MACHINE PROCESSING OF CHINESE CHARACTERS

William Stallings
Center for Naval Analyses
Arlington, VA

Chinese Characters

Chinese characters, used to encode all the dialects spoken in China as well as the historically unrelated Japanese language, present a unique machine processing and optical character recognition (OCR) problem. Written Chinese is a pictorial and symbolic system which differs markedly from written Western language systems. Chinese characters are not alphabetic; they are of uniform dimension, generally square, and are composed of strokes, each one a line that can be drawn without lifting the pen. In these highly structured characters, many regularities of stroke configuration occur. Quite frequently, a character is simply a two-dimensional arrangement of two or more simpler characters. Nevertheless, because strokes and collections of strokes are combined in many different ways to produce thousands of different character patterns, the system is rich.

Written Chinese is very difficult to learn: there are over 40,000 characters, each corresponding roughly to a word in Western languages, of which an educated person would be expected to know about five to ten thousand. The meaning of each character and its fixed monosyllabic pronunciation must be learned by rote. Usually, these two tasks are eased somewhat because one component of a character gives a clue to its meaning and the rest gives a clue to its pronunciation. But since there is no alphabetic order to Chinese characters, another difficulty is dictionary lookup; a number of special systems have been devised to impose an ordering, none of them terribly convenient. Finally, a student of Chinese must learn to draw the strokes of each character in a particular order; a character may have from one to thirty strokes with eight to twelve being typical.

Of direct relevance to the use of OCR for Chinese is the desire of the Peoples' Republic of China to simplify the written language through a series of language reforms. The first is that the government has recommended the general use of only 2000 characters. Publishers, being government-controlled, are under instructions to stay within the total of 2000 as far as possible. Secondly, the government has simplified a large number of characters with the result that the average number of strokes per character has been reduced by about a factor of two to an average of about 6 to 8 strokes per character. This continuing policy of language simplification will ease the difficulty of Chinese OCR.

Data Processing Requirements

The requirements for machine-processing of Chinese characters, whether for machine translation or other applications, are four:

- input
- storage
- data processing
- output

The requirements in the latter three areas are formidable compared to those imposed by the Latin and Cyrillic alphabets. For example, a machine translation device might be required to print out the Chinese text together with its translation. An adequate representation of each character would require a 32X32 black/white matrix. Hence the storage of the image alone of each of 5000 to 10,000 characters would require 1000 bits. Nevertheless, because of the vast improvements made in memory density and processing speed of computers, these requirements no longer present a problem.

The only remaining bottleneck is input. Because of the many thousands of characters in common use, a keyboard for Chinese (for typesetting, typewriting, keypunching, on-line computer entry, etc.) is an ungainly affair. One common model has 192 keys with 13 shifts, another simply has 2300 keys! Among their disadvantages:

Slow speed - a rate of 40 characters/minute is typical of experienced operators, compared with 70-75 words per minute for an English-language typist. It should be remembered, though, that a Chinese character corresponds roughly to a word in English, so the discrepancy is not so great.

- High error rate - error rates on Chinese typewriters are much higher than Latin letter typewriters - as high as several percent. Considering the high information content of each character, this is a serious problem.

- Training requirement - efficient use of a Chinese keyboard requires a great deal of training and is almost unattainable by those who do not know the language well.

In recent years, a number of approaches to reducing the keyboard complexity, all of which exploit some structural characteristic or the stroke order of Chinese characters, have been taken [1]. It is safe to say that none of these devices has produced an improvement in any of the problem areas listed above.

OCR for Chinese

The only alternative to keyboard entry of Chinese characters is OCR. While there is much optimism about developing satisfactory OCR devices for Latin or Cyrillic letters, the prospect for Chinese OCR is dim. Three problems arise:

- Size - to be useful, a Chinese OCR device would need to be able to recognize 5000-10,000 characters. This is two orders of magnitude greater than the number of images a Latin OCR device would have to handle.

- Complexity - a Chinese character may have as many as 27 strokes. There is so much detail that it is difficult to develop a set of features for distinguishing among characters.

- Density - compounding the complexity problem is the density of printed Chinese characters. On a given document, all characters will occupy the same amount of space, from the simplest to the most complex. The result is that the space occupied by a character is, on average, 50% black. This causes the smudging and overlap of features and strokes, even for the highest-quality printing.

Not much progress has been made in solving these problems, although a number of attempts have been made [2]. The most promising attempt currently underway is at Hitachi Ltd. in Tokyo. A group there has reported an error rate of one in a thousand with a reject rate of one in a million for a set of 1000 characters under rather ideal experimental conditions. It remains to be seen if they can go further with their approach.

Incurring the usual risks associated with such predictions, one might set the following as reasonable goals for a Chinese OCR device given a vigorous short-range development project:

- error rate: 1.5%
- rejection rate: 0.5%
- cost: 5 times the cost of a practical Latin OCR device, whatever that might be.

That these goals can be attained is questionable. If they can, then the choice between OCR and keyboard input of Chinese will be based on a tradeoff between cost, speed, and accuracy.

REFERENCES

[1] Stallings, W., "The Morphology of Chinese Characters: A Survey of Models and Applications", Computers and the Humanities, 9, 1975.

[2] Stallings, W., "Approaches to Chinese Character Recognition" Pattern Recognition, 8, 1976.

WILLIAM STALLINGS

Analyst

Center for Naval Analyses

Arlington, Virginia

William Stallings received a BS desgree in Electrical Engineering from the University of Notre Dame in 1967 and MS and PhD degrees in Computer Science from MIT in 1968 and 1971. His doctoral thesis was on Chinese character recognition.

From 1971 to 1974, he was with the Advanced Systems and Technology Operation of Honeywell Information Systems, Inc., where he worked on the development of interactive computer systems and Chinese character input/output systems. Dr. Stallings is currently on the staff of the Naval Warfare Analysis Group of the Center for Naval Analyses in Arlington, Virginia, where his principal interests are discrete-event simulation, systems analysis, and decision theory.