THE DESIGN AND APPLICATION OF THE CHINESE CHARACTER CODE FOR INFORMATION INTERCHANGE —(CCCII)—†

*Ching-chun Hsieh**
*Jack Kai-tung Huang***
*Chung-tao Chang****
*Chen-chau Yang*****

1. Introduction

The task of designning the Chinese Character Code for Information Interchange (abbr. CCCII, afterward) has two major efforts; first, to construct its code structure and second, to organize Chinese characters for the code. The volume I of CCCII,[1] published in April, 1980, has demonstrated its major characteristics of CCCII, but the characters involved are limited to the 4807 most frequently used ones.[2] Therefore, volume I is good only for certain popular applications that envolve limited characters.

The work done for the volume I of CCCII is just a part of the whole. The ultimate goal of CCCII is aimed at the interchange

---

† Paper presented at the International Workshop on Chinese Library Automation, Taipei, Feb. 14-19, 1981.

* Professor, Dept. of Electronic Engineering, National Taiwan Institute of Technology.
** Professor, Dept. of Computer Science, Ming Chuan College.
*** Professor, Dept. of Electronic Engineering, National Taiwan Institute of Technology.
**** Director, Software Center, Institute for Information Industry.

of Chinese information of any sort. This theme, we believe, is the most foundamental requirement for designing an interchange code for any languages. Otherwise, the code will not be satisfied as a general one, and it will be biased to certain area of applications. As a consequence, it may lead to be incapable of handling certain applications and even be founded impossible to be extended to cover a wider area of application if its code structure were not properly designed originally. The approach we used to design the CCCII will overcome the above drawback that has been appeared in many Chinese character codes used in commercial computer systems and even in certain standards.

In order to achieve the goal mentioned above, the complete CCCII is intended to cover all existing Chinese characters which are estimated to have more than 80,000. Since the frequencies of usage of those characters are very much unevenly distributed, we group them according to the frequencies of their appearence. In other words, user may select a suitable subset of the CCCII for its own application to elimate the inconvenience introduced by the huge volumn of characters.

For more characters, more coding rooms are required. This leads to the argument of coding efficiency. The CCCII is designed strictly following the standards of ISO646 and ISO-2022 in order to achieve international compatibility for information sharing. Therefore, basically, it requires three 7-bit bytes to represent a character. One may argue that this coding scheme may lead to certain degree of inefficiency. In this paper, it will shown that by applying similar techniques used in ISO2022, two 7-bit bytes for each Chinese character will do as well if certain extension control sequences can be arranged and the CCCII will still completely remain in a 7-bit environment defined in ISO646.

In this paper, the volume II of CCCII will be introduced. The character set of volume II is called the character set for general data processing. This characters set has collected all characters being used in data processing centers in Taiwan, the characters used in Chinese libraries, the standard characters announced by

the Ministry of Education and many others.

Many notations and symbols of this paper follow those used in the volume I of CCCII.

2. Surveys of Chinese Characters

2.1 The most frequently used characters

The 4807 characters in volume-1 of CCCII are the most frequently used characters. This set is indispensable to any application. It has been shown[1,2,3] that, for teaching, writing and newspaper printing, the frequency of using this set is over 95%.

2.2 The character set for general data processing

A survey and analysis of characters has been conducted since the end of 1949.[4] The purpose of the survey is trying to find a set of characters which will be satisfactory to the present-day business data processing need, library applications and census applications, in Taiwan. A collection list of the survey is shown in Table-1. There are totally around 21004 characters, include the most frequently ones, in this set.

2.3 Variant forms of characters

A survey and analysis of variant forms of each character collected in section 2.2 was done in the past year.[5] The variant forms of a character include its equivalent forms, ancient forms, simplified forms and variations used by some communities. For simplified forms are concerned, this collection covers those used by ancient China, main-land China and Singapore.

At present, a character is limited to have no more than 6 variant forms and the total variant forms collected are 10793.

Table 1: The survey of characters for general data processing

Item	Resource Company Name or Book Title	Number of characters (Approx.)	Applied fields
1	Taiwan Automation Co.	9,600	These systems have been applied to the data processing works in the following areas: • Electricity • Library • Gas • Small business Company • Telephone • Banking • Government Organization • Water • Tax • Schools • Police
2	Wang Industrial Co.	10,499	
3	Financial Tax Center	11,000	
4	IPX Taiwan Ltd.	9,600	
5	Feng-Chia University	16,000	
6	National Police Administration	11,825	
7	Characters for Libraries	15,000	
8	Chiao-Tung University	9,129	Basic character set for all Chinese data processing application and researches
9	The code book of the Three Corner Coding Method	8,800	The most popular and well adapted by users.
10	The Comprehensive Dictionary of Chinese Character Index	9,600	
11	Chinese Characters for Telegraph Code	8,000	for data transmissions
12	Characters for Elementary school	4,600	for teaching
13	The Ministry of Education	12,701	These are the standard forms so far published

3. The placement of characters

3.1 The partition of characters

Characters are partitioned into blocks according to their frequency of usage. The CB1, the character block one, is assigned to indicate the 4807 most frequently used characters. The CB2, the character block two, is assigned to denote the set for general data processing purpose exclude the CB1. At present, all other characters may be grouped into CB3. This partition is shown in Figure-1.

3.2 Ordering

Within each character block, the ordering of character is arranged according to the Kang-hsi radical sequence first, and then, by its stroke-count. For those characters with the same radical and stroke-count, they are sub-arranged by stroke-order. The precedance of strokes is shown by the following descending order:

(1) A dot
(2) A horizontal stroke
(3) A vertical stroke
(4) A stroke down to the left
(5) A stroke down to the right

3.3 Handling Variant forms

A character that has variant forms is somewhat a unique property of Chinese language. A character and to variant forms usually have exactly the same pronunciation and meaning but different in their stroke image. Usually, they are interchangable in writings. But, while used as an identifier to name a person, a place or a thing, they are considered as different characters and are not allowed to be mixed up. In CCCII, the codes assigned to the variant forms are designed to have identical two right-most bytes, B_2B_1, as the code of its corresponding character. In other

words, the variant forms are placed at the same section and position just as its corresponding character occupied but in different planes.[1] By this arrangement, variant forms can be uniquely identified and easily be programmed interchangeably while needed. An example of variant forms is shown in Table-2.

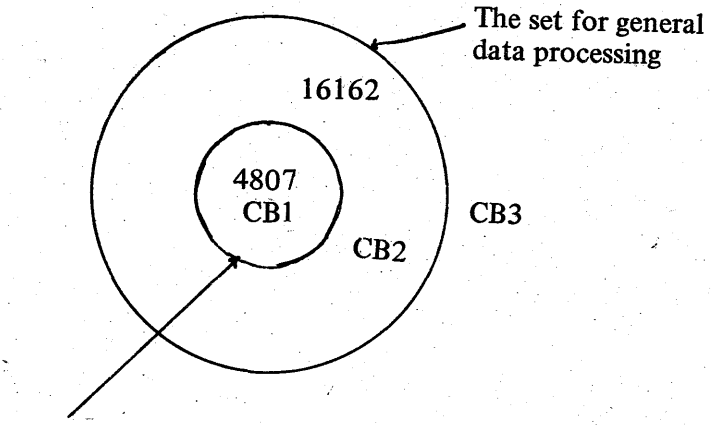


Figure 1: Partition of Chinese Characters

異體字形 CCCII 編碼結構實例

3rd byte B_3 \ B_2 2nd Byte / B_1 1st Byte	6 0 / 4 9	6 0 / 4 B	6 0 / 4 C	6 0 / 4 E	6 0 / 4 D	6 0 / 4 F	6 0 / 5 0	6 0 / 5 1	6 0 / 5 2	6 0 / 5 4	
2 1	頹	頤	顆	額	顏	題	顎	顓	類	願	normal form 此列為通用體
2 7	颓	颐	颗	额	颜	题	颚	颛	类	愿	simplified form 此列為大陸簡體
2 D	頽	𦣞		頟	顔	題	[illegible]		[illegible]	[illegible]	other variations 以下四列為同義異體
3 3	穨				顔		腭				
3 9	僓				顏		齶				
3 F	隤						齶				

Table 2: An example of the table of variant forms and their associated CCCII codes.

4. The Organization of CCCII

4.1 Compatibility

The CCCII is designed to be fully compatible with the 7-bit code environment specified by ISO646 in order to fulfill the need that some international business required. An example is the Chinese MARC format.

According to ISO2022, CCCII can also be used in 8-bit environment while all controls retain a structure compatible with the 7-bit structure. The escape sequence for CCCII is proposed to be ESC, 2/4, 4/2. This is a way to extend graphics with multiple by representation. For CCCII, each graphic character is represented by a 3-byte vector. This structure is illustrated in Figure-2.

4.2 Some revisions of the volume I of CCCII

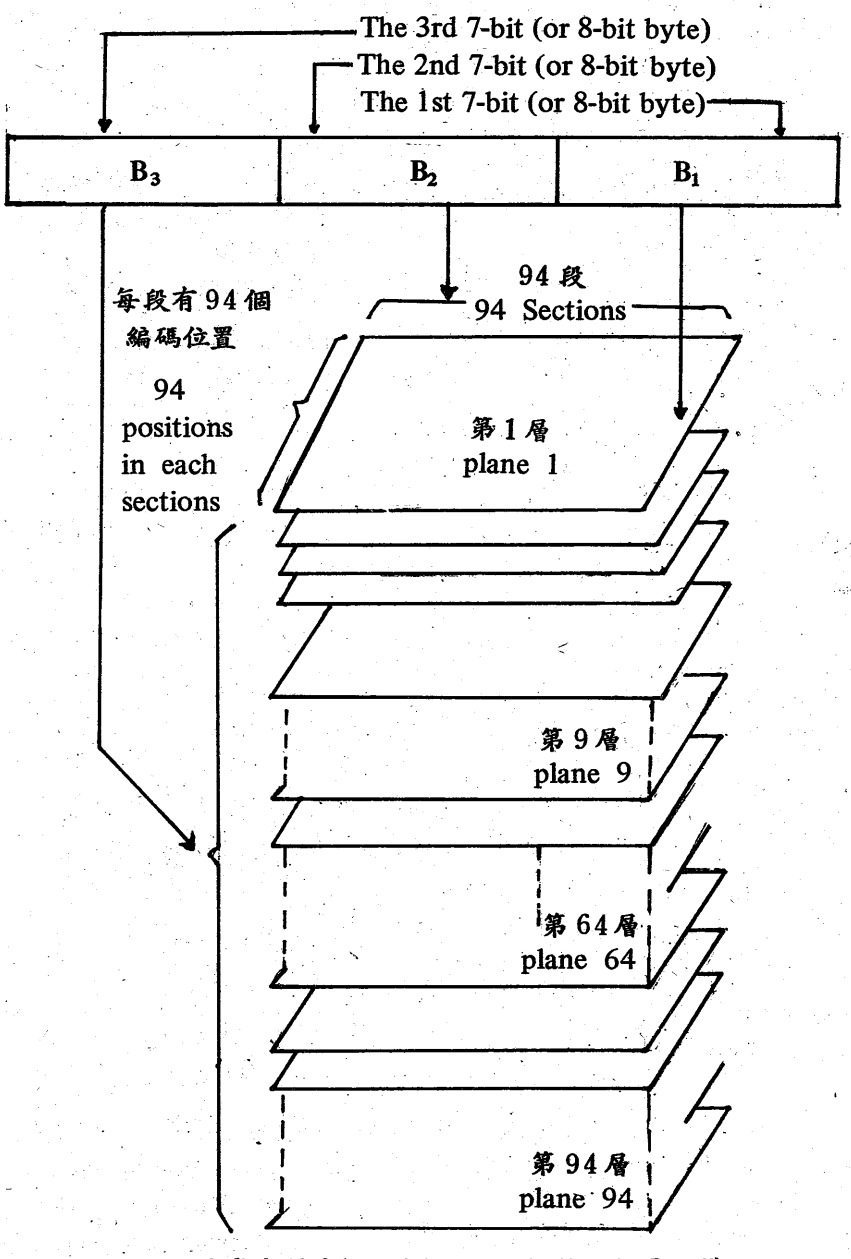
There are some technique changes made to the volume I of CCCII:

(1) The first section of each plane, as shown in Figure-3, will be reserved for control codes especially required for handling Chinese character strings.

(2) Two symbols, Ü and ¥ are added.

4.3 Layers

Started from plane 1, for each 6 consecutive planes form a layer. This structure is shown in Figure-4. The first layer is the layer for normal characters. From the second layer on are assigned for the variant forms and for future expansion.



三度空間 94 × 94 × 94個編碼位置結構圖

Figure 2 *The 3-dimensional structure of the whole 94x94x94 code*

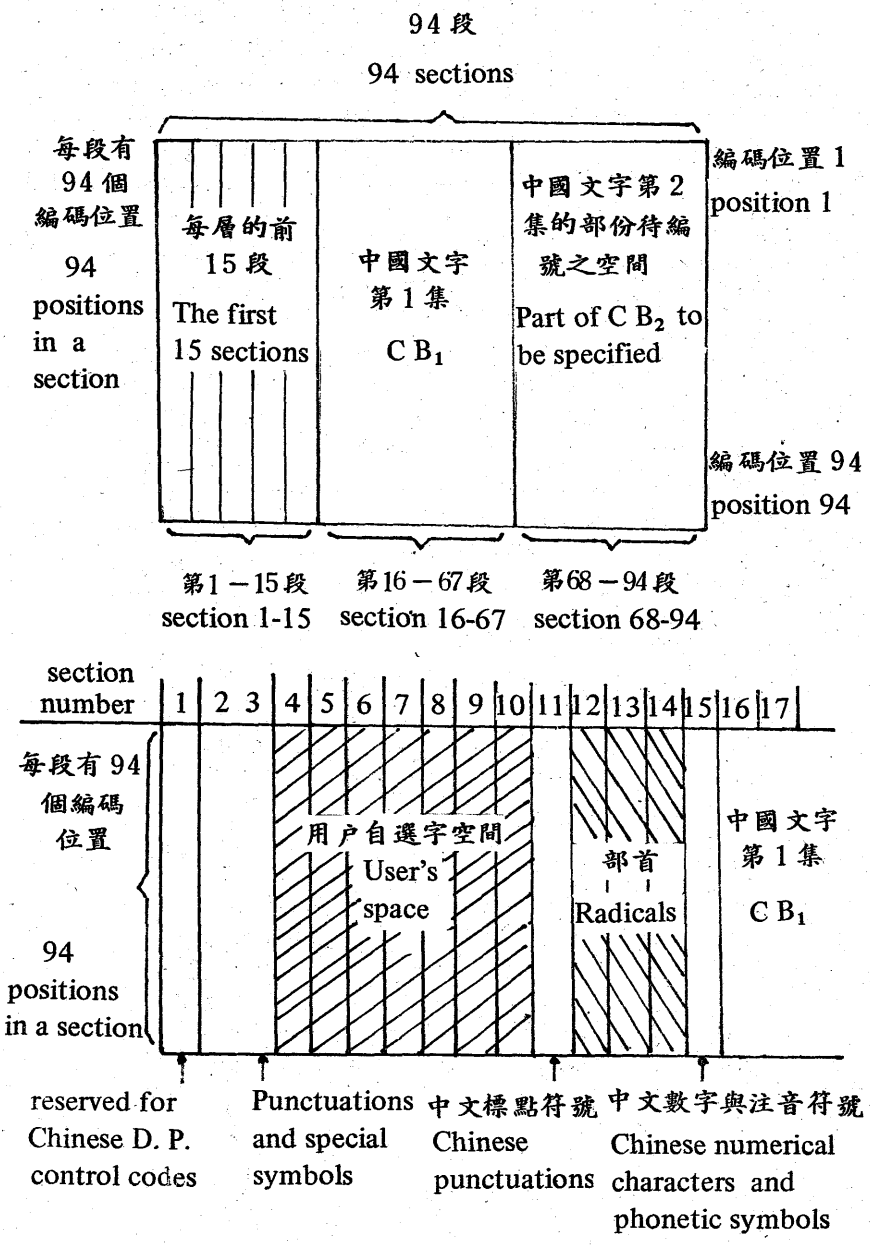


Figure 3　第一層的前 15 段結構圖

The structure of the first 15 sections in plane 1

5. Application Notes

5.1 The subsets of CCCII

The CCCII is intended to serve as an universal tool for Chinese Information interchange. Therefore, all symbols used by Chinese language and information handling will be collected and coded. But, from user's point of view, many applications do not need such a complete set of characters. In order to fulfill the above two to controversial requirements, the CCCII is so designed that the user can select a proper subset of CCCII to make the code more suitable for his own need.

The subsets of CCCII with 2-byte subcode length are listed in Table 3. Under this circumstances, by agreement between the interchanging parties, only 2-byte is needed to represent a Chinese Character. Please note that the code structure of all those subsets of CCCII is fully ISO646 compatible.

5.2 The radix-94 code for internal storage

The first 6 planes in the first layer (Fig. 4) have totally $94 \times 94 \times 6 = 53016$ positions. Since this number is less than $2^{16} = 65536$, the 3-byte code in the first layer can be compressed into a 16-bit or two 8-bit bytes as an unsigned binary number by applying radix-94 conversion. This conversion is an one to one, two way and unique conversion. Therefore, this code can be used as internal code for mass data storage to save storage space if the characters used are limited to the those in the first layer.

5.3 Code-length compression by commands

The first sections of each plane are reserved for commands used by Chinese data processing or defined by user. For some applications that 3-byte code must be used, user may define commands to switch among planes and leave the Chinese character code be two bytes. By this arrangement, the length of the code

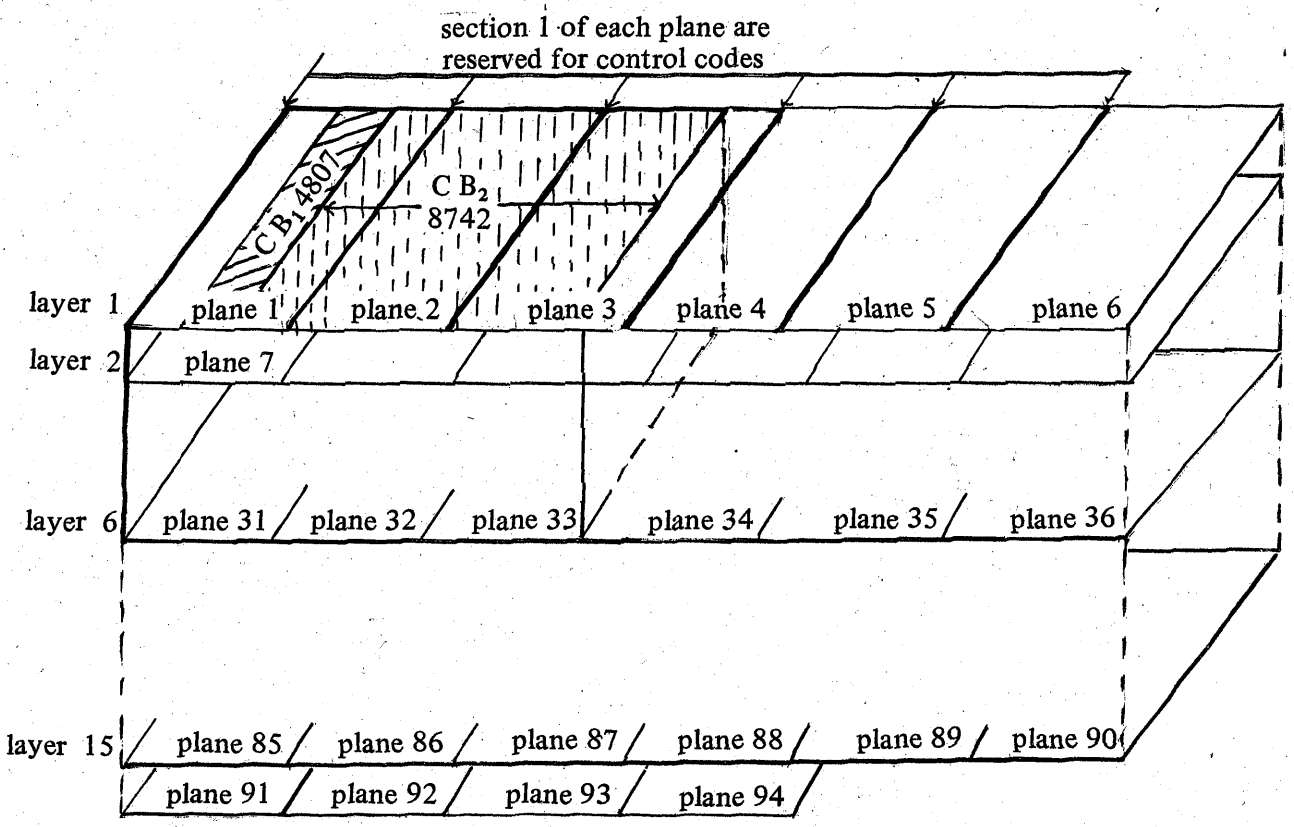


Figure 4: The Structure of CCCII

Table 3: The subset of CCCII with 2-Byte subcode length

Subset	No. of Characters	Character set Name	Publication	Sub-code length	Compatibility	User defined positions	Typical Applications
CCCII-1	4807	The most frequently used characters (plane 1)	Vol. 1 of CCCII	Two 7-bit Bytes	Fully ISO646 Compatible	658	Teaching, writing, newspaper printing, word processing
CCCII-2	22000	The character set for general data processing (plane 1,2 & 3)	Vol. 2 of CCCII	Two 8-bit Byte	Fully ISO646 Compatible	658	Business Data Processing Library, Census applications
CCCII-3	≃33600	Complete first 4 planes	Not yet	Two 8-bit Byte	Fully ISO646 Compatible	658	For more specific applications

for a block of characters may be significantly reduced.

As an example, let define two switch commands, as follows:

Command name	parameters	Comment
Switch plane, locked	plane number	4 bytes more for each switch of a group of character
Switch plane, unlocked	plane number	4 bytes more for each switch of a single character code.

Then, the third byte B_3, the pointer of the plane number, may be ommitted for all the characters. Assume the command name and parameters are 2-byte each, and there are 95% of chances to use those characters in plane 1, Then, the weighted average of code length can be computed as follows if only the unlocked switch plane command is used:

$$\begin{aligned} \text{averaged code length} &= 2 \times 95\% + 6 \times (1 - 95\%) \\ &= 1.9 + 0.3 \\ &= 2.2 \text{ bytes} \end{aligned}$$

The 7-bit bytes used to code command names and parameters are all belong to the set GO defined in ISO646. This makes the switching operation fully located within the processing operation of Chinese characters and hence fully ISO646 compatible.

6. Concluding Remark

The 31797 characters collected for the volume II of CCCII was published in two drafts, titled:

(1) Symbol and character tables of the volume II of CCCII
(2) The table of variant forms of the volume II of CCCII

We like to circulate those two drafts to all data processing

centers, organizations and persons who are interested in the automation of Chinese information. We welcome any comments, revisals, suggestions and criticisms so that we can revise the drafts and make the formal publication more suitable for the public need.

The collection of characters, the assuring of stroke image of each character, the assuring of the variant forms of each character and so many related analytical, statistical and laborious works were done by other colleagues of the Chinese Character Analysis Group. It is oweing to their endeavour that make this paper possible. Their contribution will also be presented in this workshop.

7. Acknowledgement

The Chinese Characters Analysis Group is deeply indebted to two non-profit foundations, namely: 明德基金會 and 元智基金會 for their general support.

Reference

1. The Chinese Characters Analysis Group, Chinese Character Code for Information Interchange, volume I; The Library Association of China. Taipei, 1980.
2 教育部，常用國字標準字體表；台北市，正中書局。
3. Lin, S, Basic Chinese Character Set for Computer Uses; National Chiaotung University, Hsiuchu, Taiwan, R.O.C. 1972.
4. C. F. Chou, Discussion on the Arrangement of Characters Used in Computers from the Viewpoint of Chinese Character Structure and Evolutionary Changes, this workshop, 1981.
5. C.K. P'an, A Survey of Various Forms of Chinese Characters, this workshop, 1981.
6. Related ISO publications: ISO646, ISO2022.