

CHINESE INPUT SYSTEMS WITH DYNAMIC CHARACTER COMPOSITION

B. T. G. Tan

ABSTRACT

Chinese character input systems may be classified into "Telegraphic" systems, in which a fixed character set is used, and "Calligraphic" systems, in which the character is dynamically composed on input from a set of symbols such as radicals combined with a set of composition rules. Such Calligraphic systems are also capable of dynamically forming new characters. The characteristics and merits of such Calligraphic systems are discussed.

Keywords: Chinese input systems, strokes, radicals, calligraphy

INTRODUCTION

Many systems have been described in the literature which enable Chinese characters to be input into a computer system. The relative merits of these systems have also been extensively compared and assessed [1,2,3]. These systems can be classified into two main groups:

- (a) Phonetic systems, based on the pronunciation of the character.
- (b) Graphic systems, based on the shape of the character. These systems can be further classified into various systems based on strokes, radicals, corners etc.

In practically all the systems described, whether phonetic or graphic, the objective is to uniquely identify a particular character. Each and every one of the characters within the vocabulary of the system is separately stored as a complete graphic shape in the system memory. For example, a system with 5000 characters in its vocabulary would have all 5000 character shapes stored in memory.

In all such systems, the possible vocabulary is frozen by the fixed number of character shapes stored. There is absolutely no possibility for the system to recognize or admit characters whose shapes have not been previously stored. The input problem is then basically a question of using the most practical method of identifying which of the character shapes stored is the required one.

This is a severe restriction when compared to an alphabetic system such as English. Given the 26 letters of the English alphabet, there is no restriction on the number of combinations of the 26 letters which may be used. Thus there is no fixed vocabulary, and neologisms may be created with ease, when required.

Manuscript received on November 2, 1988; revised February 21, 1989. B. T. G. Tan is with the Department of Physics, National University of Singapore, Kent Ridge, Singapore 0511.

Similarly, a human writer of Chinese characters would quite easily be able to form new characters if necessary. Using the rules of stroke formation and radical combination, he would be able to combine strokes and radicals into new characters to express new ideas. While this is not an everyday occurrence, and indeed is discouraged by conservative guardians of language purity, it means that Chinese characters are just as adaptable as alphabetic writing to the demands of an expanding vocabulary. However, very few, if any, of the existing Chinese input systems are able to cope dynamically with new characters in the same way.

TELEGRAPHIC AND CALLIGRAPHIC INPUT SYSTEMS

We would therefore like to propose a new way of classifying character input systems in terms of their ability to dynamically form new characters. We propose that all such systems be classified into two groups:

- (a) The "Telegraphic" group, for systems with a fixed set of characters whose shapes have been stored beforehand. All phonetic input systems, and the vast majority of existing graphic input systems, fall within this category. The name "Telegraphic" is appropriate because the input process is akin to the use of a telegraphic code. In these systems, the desired character is identified from a fixed set of characters by using a string of phonemes or symbols as an identification code. Telegraphic systems are closed systems which do not admit the possibility of characters not defined by the existing character set.
- (b) The "Calligraphic" group, in which the shape of the desired character is not stored in its complete form. The character shape is completely defined by the user input, which specifies the shape in terms of its component parts. These components may be any graphic shapes such as strokes or radicals (we will consider any portion of a character more complex than a stroke to be a radical), depending on how the characters are built up by the system. Each stroke or radical is input together with a composition rule, which defines its spatial relationship with respect to the previously entered strokes or radicals. In this way, the complete character is built up from a set of pre-stored strokes or radicals in a manner similar to the writing of a calligrapher. Such systems are open-ended, in that new characters may be formed dynamically by the user at the time of input.

Such Calligraphic systems should not be confused with Telegraphic input systems which use strokes or radicals to identify the desired character. In such Telegraphic systems, the stroke or radical input may not be complete and is only used to identify the character from a fixed set of characters. Furthermore, the strokes or radicals are usually not accompanied by a corresponding composition rule. Hence it would not be possible to deduce the complete shape of the character from the strokes or radicals being input.

We now propose a formal definition to distinguish between the Telegraphic and Calligraphic groups. Let N_V be the size of the vocabulary i.e. the total number of characters which the system can handle, and n_s the number of separate and distinct graphic shapes stored in the system, which we will call symbols.

Then we have, for Telegraphic systems

$$N_V \leq n_s$$

For the majority of Telegraphic systems, there is exactly one symbol stored per character in the systems vocabulary i.e.

$$N_V = n_s$$

For Telegraphic systems which have characters with more than one graphic representation e.g. for different fonts or sizes, we have

$$N_V < n_s$$

For Calligraphic systems, we have

$$N_V > n_s$$

i.e. there are always fewer symbols stored than the total number of characters which the system can handle. The extreme case of a Calligraphic system is the Pure Calligraphic system in which $n_s = 1$, i.e. only a single symbol is stored, such as a pixel, from which all the characters have to be constructed. In such a system, the entire character shape has to be input manually as a freehand drawing, as for the actual writing of a character on paper.

For any Calligraphic system, since $N_V > n_s$, the n_V symbols have to be combined in various different ways to give the N_V characters in the required vocabulary. We thus require a number of composition rules which define the positioning of each symbol with respect to the previously input symbols. Each symbol being input must thus be accompanied by a composition rule.

We define n_r to be the number of different composition rules used by the system and m the maximum number of symbols that may be used in a character. Then assuming that there is no restriction on the repetition of symbols or composition rules, there will be $n_s n_r$ different combinations of one symbol with one composition rule. N_T , the total number of different characters (whether valid Chinese characters or not) which can be generated is thus given by

$$N_T = (n_s n_r)^m$$


Thus we see that n_r is inversely proportional to n_s for fixed N_T and m . We would of course prefer to make n_s , n_r and m as small as possible, but obviously there are tradeoffs between these 3 parameters. It is also easy to show that for any fixed m , the sum of the symbols and composition rules, $n_s + n_r$, which should be made as small as possible, is at a minimum when $n_s = n_r$. It can also be seen that if $N_V < N_T$, as is most likely, then new characters can be readily admitted to the vocabulary.

In the extreme case cited above where $n_s = 1$, if the single symbol is a pixel in a square matrix, then both n_r and m are directly proportional to the resolution of the matrix. This would permit the representation of any graphical figure, including Chinese characters, in the matrix. A realistic estimation of the total number of possible characters may be obtained by assuming that not more than half the matrix elements will be filled with pixels i.e. $m = 242/2$. The composition rule accompanying each pixel is merely to select for that pixel a matrix element from those not already filled by pixels, so that n_r is not constant but decreases by

one for each successive symbol chosen. Thus we have

$$N_T = \frac{(24^2)!}{(24^2 / 2)!}$$

The closest approximation to manual Chinese writing would be to have a stroke-based method, with n equal to the number of different types of strokes. For a small repertoire of strokes, we will have $n_r > n_s$ and $m > n_s$, with m being equal to the number of strokes in the character, which can be in excess of 30 if we consider the most complex characters.

If we select radicals as symbols, n_s becomes larger than for the stroke-based case. We then have $n_r < n_s$ and $m < n_s$. For radicals only some composition rules will be applicable to each symbol. Indeed, there will be symbols such as  for which there is only one possible composition rule.

IMPLEMENTATION OF CALLIGRAPHIC SYSTEMS

The Calligraphic methods thus dynamically synthesise the final character shape dynamically from the the input. Needless to say, the complex form of the Chinese character makes this a non-trivial task. In English, for example, only 26 basic graphic shapes are employed, which are combined with each other in only one way, which is serially or in a linear sequence.

There has been much work on the analysis of Chinese characters and their structure [4,5,6]. Such analysis is usually aimed at showing that all the characters can be built up from a smaller subset of symbols. The symbols used most often are the basic strokes and radicals. Once we have chosen a suitable subset of symbols for a Calligraphic method of input, a corresponding set of composition rules is required. As shown earlier, there is a tradeoff between the two parameters; for example, one can choose to have a very small number of symbols such as the most basic strokes, but then a complex set of composition rules will be needed for combining them. On the other hand, if one seeks to have simple composition rules, then the symbols will have to be more complex such as radicals.

For a practical Calligraphic system, the choice between a stroke or a radical-based system depends on the trade-off between user complexity and flexibility in the generation of new characters. A stroke-based system would obviously enable great flexibility in the creation of new characters (A pixel-based system would be even better from this point of view, but we rule it out as being totally impractical for a computer input system). However, as characters may have more than 30 strokes, a stroke-based system would not as complex as manual calligraphy.

A radical-based system would thus appear to be more practical, since the number of radicals in a character is much smaller than the number of strokes. In this context, radical refers not just to the classically determined 214 Kang-Xi radicals, but any portion of a character. Radical-based systems, however, restrict the creation of new characters to the use of the existing radicals and composition rules, and are thus less flexible in this respect than stroke-based systems. On the other hand, these restrictions may be an advantage, as they would also ensure that any new characters would look conventional and thus be more acceptable.

An important study of Chinese character structure by Suen and Huang [4] gives one possible schema for the generation of characters from radicals and composition rules. They show that it is possible to able to generate 4,864

characters (N_V as defined above) using 687 radicals and 15 composition rules. They also limit the maximum number of radicals in any character to not more than 4. We thus have $n_s = 687$, $n_r = 15$ and $m = 4$. The resultant value of N_T is very much larger than the 4,864 characters recognized as valid, leaving much scope for the creation of new characters. Their study thus gives a possible basis for the design of a Calligraphic input system.

A Calligraphic system using the schema described in Suen and Huang's study would therefore have 687 radicals and 15 composition rules. This would require a keyboard with $687 + 15$ keys, one for each radical and composition rule. To input a character, the user would have to input each of the radicals constituting the character and its corresponding composition rule in sequence, up to not more than four radicals. For example, the character 個 is formed by the following sequence of radicals and compositional rules as defined by Suen and Huang:

1. The radical 亻 and compositional rule 2 to give 亻
2. The radical 口 and compositional rule 7 to give 亻口
3. The radical 古 and compositional rule 15 (end) to give 個

It should be noted that the compositional rule accompanying each radical as given above specifies the position of the next radical with respect to the current radical. The final radical is always accompanied by compositional rule 15 which specifies the end of the operation. An alternative convention would be for the accompanying compositional rule to specify the position of the current radical with respect to the previous radical. In this case, the first character would not need a compositional rule but may instead be accompanied by a rule to indicate the start of the operation. The above sequence would then become

1. The radical 亻 and the "start" rule to give 亻
2. The radical 口 and compositional rule 2 to give 亻口
3. The radical 古 and compositional rule 7 to give 個

A new character, not in the intended set of 4,864 (and probably never used before in Chinese) might perhaps be generated, for example, by adding a 4th radical and composition rule to the above sequence as follows (using the second convention for adding radicals defined above) :

4. The radical 辶 and composition rule 8 to give 迴

This shows that new characters or words may easily be generated by such a system.

The question of coding for storage also has to be considered for a Calligraphic system. In a Telegraphic system, each stored character usually has a unique code assigned to it for identification. This code is useful for the efficient storage and transmission of the Chinese text. In a purely Calligraphic system, there will be no such pre-defined code since the words are built up dynamically. To store each word as it is built up as part of a text, the sequence of radicals and corresponding composition rules for each word will have to be stored, constituting a dynamically created code which in itself contains all the information required for the formation of the character graphic. This is identical in principle to the storage, say, of English text, where each word is stored as the sequence of letters constituting the

word, and not as some pre-defined code.

Thus in a Telegraphic system, we can meaningfully say that system is aware that a character being input is one which it "knows" as being an established character which is part of its pre-defined vocabulary. In a Calligraphic system, however, the question of whether the system is aware that it has generated a new character need not arise at all, since as far as it is concerned all characters which it can generate are valid. However, there is no reason why we cannot pre-store the input sequence codes of a given set of characters which is defined as constituting an allowed vocabulary, so that any character with an input sequence not belonging to this set is recognised as a new character.

The large number of radicals needed for a Calligraphic system may be mitigated by using Telegraphic methods for part of the system. To keep the overall system Calligraphic in nature, we may use a Telegraphic method to identify each one of a fixed set of radicals, and then use the chosen radical with a composition rule. If we apply this to Suen and Huang's system, we need a Telegraphic method of selecting one out of their set of 687 radicals. Suen and Huang have actually assigned a two-letter code to each of their 687 radicals, so that a normal alphanumeric keyboard may be used instead of having one key for each of the 687 radicals. However, the burden now rests on the operator to remember the two-letter codes for the radicals. We may simplify the input process even further by, for example, using a stroke-based method to identify the required radical, or using an even smaller set of simpler radicals to build up the set of 687 radicals.

Some indication of how this might be done is given by the Kaihin Input Method (KIM) of Chiu and Wong [7]. They use the 214 Kang-Xi radicals as their basis, and select 36 of these radicals to form a smaller basic radical set. The other more complex radicals are then built up from these 36 — this being a Telegraphic method as the set of 214 radicals is a closed set. In the KIM system, the radicals chosen by the user are then used, in a second stage Telegraphic process, to identify the required character from an existing fixed character set whose shapes are pre-stored. The entire system is thus Telegraphic in nature.

If such a simple Telegraphic method can be used to identify radicals from a fixed radical set, these identified radicals could then be combined, using a set of composition rules, to build up a character by a Calligraphic process. Only the radical shapes would be stored, and new characters would be formed by new combinations of the radicals which had not been previously used. The overall system would be Calligraphic in nature.

PRACTICALITY OF CALLIGRAPHIC SYSTEMS

While such a Calligraphic system might eventually be made workable, particularly with the aid of artificial intelligence techniques to simplify the composition rules, it is most unlikely that any Calligraphic system could be made simpler than a Telegraphic system. Practical input systems would thus most likely remain Telegraphic in nature, with fixed character sets not allowing dynamic composition of new characters. (Of course, a fixed character set can be enlarged by adding new characters, but this is not Calligraphic as defined earlier, since it is not done dynamically.) Apart from the system of Suen and Huang [4], there is at least one other practical input system apparently of a Calligraphic nature reported in the literature by Hwang [8].

If Calligraphic systems are too cumbersome for computer input, they do form the basis of many practical character generation systems used for computer graphic output [9,10,11,12,13,14]. In such output systems, the composition of the desired character is not dependent on human intervention and thus input keying

complexities do not arise. Such systems are useful as they obviate the necessity for the output devices, such as printers and plotters, to store a complete character set for graphic output. Only the radical shapes and composition rules need to be stored. The existence of such Calligraphic character generation systems shows that Calligraphic input systems are feasible, if not totally practical. They may even be incorporated into input systems which are basically Telegraphic to generate characters.

For example, Heinzl et al [9] have described an input system which is Telegraphic in nature in that a phonetic system is used to identify one of 6,764 pre-defined characters. However, the generation of the characters is Calligraphic in principle, with the 6,764 characters being generated from 512 symbols. The composition rules are highly complex as each of the constituent symbols are positioned and dimensioned by treating them as vectors whose beginnings and endings are positioned on a 16 by 16 matrix. Thus while the character generation part of their system can generate more than 6,764 characters, their overall system is essentially Telegraphic as the phonetic input is restricted to the fixed set of 6,764 characters.

While we have considered the Chinese language in terms of individual characters, it could of course be argued that Chinese words are seldom monosyllabic but are more usually bisyllabic or trisyllabic using combinations of two or three characters. From this viewpoint, new words can be formed by a Telegraphic system, since we can coin new bisyllabic words, for example, by new combinations of the existing character set. In this paper, however, our emphasis has been on the written language and thus on the individual character, as it is the primary unit of the written language.

What then is the value of Calligraphic systems? At a philosophical level, they counter the often heard assessment of Chinese characters as being inferior to alphabetic systems, which claims that they are inflexible and not able to adapt to the introduction of neologisms. The possibility that Calligraphic systems of character input can be constructed demolishes that argument, by showing that the Chinese system of writing can also dynamically adapt to change, and that there is no inherent inferiority in Chinese writing compared to alphabetic script in that respect.

While Calligraphic input systems may never ever surpass Telegraphic systems for practical computer input, it may be useful to incorporate a Calligraphic capability in a Telegraphic input system for the composition of new characters not included in the fixed character set. For example, the commercially available Kaihin Brushwriter system [15] incorporating the input system of Chiu and Wong [8] includes a graphic composition facility for drawing new characters. This is actually a Calligraphic system with $n_s = 1$, as explained earlier. Such a drawing facility, while useful, is limited as it presupposes that the user has calligraphic skills, which may not be the case. It would be much more useful if the new character shape could be constructed using, say, a radical-based Calligraphic input system. For the Kaihin Brushwriter system, it appears that such a Calligraphic input method could use the radical selection scheme already incorporated into their system.

In fact, a sophisticated and refined Calligraphic input system could be used for computerised calligraphy, making this ancient and aesthetically highly satisfying art accessible to a much wider public. Such a Calligraphic system could be stroke-based, to allow for the widest flexibility, or radical-based for simplicity of use. Sophisticated features could be included in such a system which would allow for subtle variations in stroke shape depending on brush technique and stroke position. In addition aesthetic rules, possibly using artificial intelligence techniques, could be used to optimise the compositional balance and relative proportions of the character in accordance with the accepted rules of calligraphic aesthetics as practiced by the

great masters of calligraphy. Advanced systems of computerized calligraphy could even be designed to generate characters in the various styles of calligraphy such as lishu, kaishu, xingshu and even caoshu.

It is hoped that the ideas in this paper, while they may never be of great practical importance in Chinese language input systems, will have contributed a new angle on the assessment of input systems, by showing that Chinese characters are also amenable to change and adaptation in response to the introduction of neologisms. We also hope that this paper has drawn attention to the calligraphic aspects of the computerization of Chinese characters and given some food for thought on the relationship between Chinese writing, calligraphy and the computerization of Chinese characters.

REFERENCES

1. Cheng-Kuang Chen and Reng-Weng Gong, "Evaluation of Chinese Input Methods," Computer Processing of Chinese and Oriental Languages, Vol. 1, No. 4, 236-247, November 1984.
2. Jack Kai-tung Huang, "The Input and Output of Chinese and Japanese Characters," Computer Magazine, Special Issue on Chinese Computing Systems, Vol. 18, No. 1, 18-24, January 1985.
3. S. Y. Lo, "A Scientific Model for Comparing Various Methods of Inputting Chinese Characters into Computer," Computer Processing of Chinese and Oriental Languages, Vol. 2, No. 1, 36-58, May 1985.
4. C. Y. Suen and E.-M. Huang, "Computational Analysis of the Structural Compositions of Frequently Used Chinese Characters," Computer Processing of Chinese and Oriental Languages, Vol. 1, No. 3, 163-176, May 1984.
5. Jiben Yang, "A Psychological View on the Standardization of the Structural Elements of Chinese Characters in Information Encoding," Computer Processing of Chinese and Oriental Languages, Vol. 2, No. 1, 59-70, May 1985.
6. P. Zhang, J. Y. Suen, C. C. Deng and N. P. Chao, "Analysis of Chinese Character Component and Structure Parameters," Proceedings of the 1986 International Conference on Chinese Computing, Singapore, 196-202, August 20-22, 1986.
7. A. Chiu and F. Wong, "An Intelligent, Knowledge-Based Chinese Input System," Computer Processing of Chinese and Oriental Languages, Vol. 3, No. 1, 25-32, May 1987.
8. Yeong-wen Hwang, "A Very Compact Chinese System Using Component Method and Character Generator," Proceedings of the 1985 International Conference on Chinese Computing, San Francisco, J-2.1-J-2.11, February 1985.
9. Joachim L. Heinzl, Guo Jun Ji and Shuying Zhang, "Decentralized Terminal for Input/Output in Hanyupinyin," Computer Processing of Chinese and Oriental Languages, Vol. 2, No. 2, 85-100, October 1985.
10. K. Y. Cheng and K. J. Chen, "A System Model for Chinese Character Fonts Generation," Proceedings of International Conference on Text Processing with Very Large Character Set, Tokyo, Japan, 53-61, 1983.
11. C. C. Hsieh, C. T. Chang and P. H. Chen, "Design of a Multi-font Chinese Character Generator for Personal Computers," Proceedings of the 1985 International Conference on Chinese Computing, San Francisco, F-2.1-F.2.10, February 1985.
12. Cornel K. Pokorny, "A Low-Redundancy Electronic Generator for Chinese Characters," Proceedings of the 1985 International Conference on Chinese Computing, San Francisco, D-3.1-D-3.14, February 1985.

13. K. J. Chen and K. Y. Cheng, "A Model for the Low Cost Chinese Character Generator," Proceedings of the 1985 International Conference on Chinese Computing, San Francisco, F-1.1-F-1.10, February 1985.
14. S. J. Huang, Z. L. Wu, J. S. Yu and X. Liang, "A Preliminary Study of the Theory and Application of the Asterisk Style Writing of Chinese Characters," Proceedings of the 1986 International Conference on Chinese Computing, Singapore, 417-420, August 20-22, 1986.
15. Kaihin Brushwriter Technical Brochure, Kaihin Technology Pte. Ltd, Singapore.