t is compiled mostly from patterns written by students, as explained in Sect. 5.2.

As for the gender of the participants, Kuchibue_d is not well balanced: it is composed of 85 males (71%) and 35 females (29%), while Nakayosi_t is gender balanced, with 82 males (50%) and 81 females (50%). As for the mother tongue, Kuchibue_d is composed of 118 people with Japanese (98%) and two people with Chinese (2%) as their mother tongue. On the other hand, some foreign students also contributed to Nakayosi_t, with the result that as the mother tongue, 150 people have Japanese (92%), 11 people have Chinese (7%), 1 has Malay (1%), and 1 has Bengali. Regarding left- and right-handedness, Kuchibue_d contains handwriting of 117 right-handed people (98%) and 3 left-handed people (2%), while Nakayosi_t contains samples from 161 right-handed people (99%) and 2 left-handed people (1%).

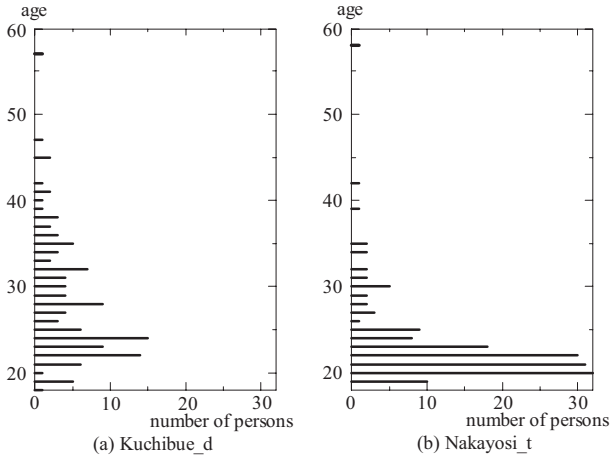


Fig. 5. Distribution of participants' age

5.5 Format

The databases were initially encoded in a binary format and accessed through libraries. When using the binary format, each participant's patterns could be stored on a 3.5-in. floppy disk, which was most popular when this project was started. Moreover, abstraction of the databases by access libraries seemed to be effective if we need to change the internal format.

It has turned out, however, that the binary format and access libraries are difficult to use on platforms other than MS Windows. Therefore, we have now made the databases available in a simple text format and in the UNIPEN format [3]. Recently available compaction methods reduce their sizes even more than the original binary versions.

5.6 Release

The full sets of Kuchibue_d ($120 \times 11,962 = 1,435,440$ patterns) and Nakayosi_t ($163 \times 10,403 = 1,695,689$ patterns), a total of 3,131,129 patterns, are now available from the bookstore of Tokyo University of Agriculture and Technology (http://tuat.coop-bf.or.jp/∼nakagawa/). Moreover, ten people's patterns in Kuchibue_d are freely available for research purposes (http://www.tuat.ac.jp/∼nakagawa/ipdb/). We can distribute this free version only by CD-ROM on the condition that the newspaper company can enclose the printed copyright for the use of its text in sampling character patterns. At present, more than 25 groups other than the original collaborators are using this version, which includes more than 10 groups from abroad.

6 Analysis of stroke-number varation

In this section, we present the results of analyzing stroke-number variation (SNV) in on-line handwritten character patterns stored in Kuchibue_d and Nakayosi_t.

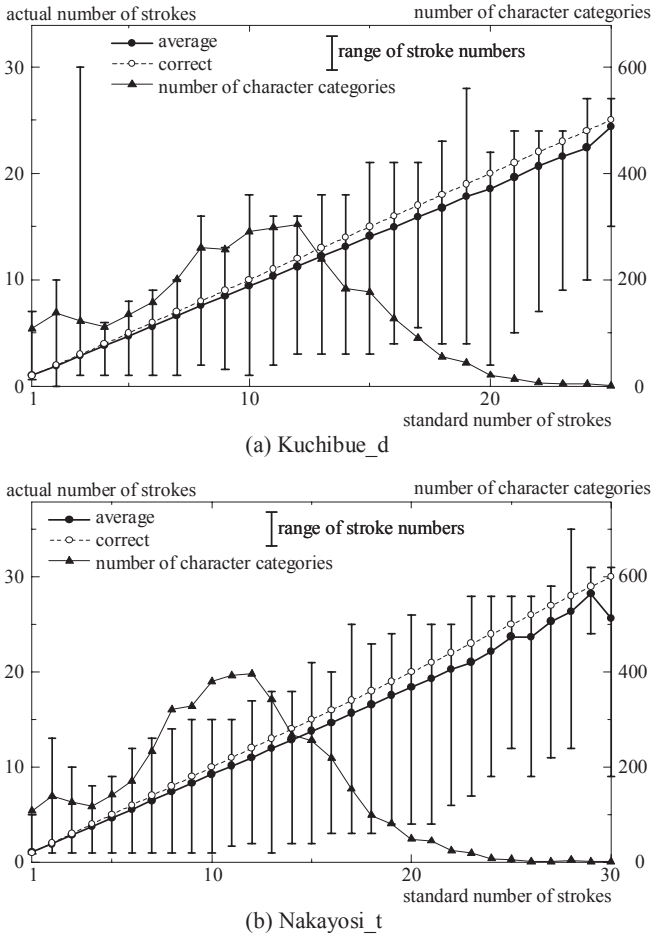


Fig. 6. Range of the number of strokes

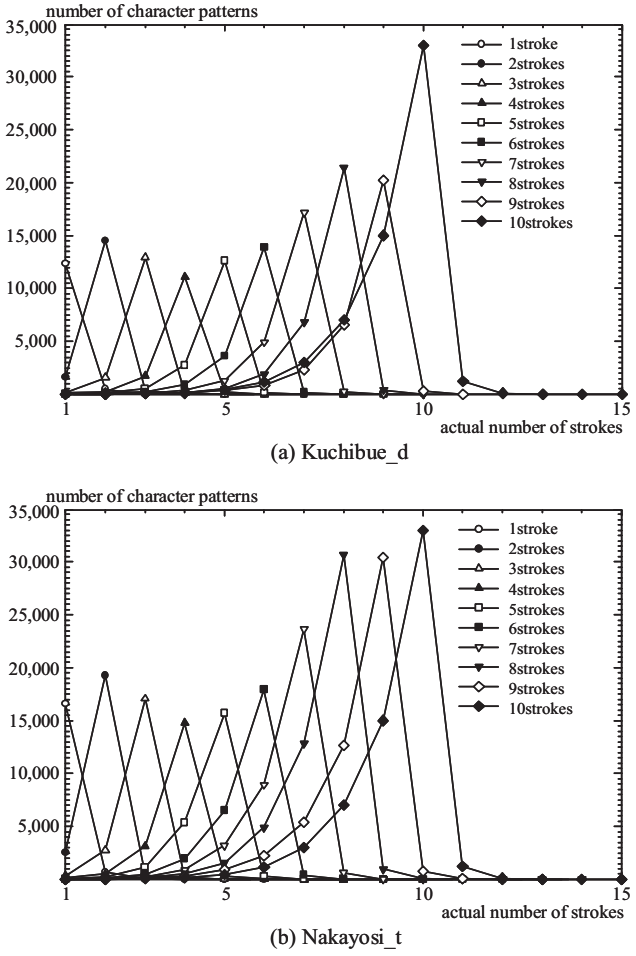


Fig. 7. Distributions of the stroke count (1–10 strokes)

6.1 Statistical view

Figure 6 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 mean and the actual range of stroke numbers are displayed on the vertical axis. The number of character categories for each correct stroke count is also shown in Fig. 6.

When people write characters casually, strokes are often connected but are sometimes broken into two or more parts. Thus, the actual number of strokes in a handwritten character pattern may be more or less than the standard. However, this range of variations is not as small as it had been supposed until recently for on-line handwritten character recognition. An extremely large range of SNVs for three-stroke patterns in Fig. 6a is due to the repeated pen-up and pen-down movements needed to mark out the three dots in the symbol shown at the top of Fig. 9a.

Figures 7 and 8 show more detailed analyses of the range and frequencies of actually occurring stroke numbers for character patterns with a standard stroke count between 1 and 10, and 11 or above, respectively. In these figures, "N strokes" means that the standard character pattern is written by N strokes in the correct way of writing.

These figures show that as the correct number of strokes becomes larger, the number of actual strokes varies more often and more widely in the range below the correct number, but there is a sharp decline in the variations above the correct number. In other words, character patterns consisting of many strokes are often abbreviated or written by connecting successive strokes.

Table 3 shows another analysis of the stroke count distribution. Although no constraint is placed on the writing style, more than half of the character patterns were written with the correct number of strokes. This may seem surprising, but there are two possible explanations for it. One is that most of the collected patterns are embedded in a sentential context so that they include many more Kana characters consisting of a few strokes than the complex Kanji characters. Patterns consisting of a small number of strokes are more likely to be written with the correct number of strokes. The other factor is that many Japanese people learn calligraphy in their childhood and continue this habit of writing characters beautifully. We have observed several people whose handwriting is quick and yet neat. On the other hand, when they deviate from the correct number of strokes, it is usually in the direction of fewer strokes.

Nakayosi_t shows a wider distribution than Kuchibue_d because the former includes about 1,100 JIS

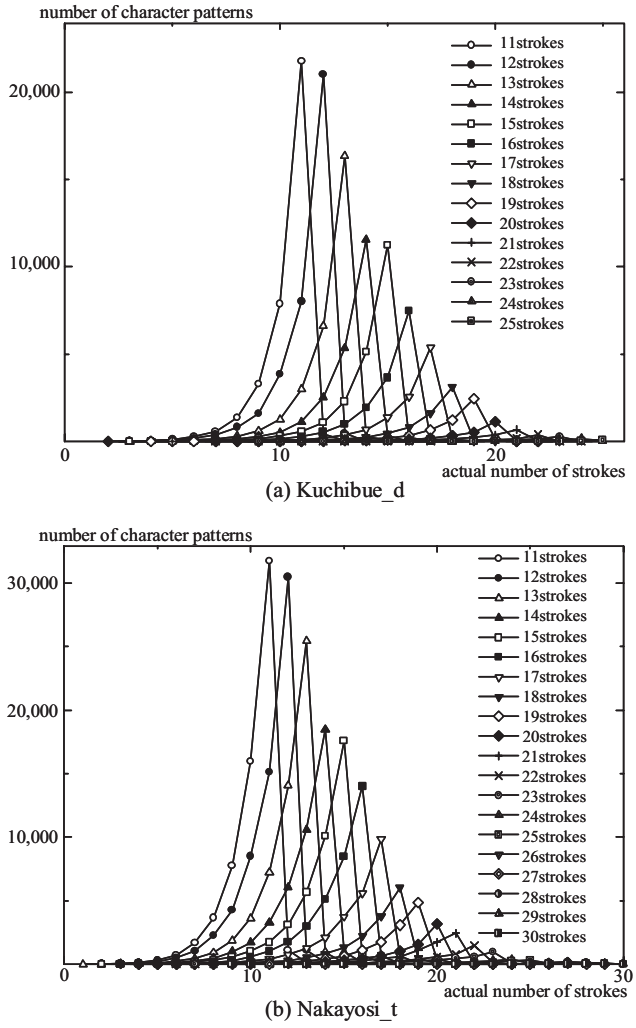


Fig. 8. Distributions of the stroke count (11–30 strokes)

Table 3. Proportion of stroke-number variations

	correct	decrease	increase
Kuchibue_d	64.90%	33.71%	1.39%
Nakayosi_t	54.12%	43.71%	2.17%

second set characters whose patterns have high stroke counts, which yields larger variations.

6.2 Categories with a large degree of SNV

We determined the smallest and the largest number of strokes for every category in the database to find the character categories with large SNV (stroke-number variations). Figure 9 shows the categories with SNVs of 14 or more. The three dots symbol appears here because some people write it by repeating the pen-down and pen-up motions to make the dots bigger.

Most of the patterns include the radicals (subpatterns) shown in Fig. 10. Those radicals are often written cursively so that we observe a large degree of SNV between neatly written and cursively written characters.

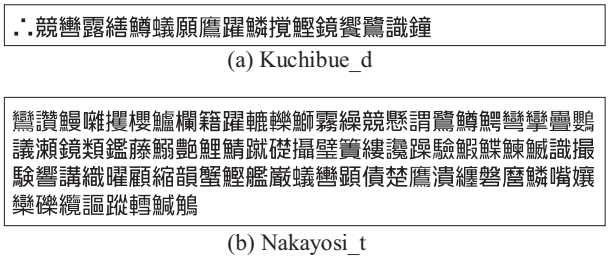


Fig. 9. Character categories written with large stroke-number variations

(a)糸 (b)幺 (c)口 (d)鳥 (e)魚 (f)灬

Fig. 10. Radicals written with large stroke-number variations

Fig. 11. A pattern written with nine strokes

Figure 11 shows the smallest stroke count pattern for the category with the largest SNV. When written correctly, it is composed of 30 strokes, but the pattern shown in Fig. 11 is written with 9 strokes.

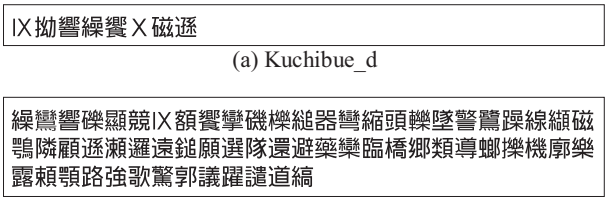
6.3 Categories with large SNV with respect to the standard

In the previous section, we showed some character categories with large SNVs. In this section, we will show some categories that have large SNVs with respect to the standard. Since character patterns with correct stroke numbers also appear, these two categories mostly overlap.

6.3.1 Categories written with fewer strokes. Figure 12 shows character categories whose average stroke numbers, when actually written, are less than the standard by three or more strokes. Many of them include the radicals shown in Fig. 10a and b, and some include those shown in Fig. 10c and d. Figure 13 shows some additional radicals that often appear in these categories. People often write them without picking up the pen, thereby producing cursive patterns.

On average, the character category with the largest stroke deviation has 4.79 strokes less than the standard. Figure 14 shows a pattern from this category written with 4 strokes. When written correctly, it is composed of 19 strokes.

6.3.2 Categories written with more strokes. Figure 15 shows character categories actually written in the databases with an average of at least 0.2 strokes more than the standard.



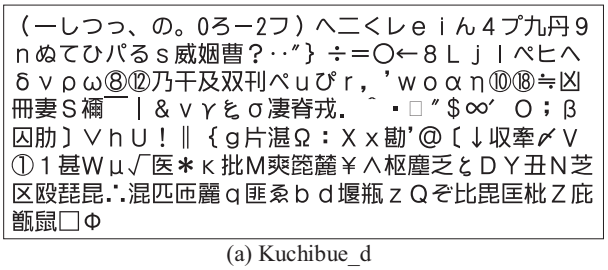
(a) Kuchibue_d

(b) Nakayosi_t

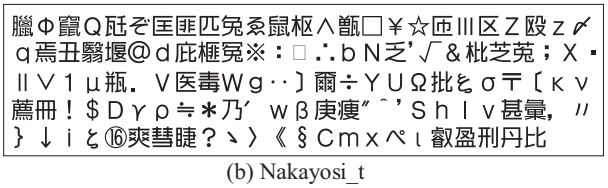
Fig. 12. Categories written with fewer strokes than the standard

Fig. 13. Radicals written with large stroke-number variations

Fig. 14. A pattern written with four strokes



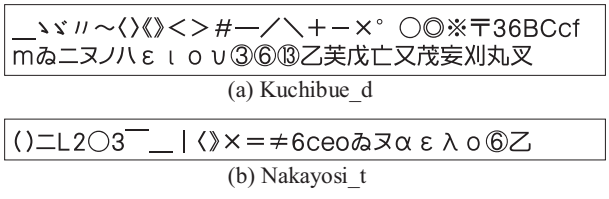
(a) Kuchibue_d



(b) Nakayosi_t

Fig. 15. Categories written with more strokes than the standard

These categories are (1) one-stroke characters for which the stroke count can only increase (irregular split of a stroke, added strokes after writing, and so on); (2) symbols, especially those containing small dots and Greek letters, for which people often add strokes to make the right shape; and (3) Kanji characters containing unfamiliar radicals. For the second and third groups, these characters are not learned in school. People do not normally write these characters, so that many people copy the font patterns when requested to write them. Consequently, several single strokes are written by two or more strokes and additional strokes are added for shaping them appropriately, with the result that the actual number of strokes exceeds the standard.



(a) Kuchibue_d

(b) Nakayosi_t

Fig. 16. Character categories written without any stroke-number variation

6.4 Categories written with the correct number of strokes

Figure 16 shows character categories actually written without any SNV in the databases. They are (1) one-stroke characters or symbols and (2) some simple Kanji or Kana characters not used often. The characters in the second group are usually written carefully stroke by stroke, as they are not familiar characters, but because they are not very difficult to write, additional strokes are unnecessary. On the other hand, the stroke count of one-stroke characters/symbols can only increase. (If they are written with a fewer number of strokes, i.e., with a null stroke, they are detected by the verification tool as a missing character and are written again). Because the number of strokes is one, however, their probabilities of stroke increase are smaller than the characters consisting of many strokes. Therefore, their SNV are confined to a very narrow range.

In any case, Fig. 16 shows only the result of writing (samples) by the participants to our databases rather than the patterns that may occur across a large population, as the other studies show. People in general and even the participants may write them using more than the standard number of strokes to arrange their shapes as discussed above. Therefore, the same character may appear in Figs. 15a and 16b, or in Figs. 15b and 16a, if we relax the threshold for Fig. 15. However, the same character cannot appear in both Figs. 15a and 16a or in both Figs. 15b and 16b since character categories from each database are classified into either of them.[1]

7 Analysis of stroke-order variations

In this section, we present the result of analyzing stroke-order variations (SOV) in on-line handwritten character patterns stored in Kuchibue_d and Nakayosi_t.

7.1 Statistical view

In order to investigate SOV, we utilized our recognition engine [6], which uses a strict stroke-order dictionary containing a single correct stroke order for each character based on the ideal script set and a general dictionary

[1] The careful reader may find similar looking patterns "◯" in both Figs. 15a and 16a, but they are actually different characters. The first one in Fig. 15a is the Kanji numeral denoting "zero", while the second occurrence in Fig. 15a is the letter "O", while the occurrence in Fig. 16a is a symbol.

```
if(score_G - score_S >= 30) then
    Detect
endif
if(score_S <= 800)
    Detect
endif
```

Fig. 17. Logic for detecting suspicious patterns

where several SOVs are registered for each subpattern and shared among character patterns.

If the recognition engine using the general dictionary produces a certain score (score_G) for an input pattern with respect to the correct category, but using the strict dictionary results in a lower score (score_S), this implies that the input pattern may contain SOV. Moreover, an input pattern whose score_S is less than a certain threshold may also contain SOV. Figure 17 shows the logic of detecting suspicious patterns. With the highest score of the recognition engine being 1024, the margin at which score_G is judged to be larger than score_S is set at 30 and the threshold for score_S is set at 800 (we know empirically that a pattern recognized with a score higher than 800 is a correct and neat pattern with high probability).

The basic idea is to detect all such patterns and inspect their stroke order visually. However, the databases are so large that visual inspection is not feasible. Therefore, we select a few data sets (the term "data set" denotes a collection of patterns given by each participant) that represent the databases and determine the ratio of detected patterns that represents truly wrong stroke order patterns. We apply the detection to all the data sets in each database and multiply the average ratio by all the detected patterns to estimate the total number of patterns with SOV.

We selected the three sets in each of Kuchibue_d and Nakayosi_t, respectively. They produce average recognition rates and the majority of data sets produce similar recognition rates with a small number of exceptional data sets, so that they are expected to represent the databases.

The result of the detection is shown in Table 4. A significant number of Kanji characters may include SOV, while Kana characters seem considerably more stable. Moreover, nearly half of the symbols may include SOV. This is reasonable since the stroke order of symbols is not defined. Therefore, we consider SOV for Kanji characters. We applied the visual inspection to the detected Kanji characters as shown in Table 5.

In Kuchibe_d, about 54% of detected Kanji patterns have SOV, and in Nakayosi_t about 49% of detected Kanji patterns have SOV. SOV may also exist among undetected patterns. We applied the detection process to all the data sets of Kuchibue_d and Nakayosi_t, with the result that 32% of Kanji characters in Kuchibue_d and 36% of Kanji characters in Nakayosi_t were detected on average, respectively. From these numbers, we estimate that at least $0.32 \times 0.54 \times 100 = 17.3\%$ of all the Kuchibue_d Kanji patterns and at least $0.36 \times 0.49 \times 100 = 17.64\%$ of all the Nakayosi_t Kanji patterns are written with a nonstandard stroke order.

Figures 18 and 19 show categories written with incorrect stroke orders and radicals that often appear among the incorrect stroke-order patterns, respectively. These radicals seem to be responsible for causing SOV.

Table 4. Detected patterns

(a) Kuchibue_d

Data set	No.(#) of detected patterns			
	Kanji (#/ 5,643)	Kana (#/ 5,068)	Symbols (#/ 1,085)	Alpha-numerals (#/166)
MDB0033	1,668	243	432	37
MDB0037	2,177	780	291	33
MDB0052	1,523	673	365	59

(b) Nakayosi_t

Data set	No.(#) of detected patterns			
	Kanji (#/ 5,799)	Kana (#/ 3,723)	Symbols (#/ 816)	Alpha-numerals (#/ 65)
NKY0008	1,978	289	523	12
NKY0043	1,677	299	531	15
NKY0068	2,263	184	515	11

Table 5. Verified patterns

(a) Kuchibue_d

	Detected	Verified as correct stroke order	Verified as incorrect stroke order
MDB0033	1,668	750 (45.0%)	918 (55.0%)
MDB0037	2,177	1,015 (46.6%)	1,162 (53.4%)
MDB0052	1,523	683 (45.0%)	840 (55.0%)
Average	1,790	817 (45.6%)	973 (54.4%)

(b) Nakayosi_t

	Detected	Verified as correct stroke order	Verified as incorrect stroke order
NKY0008	1,978	1,239 (62.6%)	739 (37.4%)
NKY0043	1,677	927 (55.3%)	750 (44.7%)
NKY0068	2,263	828 (36.6%)	1,435 (63.4%)
Average	1,972	998 (50.6%)	975 (49.4%)

威違盈燕感虐虚犀剤斉牒凸薄縛櫨敷淵蔑簿命幽慮虜熙粟帷閏衛越堰欧殴鴎過臥階樫鰹樺憾監艦鑑犠巨拒渠距勤緊区躯駆虞劇堅賢減虎護昆混済堺珊鹿嫉繍衝情審臣腎垂枢性牲戚全痩臓蔵唾竪跳登匿博藩被枇昆畢姫膚陛郵要濫藍覧離臨麟

Fig. 18. Character categories written with incorrect stroke orders

(a)臣 (b)巨 (c)区 (d)比 (e)广 (f)皿

Fig. 19. Radicals causing stroke-order variations

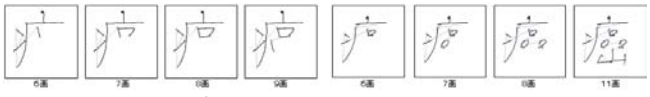


Fig. 20. Incorrect stroke order caused by pattern simplification

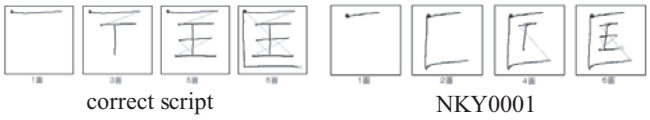


Fig. 21. Incorrect stroke order caused by uniting bracketing radicals

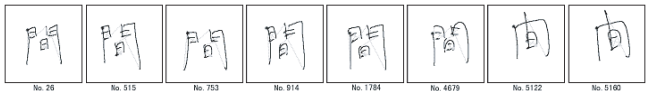


Fig. 22. Neat patterns and simplified patterns for the same character

7.2 A closer look at incorrect stroke-order patterns

As shown in Fig. 20, many people simplify the "mouth" radical shown in Fig. 10c by a single-stroke square, thereby changing its stroke order. Very often, pattern simplification accompanies change in stroke order.

Figure 21 shows some radicals that are often written by consecutive strokes as a unit, although the correct order is to insert another radical or radicals in between, for example in brackets.

Patterns with SNV within individual handwriting patterns are pretty common, but those with SOV within them are 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. Figure 22 shows such an example where after writing neatly the participant writing the pattern has switched to a simplified form.

8 Suggestions for the design of recognizers

Although there are some differences between Kuchibue_d and Nakayosi_t, they reveal many important characteristics of handwriting.

With regard to SNV, as the number of strokes in a character pattern increases, the range and occurrences of the actual number of strokes made by the participants that are less than the correct number increase, while the range and occurrences of the actual number of strokes that are more than the correct number drops sharply. This fact implies that the matching algorithm should be robust with respect to stroke connections. This fact can also be utilized to narrow down the search space. Moreover, some radicals cause a large reduction in stroke number and deformation, so that it would be useful to allow the system to have multiple representations for them.

The SOVs are mainly caused by common variations and added strokes. The common variations can be registered in the character pattern dictionary. Some structural organization of the dictionary that allows sharing of the radical variations among character patterns containing that radical would make the registration more efficient [7]. On the other hand, added strokes are not easily treated by on-line recognizers. The best solution would be to employ or combine off-line recognizers that are essentially stroke-order free [11,13]. A more serious problem in SOV is in writing symbols. Their writing orders are not defined, so they are often written with unusual stroke orders, which makes their on-line recognition difficult. To make matters worse, they also have large SNVs, with the actual stroke numbers being more than the correct number. Since they have simple shapes and not enough distinguishing features, their recognition by off-line methods is also not easy when many symbols must be distinguished. Moreover, there is little contextual information available for symbol recognition. Therefore, a distinct treatment might be necessary.

9 Conclusion

In this paper we described the design of on-line handwritten Japanese character pattern databases, focusing on software tools, method of pattern collection, and analysis of patterns. Our efforts resulted in two databases, one with 11,962 patterns from each of 120 participants and the other with 10,403 patterns from each of 163 participants, with a total of over 3 million patterns. As a matter of principle, we collected handwritten character patterns in the context of sentences as much as possible, and we verified the collected patterns by both machine and human inspection to ensure a high reliability of our databases. So far we have not received any report from the users of our databases on incorrectly tagged or missing patterns. The sentences for character patterns were compiled from a major newspaper to include frequently appearing characters, while less frequently used characters were written one by one without any sentential context. This was a practical compromise. The same text was used for collecting script patterns from all the participants. The patterns were inspected, and the participants were asked to rewrite any omitted or incorrectly written patterns if they agree that it was mistaken. Then we applied the same scheme to a different text so as to include more character categories while suppressing the amount of text. Analyses of the stored patterns were presented to show the quality of collected patterns and to consider future research on character recognition. They reveal that as the number of strokes in a character pattern increases, the variability in the number of strokes in the handwritten version of that character increases widely, with the handwritten pattern having fewer strokes than normal: only in 1–2% of cases did the number of strokes in the handwritten character exceed the normal, and that mostly for symbols and unfamiliar Kanji patterns whose stroke orders are not taught in school. Also, stroke-order variations

are mostly caused by common habits and strokes added after writing a character, and unusual stroke orders are often employed for writing symbols. Although the stroke orders for writing common sets of Kanji and Kana are defined and standardized in Japanese education, those for symbols are not. This makes people often employ any writing order when writing symbols, which makes their on-line recognition difficult when many symbols must be distinguished.

Needless to say, the analyses presented here were made from a certain perspective, and there remain many other aspects for further analysis.

Acknowledgements. We would like to express our sincere thanks to all those who have contributed to pattern collection projects. Companies (in alphabetical order): Brother, Canon, Fujitsu, Fuji Xerox, Hitachi, Hitachi Software Eng., IBM Japan, Matsushita Electric, Mitsubishi Electric, NEC, Oki Electric, Ricoh, Sanyo Electric, Seiko Epson, Sharp, and Toshiba and university laboratories (in alphabetical order): Matsumoto Laboratory at Waseda University, Miyahara Laboratory at Nagasaki University, Nakano Laboratory at Shinshu University, Sakoe Laboratory at Kyushu University, and Yamamoto Laboratory at Gifu University each contributed patterns from five persons. Professor Ichimura at Tokyo National College of Technology and Associate Professor Takada at Edogawa University helped us to collect patterns from their students. Most patterns from female students in Nakayosi_t were collected at Tsuda University. Thanks are due to Kishi Laboratory and its students. Thanks are also due to our students: K. Akiyama, L. Higashigawa, T. Higashiyama, Y. Yamanaka, Y. Nishimura, and T. Fukushima for their involvement in the collection, verification, and maintenance of the databases. Preparation for collecting patterns was partially funded by the Grant-in-Aid for Scientific Research under contract number 05558027 and 07207207 and the collection of Nakayosi_t was funded by the Advanced Software Enrichment Project of IPA under MITI, Japan.

References

1. Guyon I, Schomaker L, Plamondon R, Liberman M, Janet S (1994) UNIPEN project of on-line data exchange and recognizer benchmarks. In: Proceedings of the 12th ICPR, 2:29–33
2. Hull JJ (1994) A database for handwritten text recognition research IEEE Trans PAMI 16(5):550–554
3. Jaeger S, Nakagawa M (2001) Two on-line Japanese character databases in Unipen format In: Proceedings of the 6th ICDAR, pp 566–570
4. Matsumoto K, Fukushima T, Nakagawa M (2001) Collection and analysis of on-line handwritten Japanese character patterns. In: Proceedings of the 6th ICDAR, pp 496–500
5. Nakagawa M (1990) Non-keyboard input of Japanese text – on-line recognition of handwritten characters as the most hopeful approach. J Inf Process 13(1):15–34
6. Nakagawa M, Akiyama K, Tu LV, Homma A, Higashiyama T (1996) Robust and highly customizable recognition of on-line handwritten Japanese characters. In: Proceedings of the 13th ICPR, 3:269–273
7. Nakagawa M, Tu LV (1996) Structural learning of character patterns for on-line recognition of handwritten Japanese characters. In: Perner P et al (eds) Advances in structural and syntactic pattern recognition. Lecture notes in computer science, vol 1121. Springer, Berlin Heidelberg New York, pp 180–188
8. Nakagawa M, Higashiyama T, Yamanaka Y, Sawada S, Higashigawa L, Akiyama K (1997) On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions. In: Proceedings of the 4th ICDAR, pp 376–381
9. Saito T, Yamada H, Yamamoto K (1985) On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis (in Japanese). Trans IECE Jpn J68-D(4):757–764
10. Smith SJ, Bourgoin MO, Sims K, Voorhees HL (1994) Handwritten character classification using nearest neighbor in large databases. IEEE Trans PAMI 16(9):915–919
11. Tanaka H, Nakajima K, Ishigaki K, Akiyama K, Nakagawa M (1999) Hybrid pen-input character recognition system based on integration of on-line-off-line recognition. In: Proceedings of the 5th ICDAR, pp 209–212
12. Tappert TC, Suen CY, Wakahara T (1990) The state of the art in on-line handwriting recognition IEEE Trans PAMI 12(8):787–808
13. Velek O, Jaeger S, Nakagawa M (2002) A new warping technique for normalizing likelihood of multiple classifiers and its effectiveness in combined on-line/off-line Japanese character recognition. In: Proceedings of the 8th IWFHR, pp 177–182
14. Viard-Gaudin C, Lallican PM, Knerr S, Binter P (1999) The ireste on-off (Ironoff) handwritten image database. In: Proceedings of the 5th ICDAR, pp 455–458
15. Yokota T, Kuzunuki S, Gunji K, Hamada N (2001) User adaptation in handwriting recognition by an automatic learning algorithm. In: Proceedings of HCI International 2001, 1:455–459

Masaki Nakagawa was born on 31 October 1954 in Japan. He received his B.Sc. and M.Sc. from the University of Tokyo in 1977 and 1979, respectively. During the 1977–1978 academic year he pursued a degree in computer science at Essex University in England and received his M.Sc. with distinction in computer studies in July 1979. He received his Ph.D. in information science from the University of Tokyo in December 1988. Since April 1979 he has been working at Tokyo University of Agriculture and Technology. Currently, he is a professor of media interaction in the Department of Computer, Information and Communication Sciences. In the last 10 years, he has been collaborating with industry to advance pen-based human interactions composed of on-line handwriting recognition, pen-based interfaces, and applications, especially educational applications on an interactive electronic whiteboard. He has served on several committees of the Japanese government on industry and university partnerships and those on IT-oriented and IT-supported learning.

Kaoru Matsumoto was born in November 1973. He received his B.Sc. and M.Sc. from the Tokyo University of Agriculture and Technology in 1997 and 1999, respectively. During 1993–2001, he pursued his Ph.D. at the same university in on-line handwritten character recognition, focusing on coarse classification methods and on-line handwritten character pattern databases. At present, he is a researcher at the Research Institute of The SANNO Institute of Management.