# A Description of Chinese Characters Using Sub-patterns

TOSHIYUKI SAKAI*, MAKOTO NAGAO* AND HIDEKAZU TERAI*

*Abstract*

A method of synthesis of Chinese characters is presented in this paper. Chinese characters have too complicated structures to be described by a simple algorithm. We have extracted about 250 basic components of Chinese characters. From these components and with 10 synthesis operators we have succeeded in writing 2,000 Chinese characters on the XY-plotter controlled by a digital computer. In this system we have subprograms such as enlargement, vertical/horizontal writing control.

## 1. *Introduction*

Information processing techniques by digital computer have been much developed, and various problems can be solved accurately and quickly in the computer. There are, however, many problems to be solved. One of them is to develop an easy method to feed data into computer. Researches on character recognition are now in progress for printed letters or special font of alphabet and several symbols. As for KANJI or Chinese letter, it is very difficult to transform them automatically into machine readable form, because the number of them are too many, i. e. more than 2,000, and their structures are too complex. Conventional method of KANJI coding used in Japan is to use KANJI encoder which has about 2,000 keys and several function keys. The speed of typing is much slower than alphabet typewriter.

In this paper we described a method of encoding and displaying KANJI with the reduced number of keys on typewriter; KANJI is decomposed into sub-patterns or elements, and operators combine them into complete KANJI.

## 2. *Description of KANJI*

A KANJI can be looked on as a geometrical figure which is characterized by a special arrangement of some sub-patterns or elements. It is enough to select such a set of operators shown in Table 1 in order to describe the structure of any KANJI. As for sub-patterns or elements, about 250 units illustrated in Table 2 are picked up from the experimental analysis of usual KANJI. These

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 5 (1969), pp. 285-293.

* Department of Electrical Engineering, Kyoto University.

Table 1. Operators for the composition of chinese characters.

| operator | meaning | example | operator | meaning | example |
|---|---|---|---|---|---|
| H | left · right ○ ○ | 休 | S | surroundings | 困 |
| V | up · down ○/○ | 電 | K | KANMURI | 字 |
| P | left · right ○○ | 非 | L | SINNYO | 進 |
| Q | up · down ○/○ | 否 | J | TASUKI | 司 |
| I | penetration | 中, 丹 | G | TARE | 歴 |

sub-patterns do not necessarily coincide with the conventional classification of HEN, TUKURI, KANMURI, etc., but in some cases more detailed sub-division is tried.

As stated in Table 1, an operator II means that it arranges two operands holizontally in separated position, and V arranges them vertically. When two operands must be arranged connectedly, operator P or Q is used according to horizontal form or vertical form. Functions of other operators will be easily understood from the example shown in Table 1.

An example of structural description of KANJI or parsed form of KANJI is shown below;

/算/……V (H (𥫱, 𥫱), 目, I (H (丿, 丨), 一))

In this case five elements 𥫱, 目, 丿, 丨, and 一 are used. It may happen that there are several different parsed forms for the same KANJI, for example H (V (立, Q (口, 儿)), V (立, Q (口, 儿)) and V (H (立, 立), H (Q (口, 儿), Q (口, 儿)) for 競 but this does not prevent the unique coding and displaying of that KANJI if an appropriate dictionary or a parsed form translator is used. Other examples of parsed form are shown in Fig. 1.

密　V (K (宀, I (心, 丿)), 山)
好　H (女, 子)
奪　V (Q (一, K (𠂉, 隹)), 寸)
痴　G (疒, H (矢, 口))
致　H (Q (一, 厶, 士), 夂)

Fig. 1. Examples of the coding of Chinese character.

3. *Domain Partition*

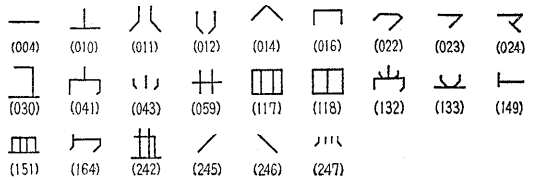
If an structural description form or a parsed form is given, each sub-pattern or element is allocated their own domain according to the specification of operator and the class to which the element belongs (the weighting of partition is shown in Table 3). Domain partition is shown in Fig. 2 for each operator.

Table 2. List of Sub-patterns.

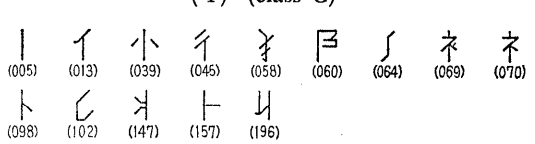
(1) Example of Sub-pattern

| No. | SUB Pattern | No. | SUB Pattern | No. | SUB Pattern |
|---|---|---|---|---|---|
| 1 | ノ | 34 | 又 | 67 | ヒ |
| 2 | 𠂆[B] | 35 | 女 | 68 | 氏 |
| 3 | 𠂆[C] | 36 | 子 | 69 | ネ |
| 4 | 一 | 37 | 𠂇 | 70 | ネ |
| 5 | 丨 | 38 | 土 | 71 | 示 |
| 6 | ノ | 39 | 艹 | 72 | 疒 |
| 7 | 丶 | 40 | 寸 | 73 | 癶 |
| 8 | 乙 | 41 | 宀 | 74 | 祭 |
| 9 | 乚 | 42 | 小 | 75 | 士 |
| 10 | 亠 | 43 | 灬 | 76 | 矢 |
| 11 | 八 | 44 | 尸 | 77 | 木 |

(3) (class B)




(4) (class C)




(2) (class A)

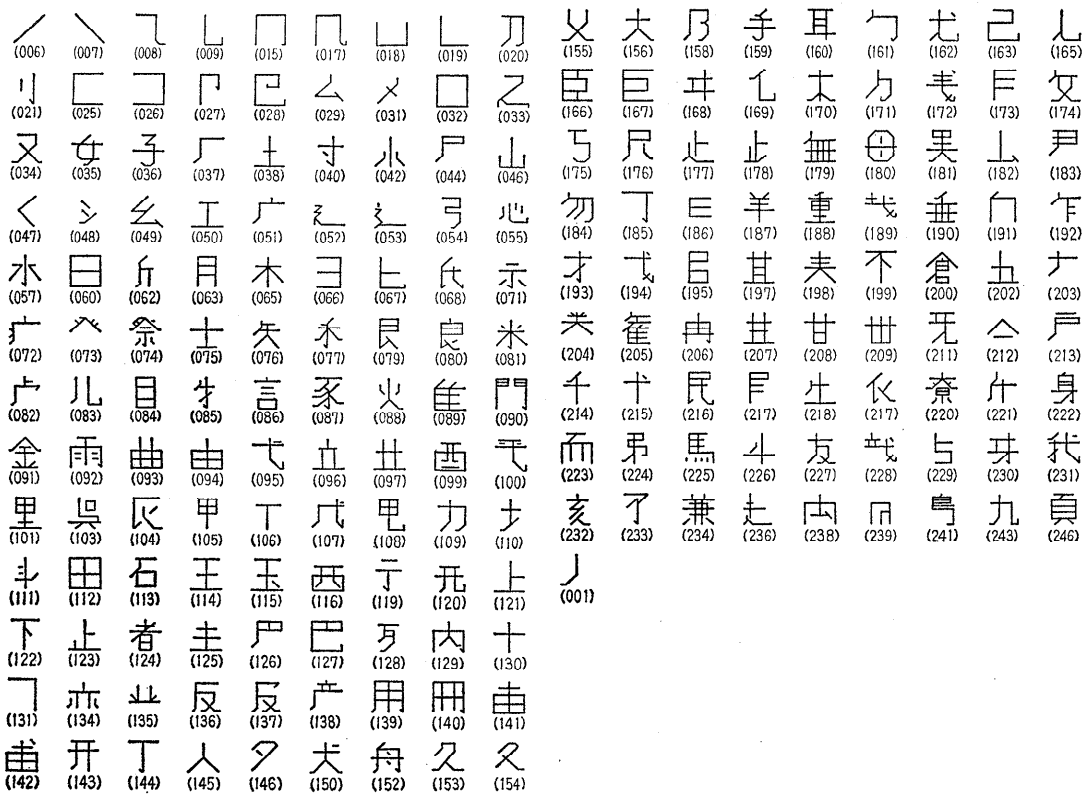


Table 3. Partition weighting.

| Operator | operands | weight |
|---|---|---|
| H | L, G, J, P | 4 |
| | class A, class B | 2 |
| | class C | 1 |
| V | K, Q | 4 |
| | class A, class B | 2 |
| | class B | 1 |
| P | L, G, J, H | 4 |
| | class A, class B | 2 |
| | class C (「\|」) | 1 (0) |
| Q | K, V | 4 |
| | class A, class C | 2 |
| | class B (「—」) | 1 (0) |

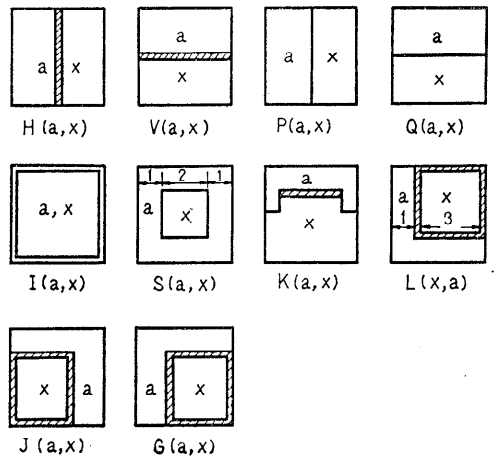



Fig. 2. Operators of the domain partition.

Hatched zones indicate the dead zone where no line element is allowed to enter. To express the domain partition for the operator of penetration- I is difficult, so that it is realized by overwriting the penetration line on a figure. When the domain for each element described in parsed form are settled, the size of each element is deformed to be parked into that area, i.e. reduction, enlargement or and distortion are applied.

## 4. *Algorithm of Synthesis of KANJI*

When a parsed form is fed into computer, the syntactic error is checked, then structural analysis or leveling is carried out. According to this level of each operator and element, domain allocation is executed. As for the case of 踈, for example, structural analysis of parsed form becomes like this;

H(Q(一, 止), I(V(一, 口, H(／, ＼)), |))
0 1 2 2 1 2 3 3 3 4 4 2

First, the whole area for one KANJI (20 mm×20 mm) is divided into two equal parts A and B by the operator H, which is level zero. Then the part A is sub-divided by the level 1 operator V into two parts A1 and A2. The difference of size in A1 and A2 depends on the class of elements which the operator V governs. The same way it proceeds for sub-area B. Final domain partition is shown in Fig. 3.

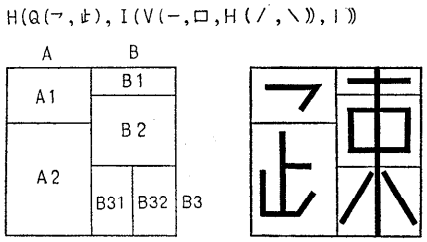



Fig. 3. Domain Partition of「疎」.

## 5. *Experiment*

Experiment was carried out using a medium size general purpose computer: NEAC 2200/200, core 16k bytes, access time 2$\mu$s, drum as a back-up store. As an output device an on-line XY-plotter is used, whose pen moves to one of eight direction with 0.2 mm pitch. Program is written in assembler language. The total number of instructions is about 2000. Average processing time for one KANJI is 1.5 sec, this slowness depends on the low speed of the XY-plotter.

## 6. *Discussion*

There are several information storage methods for KANJI data, such as dot patterns, stroke sequences, and our subpatterns. The memory size required for our method is about one-fourth or one-third of other methods. The characteristic points of our method are (1) flexibility in size and proportion of output KANJI, (2) small number of keys i.e. 250 subpatterns and 10 operators, (3) pronounciation-free coding because parsed form is obtained from visual structure of KANJI. On the other hand, for some KANJI's it is difficult to describe their structures by the operators. Sometimes, unnatural form or unbalanced form appears. This description method may be applicable to Hangul letters.

### *References*

[1] Sakai, T., M. Nagao and H. Terai, Description of KANJI by sub-patterns, *Report for Information Theory Group of IECEJ*, Jan. (1969).

[2] Terai, H., Synthesis of KANJI by digital computer, Master Thesis, Kyoto University, March (1969).