E. F. Yhap
E. C. Greanias

An On-Line Chinese Character Recognition System

This paper describes an experimental system for the on-line recognition of handwritten Chinese characters. Constituent shapes of the characters are recognized as they are formed on an electronic tablet. The 72 constituent shapes can define a very large set of Chinese characters. The implementation described recognizes more than 2200 Chinese characters. More symbols can be added with relative ease. Experimental results are given for two writer populations.

Introduction

The written Chinese language has been an imposing challenge to information technologists for many years. Chinese character recognition and automatic translation have received the attention of numerous workers during the past two decades [1]. Although some progress has been made, much more is required to implement practical recognition systems.

This paper describes a system for the on-line recognition of handwritten Chinese words. The writers are presumed to be employed in professional, technical, or administrative work that requires them to occasionally prepare reports, messages, and other text. Since Chinese symbols form an integral part of the Japanese language, the system is applicable both to Chinese symbols and to Kanji, the Chinese symbols in the Japanese language. This automatic recognition function can facilitate the recording, editing, and transmission of text by authors who are not expert in the use of Chinese keyboard devices. These keyboard devices, developed for a practical Chinese vocabulary of 2000 to 3000 symbols, require operators with specialized skills and training. The intended users of our system would not operate such special keyboards often enough to become proficient with them.

Several methods for on-line recognition of Chinese symbols that use number of pen strokes as a primary selection parameter have been described in the literature [2–5]. Since simple stroke counts are not enough for this selection, other parameters must also be used. The resulting recognition systems have required various combinations of stroke count, shape of specific strokes, and relative stroke position to achieve a vocabulary of about 1000 words. The recognition system described in this paper employs a set of 72 constituent shapes to specify and recognize more than 2000 Chinese words. The use of these shapes, called alphabetic elements, makes it possible to extend the vocabulary with relative ease.

Alphabetic elements

Text written in European languages can be read by machines that recognize the relatively small set of alphabetic symbols that make up every conceivable word in these languages. Machine reading of Chinese text requires the recognition of a much larger number of symbols, which correspond to the words themselves. The on-line Chinese recognition system described here decomposes the symbols that represent the words into a much smaller set of alphabetic elements that can be recognized individually. The main consideration in the definition of these alphabetic elements is to find a fixed set of constituent shapes that can be used to describe or synthesize a very large number of common Chinese words. These alphabetic elements should account for all of the shapes that occur in all the Chinese words that make up a practical lexicon. The words handled by the system are represented by strings of these elements, with each string corresponding to a single word. The alphabet must retain element by element

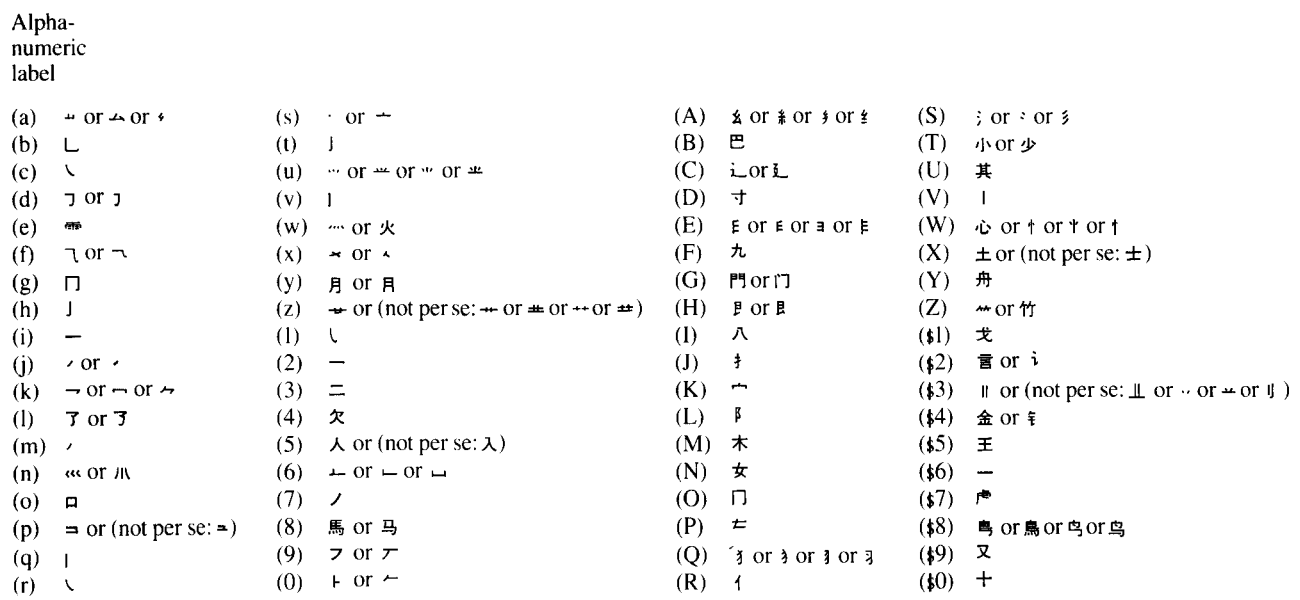

Figure 1 72 alphabetic elements.

correspondence between the shape of the words written on paper and the strings of elements that represent the words in the system. In addition, the strings of elements must provide sufficient discrimination between words with similar shapes to minimize possible conflicts.

In order to keep the tasks of word decomposition and pattern recognition manageable, we found it necessary to restrict the size of the set of alphabetic elements. A limit of fewer than one hundred elements was found to be both adequate and convenient.

Several decomposition approaches are applicable to the problem [6–10]. The alphabetic element set that was selected is similar to the set of 72 elements previously described by Yhap [11]. However, in that prior work the positions of symbol elements had to be specified precisely in order to construct acceptable *output* font displays of Chinese symbols. In this work the positions of alphabetic elements only need to be specified precisely enough to insure that the description of the *input* character is not confused with any other symbol in the vocabulary of the system.

To define the set of alphabetic elements for this system, all of the 214 standard radicals that are used to classify words in Chinese dictionaries were studied with regard to frequency of occurrence in common usage. A radical is a Chinese character or part of a Chinese character which usually indicates some fundamental property of the symbol in which it is imbedded. Some examples are:

石 for stone,

木 for wood,

金 for metal, etc.

We concluded that these 214 radicals were not suitable for a practical system. Some of these radicals occurred in very many words, but others were only found in relatively few uncommon words. In order to keep the alphabetic element set reasonably small, a first set of basic shapes was selected. These included those radicals that occur frequently plus special shapes that can be pieced together to form all of the remaining radicals. The first set of alphabetic elements was applied to approximately 7000 Chinese words and then refined in order to select the final shapes. The 72 alphabetic elements are shown in Fig. 1. It was found that these elements can also be used to represent the Japanese Katakana phonetic symbols. Each symbol has an assigned alphanumeric label as indicated in the figure.

The experimental Chinese/Kanji data entry system

The experimental on-line recognition system was implemented on an IBM System/370 using the Virtual Machine Facility (VM) and the Conversational Monitor System (CMS). The recognition time of the experimental system is compatible with the writing time for the character—one to three seconds. The virtual storage required for the CMS machine is less than 2.5 megabytes. An experimental electronic tablet was used to collect pen position infor-

mation as the users wrote Chinese symbols on the tablet. The pen position signals from the tablet are transmitted to the computer using an IBM 7406 Device Coupler, a 9600-baud start-stop line, and an IBM 3705 Communication Controller into the System/370 channel. The recognition system includes two displays, one for input and one for output, to provide the user with prompt feedback about the writing. The input display plots the sequence of pen positions and creates a facsimile of the symbol as it was written. The output display presents a standardized font representation of the symbols that were recognized. If a symbol is not recognized, the output display presents a list of codes representing the symbol elements that have been detected. These presentations are used to diagnose problems in the logic or in the writer's symbol construction technique.

The basic recognition system is shown in Fig. 2. Its major units include the tablet and a five-part recognition program that processes the tablet pen position information.

- *Electronic tablet*

The electronic tablet was specially constructed for this study. It was designed to give high resolution, reliable pen up/pen down indications and minimal distortion due to changes in the angle between the pen and the tablet surface. This tablet produces a varying pattern of electromagnetic signals, which is picked up by the pen. The relative amplitudes and timings of the signals seen by the pen are used to establish the instantaneous pen positions. The tablet was operated with 200 points/inch spatial resolution and approximately 80 points/second sampling speed.

- *Signal filtering*

The sequence of pen positions is filtered to remove unwanted fluctuations which occur due to the finite spatial resolution and sampling rate of the tablet and to the unsteadiness of hand motions. In order to eliminate extraneous displacements, a threshold is set for the minimum separation between successive pen positions. When a pen position is not separated from the preceding pen position by more than this threshold, it is rejected by the filter. As the pen is placed down on the paper, the first few points tend to show larger extraneous displacements. Therefore, the displacement threshold is made twice as large at the time of pen down and then reduced to its normal level after the pen has moved beyond the initial displacement threshold. This filtering action acts to smooth out expected fluctuations in the data points. Figure 3 contains an illustration of the filter action.

- *Line segments and line directions*

The filtered pen position signals for each stroke (pen down to pen up) are analyzed to identify the individual line segments that make up the stroke. A line segment is a portion of a stroke that does not contain a significant change in direction. In order to identify the end of a line segment, the slope of the displacement from the first point of the segment to a succeeding point (the "average" line direction) is calculated for each point. The slope between adjacent points (the "instantaneous" line direction) is also calculated for each point. As long as the average and

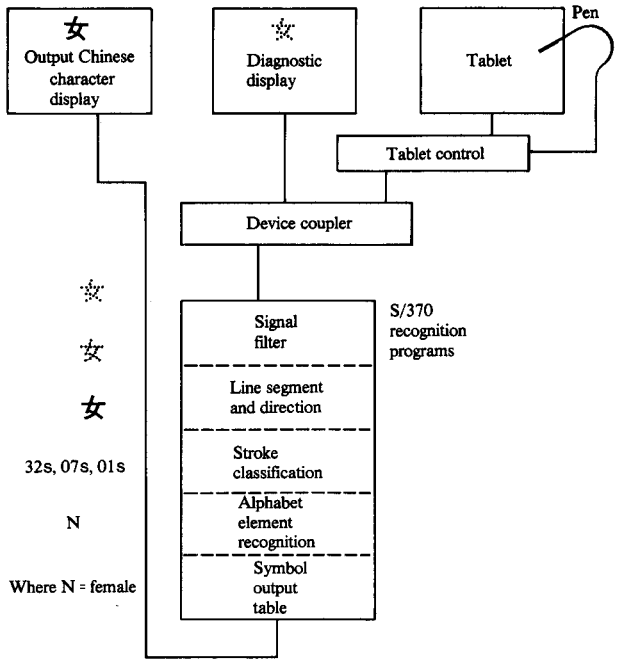

Figure 2 Simplified system diagram. Tablet and noncoded diagnostic displays are connected to the System/370 through the device coupler. The output display receives coded recognition information from the 370. Progressive changes in the signal leading to recognition are shown at left.

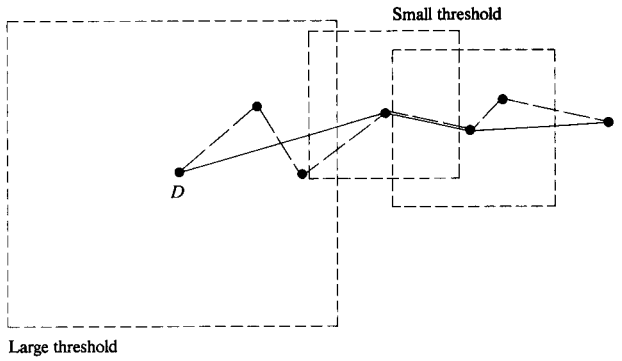

Figure 3 Filter action.

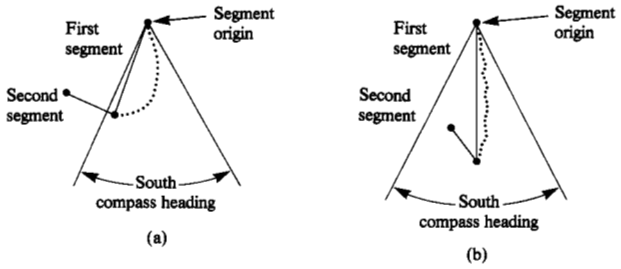

Figure 4 Line segment determination: (a) gradual change; (b) abrupt change.

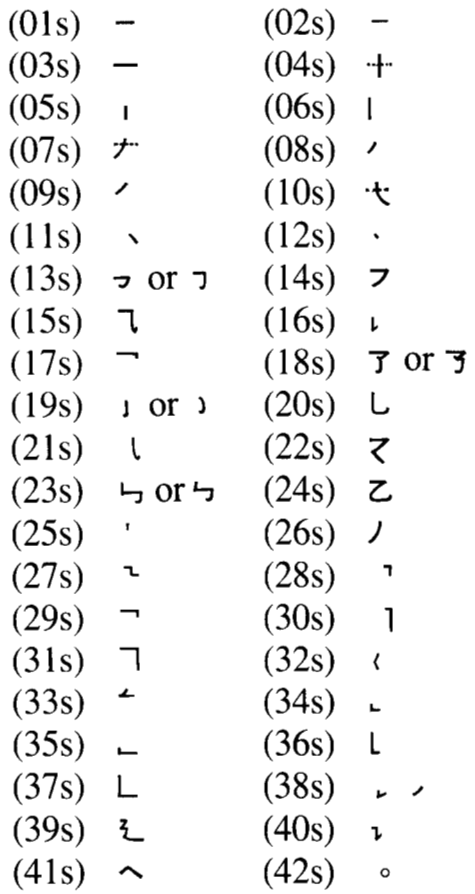

Figure 5 42 allowed strokes.

the instantaneous line directions do not change "compass headings," the line direction is deemed to be unchanged and the segment continues. If the average line direction's compass heading changes (*e.g.*, from south to southwest), the segment ends at the point preceding the change. In selected cases, the segment ends when the instantaneous line direction heading changes without an appreciable change in the average line direction. These cases represent short, abrupt changes in stroke direction that are important for specific classes of symbols. The criteria for these segment terminations depend upon the directions and relative magnitudes of the change and the segment. An example of an important short abrupt change that terminates a line segment is a short northwest move following a generally south move. Figure 4 illustrates this case.

• *Stroke classification*

Pen strokes (pen down to pen up) are classified into 42 categories. These categories, referred to as "allowed strokes," are specifically designed to distinguish among important shape components of Chinese/Kanji symbols. Shape is the most important parameter for the classification of strokes that contain multiple segments with changing directions. Position and intersection are most important for the classification of strokes that are single-segment straight lines. The allowed stroke categories are shown in Fig. 5.

To classify single-segment strokes, their position must be analyzed relative to other strokes. The classification of some single-segment strokes depends upon the positions of the strokes that precede or follow them. The logic for this analysis defines a rectangular boundary, called the subbox, that encloses the preceding portions of the symbol. The position of the stroke with respect to the subbox is a key factor for several of the stroke categories. In other cases, the paths of single-segment strokes are compared with the end points of other strokes to detect the occurrence of intersections. For these cases the intersection information is used to differentiate the stroke categories.

The first 12 strokes in the list of 42 allowed strokes are the single-direction strokes, and their recognition is described below under their four directional headings.

The single-segment strokes labeled 01s, 02s, and 03s are all east strokes. They are differentiated as follows:

01s: Stroke is not close to bottom margin of subbox, *e.g.*,

02s: Stroke is close to bottom margin of subbox but does not extend beyond left-right margins of subbox, *e.g.*,

03s: Stroke is close to bottom margin of subbox and extends beyond left-right margins of subbox, *e.g.*,

or it extends only beyond the left margin of the subbox, *e.g.*,

The single-segment strokes labeled 04s, 05s, and 06s are all south strokes. They are differentiated as follows:

04s: A previous stroke is crossed, *e.g.*, 

05s: A previous stroke is not crossed, and stroke is not close to left margin of the subbox, *e.g.*, 

06s: A previous stroke is not crossed, and stroke is close to left margin of the subbox, *e.g.*, 

The strokes labeled 07s, 08s, and 09s are all one-segment southwest strokes. They are differentiated as follows:

07s: A previous stroke is crossed, *e.g.*, 

08s: No previous stroke is crossed, and the center point of the following stroke is to the right of the center point of this stroke, *e.g.*, 

09s: No previous stroke is crossed, and the center point of the following stroke is not to the right of the center point of this stroke, *e.g.*,  or

The single-segment strokes labeled 10s, 11s, and 12s are all southeast strokes. They are differentiated as follows:

10s: A previous stroke is crossed, *e.g.*, 

11s: No previous stroke is crossed, and it is not a short stroke, *e.g.*, 

12s: No previous stroke is crossed, and it is a short stroke, *e.g.*,

To classify strokes that have changing directions, the successive line segment directions are examined by tree-structured logic that is designed to classify multisegment strokes. If the values of a stroke do not satisfy all the conditions along some path in the tree-logic, the line directions for that stroke are revised by assimilating the smallest segments into their preceding neighbors. The revised set of line segments are then analyzed with the same tree-logic. If the stroke still fails all the logic conditions, the line directions are revised again by assimilating the next smallest directions into their preceding neighbors. If all of the logic conditions are still not satisfied, the stroke is rejected. When a stroke is rejected, the entire symbol is not recognized.

• *Alphabetic element determination*

Alphabetic elements are determined by the occurrence of one or more allowed strokes. Twenty-two of the alphabetic elements are made with a single stroke. These are classified as both allowed strokes and alphabetic elements. Each of the 72 alphabetic elements in Fig. 1 is created by the occurrence of the indicated allowed strokes. Many of these elements also require the allowed strokes to have a particular location with respect to the other portions of the symbol. In some cases a complex element might contain the strokes of a simpler element. The presence of the additional strokes changes the identity from the simpler to the more complex category.

The two-dimensional area that contains a Chinese character is partitioned into nested subboxes, similar to those described in [11] for the construction of output display fonts. However, for this system, it is only necessary to retain enough information to provide unambiguous representations of the symbols.

If we consider each Chinese word to be enclosed by a box, we can subdivide this box into subboxes. For example, the word

can be broken down into

and the right subbox, in turn, can be broken down into

The representation of these subboxes follows the sequence established by the way in which the word is written—left to right and top to bottom. We use a comma (,) and a period (.) to describe the relative positioning of subboxes—the comma between vertically adjacent subboxes and the period between horizontally adjacent subboxes.

If alpha and beta are the representations of the contents of two subboxes, then (alpha).(beta) means that the beta subbox is to the right of the alpha subbox. Similarly, (alpha),(beta) means that the beta subbox is below the alpha subbox. For the word

we thus have

(口).((亠),(口),(子))

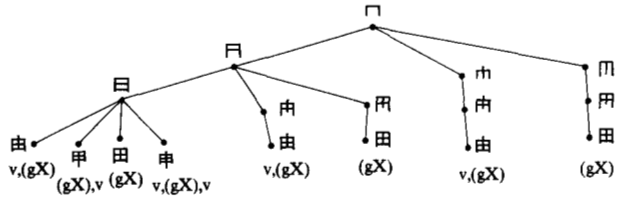

Figure 6 Logic tree example for a shared stroke.

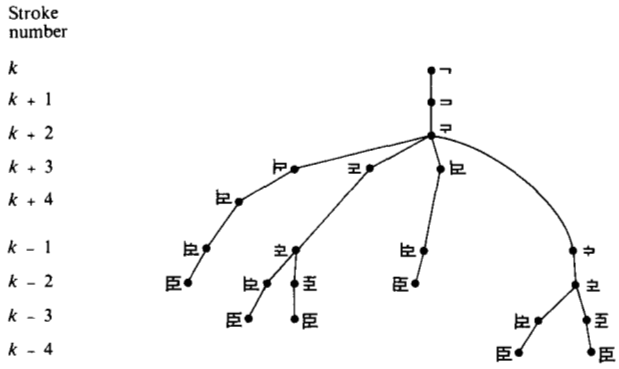

Figure 7 臣 : example of forward-backward logic tree.

There are exceptions to the left-to-right and top-to-bottom order of forming characters. These are handled by the use of empty subboxes, where the symbol (*) denotes an empty subbox. Empty subboxes are used to designate the superposition of symbol elements within a larger portion of the character and to represent the word with an alternative alphabetic element string. For example, using the nomenclature of Fig. 1,

樂

would have a normal spelling of ((A).(joi).(A)),(M). Alternatively, however, we can have the following formation sequence for the first part of the top subbox:

(*).(joi)

[Note that the simplified notation (*.joi) is used below for (*).(joi).]

(*.joi)(A.*),

where (A.*) stands for ((A).(*)) and where the omission of both a comma and a period between (*.joi) and (A.*) means simple superposition of these subboxes on each other.

The rest of the symbol remains the same:

(((*.joi)(A.*)).(A)),(M)

• *Alphabetic element recognition*

Logic trees, which specify all allowed shape variations, implement the alphabetic element recognition. Three kinds of trees are used:

1. Trees that detect the 72 alphabetic elements when they occur with independent strokes. These trees analyze the sequence of allowed strokes, the intersection information, the subbox boundaries, and the location of the allowed strokes with respect to the subbox. The paths through the trees define the conditions for each alphabetic element: the occurrence of the allowed strokes and their relative positions with respect to adjacent strokes. In some cases, there is more than one path through a tree, or more than one tree, for a single alphabetic element. This is one of the ways that alternative stroke writing sequences are accommodated. Of the 72 alphabetic elements the following 16 allow for alternative stroke sequences: w, x, z, G, L, N, Q, X, Y, 0, 8, $1, $5, $7, $8, and $0.

2. Trees that detect alphabetic elements which share strokes with multiple crossings. A simple example of such a tree is given in Fig. 6 for the detection of (v),(gX); (gX),(v); (gX); or (v),(gX),(v).

3. Trees that begin by detecting certain imbedded kernels within an alphabetic element pattern scan forward to detect the last part of the pattern and then backward to detect the first part of the pattern. This kind of tree is especially useful when we have to deal with a pattern which can be written with different stroke sequences or with a different number of strokes due to joining or not joining of certain successive one-segment strokes. An example of this is given in Fig. 7 for the pattern which we shall label as ivpv6.

Besides the logic described by the trees that have been discussed, special conventions have been adopted for the recognition and representation of the following groups of patterns. The representation that is used depends upon whether the patterns occur by themselves or as part of more complex patterns. When they occur by themselves, the alphabetic representations generated for them by the alphabetic recognition logic are:

(土 : X) and (士 : $02)

(四 : g$3) and (皿 : g$3$6)

(人 : 5) and (入 : mr) and (八 : I)

(日 : oi) and (曰 : gi2)

(己 : pb) and (已 : vpb) and (巳 : ob)

But whenever the above patterns form part of a more

complex symbol, they are assigned the following uniform alphabetic representations:

X for both 土 and 士, as in (地 : (X).(dVb)) and
(志 : (X),(W))

g$3 for both 四 and 皿, as in (買 : (g$3),(y9r)) and
(盆 : (ldj),(g$3))

5 for both 人 and 入, as in (卒 : (s),(5.5),($0)) and
(全 : (5),($5))

oi for both 日 and 曰, as in (時 : (oi),(X,D)) and
(冒 : (oi),y2))

pb for 己, 巳 and 巳, as in (紀 : (A).(pb)) and
(祀 : (s9vs).(pb))

The above convention avoids unnecessary alternative spelling entries in the symbol recognition output table described below.

• *Symbol recognition output table*

The outputs of this recognition system are produced in the form of machine codes which represent each of the symbols that have been included in the repertoire, or lexicon, of the system. These codes correspond to the information system codes for input/output devices, and they are stored in the symbol recognition output table. When the recognition logic has completed the analysis of a Chinese/Kanji character, the sequence of symbol elements that were detected is used to search the recognition output table to find the machine code that represents the symbol that was written. In many cases there will be more than one sequence of symbol elements for each machine code. These correspond to acceptable alternative writings of the symbol. If it is desired to relax the constraints for writing a particular symbol, an alternative sequence can be provided to accept the desired writing variation. (This alternative must not be subject to confusion with other symbols.) The alternative sequence is then used as another access to the appropriate machine code.

For describing only a few thousand words we found it convenient not to utilize the parentheses, commas, or periods except in the few cases where they are needed, *e.g.*,

(吧 : o.B) and (邑 : o,B)

Moreover, in such cases we can even drop the comma(s) or dot(s) for the more frequent word, *e.g.*,

(吧 : oB)

Preliminary recognition results

Recognition criteria have been implemented for 2249 Chinese/Kanji symbols and the Japanese Katakana symbols. By using this alphabetic element approach, it was possible to start with a small set of Chinese/Kanji symbols and progressively add new symbols without causing significant conflicts in the previously designed recognition criteria. Two significant tests have been made with this system. The first test covered 920 Chinese/Kanji characters plus the Katakana symbols. These were written by six native Japanese subjects working in the United States. Data were collected and the recognition results analyzed. The following summarizes the results obtained for 5958 written characters:

1. 5426 characters were recognized correctly.
2. There were zero misrecognitions (*i.e.*, zero substitutions).
3. 102 characters were not recognized that should have been recognized. This indicates that improvements are required in the alphabet recognition logic.
4. 19 characters had one or more strokes not recognized that should have been recognized. This indicates the need for improvements in the stroke recognition logic.
5. Of the 411 remaining characters that were not recognized,
   (a) 111 characters had poorly written strokes or poor signal recordings.
   (b) 24 characters had linking of two or more strokes.
   (c) 107 characters had extra strokes, missing strokes, or exotic stroke sequences.
   (d) 97 characters had improperly crossed or uncrossed strokes.
   (e) 72 characters were labeled incorrectly due to experimental error.

Excluding the 411 poorly written or recorded characters listed above under (5), the recognition rate was 97.8%; including them, the recognition rate was 91.1%.

The experimental system was modified to expedite data collection and improve performance. The data-taking protocol was changed to reduce experimental errors and the logic was extended to enhance the tolerance for imperfect writing. An expanded model of the system was sent to Japan for further tests with a 2260-character lexicon. One purpose of these tests was to learn if there are significant differences in the writing characteristics of another population. Different Japanese subjects participated

Table 1 Test results for nine writers.

Subject	*No. of char. written*	*Recog. % at init. writing*	*Recog. % after up to 2 rewritings of rejects*
A	184	59.1	85.1
B	278	84.8	99.3
C	283	89.8	99.3
D	214	83.5	98.6
E	211	81.7	98.6
F	251	78.0	95.7
G	278	81.3	95.4
H	197	81.2	98.9
average of above	237	79.9	96.4
I	349	89.0	99.1

in the testing in Japan. A summary of results for one test with nine writers is presented in Table 1. Each subject was asked to write for a fixed period of time and to make two tries to correct any failures. Since the last writer had prior experience with the system, his results are not included in the average figures.

The corrections were made with guidance from experimenters who made visual observations of the initial writing and then made suggestions for improvements.

It was interesting to compare the data of the two writer populations of the tests above with the original data that were gathered from a Chinese writing population for the initial recognition logic design. We observed some differences in writing styles which may be related to the writer's national origin. (We have not tried any Korean populations.) Differences were noted in some 20 basic subpatterns that occur in many characters. They involve variations in both the shapes of some symbols and the sequence of strokes. These variations often require several alternative alphabetic element sequences for a single Chinese symbol. By gaining a better understanding of these variations, it should be possible to extend the tolerance of the recognition logic to serve a broader writer population. It also has been suggested that an on-line tablet recognition system might be used to teach the Chinese language to school children. Short periods of exercise with a real time interactive drill system might be expected to accelerate the language learning process.

Conclusion

The results that we have obtained, although not conclusive, suggest that this Chinese character recognition method can be developed into a practical method for on-line entry of Chinese text by casual writers. The most significant features of this approach are:

1. It provides means to handle an open-ended vocabulary. New words can be added by simply specifying their allowed "spellings" in a symbol recognition output table.
2. On-line tablet recognition offers considerable promise as a natural data entry device for casual users of information systems. Many writers have shown that they adapt easily to the modest constraints of on-line recognition logic. The provision of computer-mediated feedback to the writer in real time should expedite the achievement of practical recognition rates and greatly reduce the need for human supervision of the training.

Acknowledgment

The authors want to acknowledge the contributions of the following people: G. Purdy, G. Secor, L. Tan, M. Ebihara, P. Greier, J. Kim, and T. Miyazaki for the implementation of the experimental system; H. Dym, J. Seeland, and S. Kambic for the implementation of the tablet device; and Y. Ohi and M. Ebihara for the experimental testing in Japan.

References

1. W. Stallings, "Approaches to Chinese Character Recognition," *Pattern Recognition*, Vol. 8, Pergamon Press, Inc., Elmsford, NY, 1976, pp. 87–98.
2. S. Chang and D. Lo, "An Experimental System for Recognition of Handwritten Chinese Characters," *Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems*, Taipei, Taiwan, 1973, pp. 257–267.
3. K. Ikeda, T. Yamamura, Y. Mitamura, S. Fujiwara, Y. Tominaga, and T. Kiyono, "On-line Recognition of Handwritten Characters Utilizing Positional and Stroke Vector Sequences," *Proceedings of the Fourth International Joint Conference on Pattern Recognition*, Institute of Electrical and Electronics Engineers, New York, 1978.
4. H. Arakawa, K. Odaka, and I. Masuda, "On-line Recognition of Handwritten Characters—Alphanumerics, Hiragana, Katakana, Kanji," *Proceedings of the Fourth International Joint Conference on Pattern Recognition*, Institute of Electrical and Electronics Engineers, New York, 1978.
5. O. Kato, M. Niwa, and N. Fujii, "Real-time Recognition of Handwritten Characters using Stroke Relative Position Matrix," *1979 National Convention Records*, IECEJ no. 1308, pp. 5-297 (in Japanese).
6. S. K. Chang, "An Interactive System for Chinese Character Generation and Retrieval," *IEEE Trans. Systems, Man, Cybernetics* **SMC-3**, 257–263 (1973).
7. N. Chou, "A New Alphameric Code for Chinese Ideographs," *Proceedings of the First International Symposium on Computers and Chinese I/O Devices*, Institute of Mathematics, Academie Sinica, Taipei, Taiwan, 1973.
8. C. Hsieh, Y. Huang, S. Lin, and H. Hsu, "The Chiao-Tung Radical System," *Proceedings of the First International Symposium on Computers and Chinese I/O Devices*, Institute of Mathematics, Academie Sinica, Taipei, Taiwan, 1973.
9. K. L. Hsu, "The Creation of a Set of Alphabets for Chinese Character Written Language," *Proceedings of the First International Symposium on Computers and Chinese I/O Devices*, Institute of Mathematics, Academie Sinica, Taipei, Taiwan, 1973.

10. C. Leban, "Graphemic Synthesis: The Ultimate Solution to the Chinese I/O Problem," *Proceedings of the First International Symposium on Computers and Chinese I/O Devices*, Institute of Mathematics, Academie Sinica, Taipei, Taiwan, 1973.
11. E. F. Yhap, "Keyboard Method for Composing Chinese Characters," *IBM J. Res. Develop.* **19**, 60–70 (1975).

Received March 26, 1980; revised November 25, 1980

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.