ON-LINE COMPUTER CLASSIFICATION OF
HANDPRINTED CHINESE CHARACTERS AS
A TRANSLATION AID

G. F. Groner
J. F. Heafner
T. W. Robinson

April 1967
ABSTRACT

This Paper describes a method for using sequential positional information to recognize handprinted Chinese characters, and a computer program which uses this method to provide a translation aid. The method is based on the fact that Chinese characters are drawn as a definite sequence of a few types of strokes. The program, which provides a direct man-computer communication via a RAND Tablet and a CRT display, simplifies using a Chinese dictionary by reducing the number of steps and the time required.
ON-LINE COMPUTER CLASSIFICATION OF
HANDPRINTED CHINESE CHARACTERS AS
A TRANSLATION AID*

G. F. Groner
J. F. Heafner
T. W. Robinson

The RAND Corporation, Santa Monica, California

Several recent investigations have revealed the advantages of using time sequential information for computer recognition of handprinted letters and numbers [1-5]. However, character variations are introduced that would not arise if only static image information were used. The letter A, for example, may be (and is) written as one, two, or three strokes, with various stroke orders and directions, and yet may look the same when completed. Applied to the recognition of Chinese characters, however, sequential information greatly simplifies the recognition process without introducing any variations.

THE RECOGNITION OF CHINESE CHARACTERS

Although a Chinese character may have a complicated static image, it is drawn with a number of simple strokes. These strokes are always drawn in the same order, following

*Any views expressed in this Paper are those of the authors. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

This Paper was prepared for publication in IEEE Transactions on Electronic Computers.
particular rules [6]. Figure 1b shows the order in which the character in Fig. 1a is to be drawn.

Caldwell [7] has built a Chinese typesetting machine based on the following principles:

- "Chinese has an 'alphabet,' in the sense that all Chinese characters are written by selecting strokes from a relatively small number (21) of basic strokes."
- "Chinese has a 'spelling,' in the sense that sequence of strokes used in writing a given character is invariant. . . . Every Chinese learns to write a character by using exactly the same strokes in the same sequence."

Caldwell's Sinotype selects 2333 Chinese characters based only on a sequence of strokes\(^1\) entered via a keyboard—its "recognizes" in the sense that it encodes a stroke sequence as a single character. It makes use of the complete set of basic strokes so that an operator can reproduce a character pictorially, whereas a smaller set of strokes would probably be sufficient for discrimination among characters.

Only five stroke types ((1) horizontal, (2) short, (3) vertical or downward-left, (4) downward-right, and (5) anything more complicated), as shown in Fig. 2, seem to be required for purposes of computer recognition. These strokes can be accurately and easily identified since they are defined by direction sequences. Strokes can be counted

\(^1\)There are actually 20 pairs and 4 triads of characters which have the same stroke sequence, but this only amounts to 2 percent of Sinotype's characters.
Fig. 1--Construction of a Chinese character:
(a) completed character.
(b) stroke order

Fig. 2--Stroke types: (a) horizontal. (b) short. (c) vertical or downward-left. (d) downward-right. (e) anything complicated
by a computer with no error if the input device sends a
signal upon stroke completion.

A Chinese character may thus be described as a well-
defined sequence of accurately identifiable strokes which
have a particular geometrical arrangement. Although the
initial recognition problem is formidable, each time a
stroke is drawn and identified this problem is reduced by
a factor of approximately 5. Figure 3 (derived from Ref.
8) illustrates, for characters with various numbers of
strokes, how the magnitude of the character recognition
problem is reduced as successive strokes are identified.
This example is a "worst case" situation because the
horizontal stroke accounts for about one third of all
Chinese writing [7].

A more difficult computer recognition problem arises
when characters, such as those in Fig. 4, must be dis-
tinguished by the relative sizes and positions of the
strokes rather than by the types of strokes. This is
similar to the problem of recognizing a Latin letter con-
structed of two vertical strokes and a horizontal stroke--
it is easy to decide that the letter is an A or H, but it
is more difficult to decide which one. The recognition
algorithm at this point must discriminate, yet be tolerant
enough to allow for variations. This type of algorithm is
also more difficult to automate than one which merely
checks for a sequence of stroke types.

Reference 8, the Chinese-English dictionary most
commonly used in this country, contains 7785 characters.
The Chinese language has been estimated to comprise as
many as 50,000 characters.
Fig. 3 -- Frequency of occurrence of characters with the same description
Fig. 4--Different characters with the same stroke sequence
AN AID TO USING A CHINESE DICTIONARY

It is usually a long and arduous task to find Chinese characters in a dictionary because the characters have no natural ordering. In order to demonstrate the feasibility of automating this procedure, we are developing a computer program for cataloging and retrieving related groups of Chinese characters. The program is written in IBM 360 Assembly Language and runs on an IBM 360/Model 40. It makes use of much of the software and techniques developed for the GRAIL Project [5,9]. The input device is a tablet [10]; the output device is a high-performance cathode ray tube (CRT) display.

Conventional Character Lookup

Traditionally, access to Chinese characters has been by means of a "radical" index. Each Chinese character contains at least one "radical," which is a prominent part of the character and which appears as a part of a large number of characters. In the most commonly used Western system [8], the dictionary user inspects the character, picks out the most prominent radical, counts the number of strokes (as many as 17) in the radical, and looks up the radical number (from 1 to 214) in the radical index. He then turns to an adjacent section of the dictionary where all the characters are arranged by radical number. Under

---

3One "natural" ordering is by sound, but since the number of Chinese verbal words is small—about 400—and the number of characters is very large—about 50,000—ordering by this method does not solve the problem. Also, a person using the dictionary is not likely to know the character's sound.
each radical may be listed as many as several hundred characters organized in order of the remaining stroke count exclusive of this radical. The user scans the list of characters under the given radical and remaining stroke count (there may be as many as 50 characters in that list) until he locates the character of interest. Adjacent to this character is a "dictionary number" which indicates the character's location in the body of the dictionary. He can now look up the character.

There are seven steps required to lookup each character (inspection for radical, counting radical stroke number, looking up in radical list, turning to character list, counting remaining strokes, inspecting character list, and turning to the proper page) as compared to one step (turning to the proper page) in the case of an alphabetical language. This procedure may take several minutes. Often one finds that the radical first chosen is the "incorrect" one (characters are listed under only one radical). Often, too, the radical is "obscure"; in this case one can (as a last resort) turn to a section known as "list of characters with obscure radicals," which is ordered by total stroke count. The computer program simplifies this translation procedure.

Using the Program

The Hardware. A user communicates with the computer via a tablet in conjunction with a CRT display as shown in Fig. 5. The tablet hardware consists of a 10-in. square writing surface and a pen-like writing instrument. As the user moves the pen near the tablet surface, a (hardware generated) dot on the CRT follows the pen motion--
this direct feedback helps him position the pen for pointing or drawing. When he presses the pen against the tablet writing surface a switch in the pen closes, notifying the computer that he is beginning a stroke. As he moves the pen across the tablet, the pen's track is displayed (via software) on the CRT--the pen thus seems to have "ink." When the pen is lifted, the pen switch is opened notifying the computer of a stroke completion, and "inking" ceases. A user may "point" at an area on the CRT by closing and opening the pen switch on the corresponding area of the tablet surface.

The Display. A user generates "scrolls" of characters which have the same (computed) description by operating on the display shown in Fig. 6. The description presently consists of the first stroke type (FIRST ID) and the total number of strokes in the character (STROKES), but will be extended to include the first two or three stroke types in order to reduce the number of characters on a scroll. When looking up a character, a scroll containing the character may be retrieved by writing the character on the tablet. A "page" of the scroll, consisting of 21 characters along with their (user entered) dictionary numbers, appears in the lower portion of the display.

Entering Characters. A character to be entered on a scroll is drawn, as "ink," in the ENTER box. The program computes and displays the character's description, retrieves the corresponding scroll, displays the first page of the scroll which has a vacant character space, then moves the "inked" character from the ENTER box to a vacant space. If the character's description is correct, the user then
Fig. 5--A user interacting with the program

Fig. 6--The display
writes its dictionary number and any comments (the character's number in another dictionary, pronunciation, or translation) below the character. If the description is not correct, the user may erase its ink (by a scrubbing motion) and try again. The entire contents of a character space—character ink, dictionary number, and comments—are automatically catalogued whenever another character is entered or retrieved. Characters may be entered in any order regardless of description—it is convenient to enter them in the order they appear in a dictionary.

In Fig. 7, a user has just written a character in the ENTER box, and the program has displayed a page of characters having this description. The user will see the character disappear from the ENTER box, then reappear in line with the three characters previously entered.

**Retrieving Characters.** When a user wants to find a character in a dictionary, he first writes it in the RETRIEVE box. The program then displays the first page of the scroll of characters having this description. This display appears very quickly (after a one-second delay to insure that the character has been completed, plus a few milliseconds to retrieve the corresponding scroll). The user may then visually scan this scroll to locate a likeness of the character he has just drawn together with its dictionary number. When scanning, the user may "flip" pages of the scroll by "pointing" at PAGE. Although all characters of a scroll have the same description, they differ in appearance because of varying stroke shapes and geometrical arrangements—making it easy to find the character of interest. The program reduces the conventional seven steps to a three-step procedure (draw
character, find character with its dictionary number on
the scroll, and look in the dictionary under this number).

In Fig. 8, a user has written a character in the
RETRIEVE box, and the program has displayed the scroll for
FIRST ID=1, STROKES=2. The user observes that his charac-
ter is number 5807 and is pronounced shì. When he looks
in the dictionary [8] under 5807, he finds that the charac-
ter means "ten" or "complete."

THE PROGRAM ORGANIZATION

The control section of the program monitors three
activities: (1) character recognition, which resolves the
input; (2) display managing, which updates the output; and
(3) cataloging and retrieving display data from secondary
storage (an IBM 2311 Disk Storage Drive). Although applied
to Chinese characters, the program provides for cataloging
and retrieving arbitrary graphics.

One of two character recognition routines is invoked
depending on whether the pen is in an area where Chinese
characters are drawn (e.g., in the ENTER or RETRIEVE box)
or where letters and numbers are written (e.g., just above
a dashed line in a character space). One routine [5]
recognizes Latin letters, Arabic numerals, and special
symbols; the other identifies the five Chinese stroke
types. Both make use of a routine which analyzes the
point-by-point pen-location data (provided by the tablet)
as a stroke is being drawn. A stroke is identified within
a few milliseconds after its completion.

The display manager segment presents the appropriate
calendar scroll display upon request. A scroll of
Fig. 7--Entering a character

Fig. 8--Retrieving a character
characters has an internal identification number which is the concatenation of stroke count and first stroke type. The display pages for each scroll are maintained on a disk in their CRT display format.

The directory, which relates disk location to a scroll internal identification, is in the form of a data ring structure (Fig. 9). The ring structure is comprised of elements which contain (1) data and (2) links to similar elements, fashioned in such a way as to form a circular (ring) conglomerate.

The catalogue and retrieve segment of the program creates and edits the ring structure. For a cataloging operation, the display scroll currently in core is recorded on disk and its disk location is recorded in the ring structure. For a retrieve operation, the scroll identification, as determined from the input, is located on the ring structure, and the accompanying disk address is obtained—the display scroll is read into core and presented to the display manager. Upon termination, the ring structure is automatically recorded on disk for later use.

POSSIBLE EXTENSIONS

The computer program may be used as an aid in organizing and preparing a hardcopy dictionary index based upon stroke type and stroke count. Such an index would simplify the lookup procedure without requiring a computer. The program may be used to enter the dictionary numbers for any (or all) available dictionaries along with the characters. The computer displays thus generated could
Fig. 9--Ring structure relating display scroll to its disk location
then be printed in booklet form--e.g., by using a Stromberg-Carlson 4060 to produce microfilm, which, in turn, could be used to generate printing masters.

Other computer programs based on on-line Chinese character classification and graphical information retrieval procedures could be used to set type, to instruct students in the reading and writing of Chinese, or to update special purpose dictionaries.
REFERENCES


