EXTENSION OF NAPLPS VIDEOTEX/TELETEXT STANDARD TO THE CHINESE LANGUAGE

Y.F. Lum
Department of Communications
Ottawa, Ontario
Canada

Cai Hengsheng and J.S. Riordan
Carleton University
Ottawa, Ontario
Canada

ABSTRACT

The paper describes the major features of the NAPLPS videotex/teletext standard and its extension to include the Chinese Hanji character set, creating what is termed C-NAPLPS. Various Hanji keyboard input techniques are examined. A C-NAPLPS videotex service can potentially interwork internationally.

1. INTRODUCTION

The NAPLPS (North American Presentation Level Protocol Syntax) is a presentation layer standard for encoding of information in videotex and teletext services. It is a fully recognized international standard as Data Syntax III of CCITT Recommendation T.101 for International Interactive Videotex Service[1]. It is also the information encoding standard for the System "C" Teletext as defined in CCIR Recommendation 653[2]. The NAPLPS was jointly developed by the Canadian Standards Association and the American National Standards Institute[3]. At the CCITT Plenary Assembly, November 1988, Melbourne, Australia, a Revised Recommendation T.101 was approved which included an extension of NAPLPS to all other languages that have code tables registered in accordance with the ISO 2375 standard. One set of Chinese Hanji characters has been registered as ISO Registered Set No. 58. The marriage of the Chinese character set with NAPLPS could result in a set termed C-NAPLPS.

2. THE KEY FEATURES OF NAPLPS

2.1 Independence of Hardware Technology

A fundamental design philosophy of NAPLPS is its independence from the terminal equipment implementation technology. Terminals may have high, medium or low resolutions and have different ranges of color capabilities. For example, two terminals, one with medium resolution such as terminals using the TV set as display, and one with high resolution such as that used in computer graphics, would both decode the NAPLPS signals and would show all the essential information, but the terminal with the better resolution will present a more pleasing picture. The NAPLPS design approach will ensure both forward and backward compatibility. Backward compatibility means that future terminals will be able to access old data. Forward compatibility means that existing terminals can receive all future command formats, including future enhancements of the current standard in an intelligent manner.

An important concept used is that of the Unit Screen. Coordinate systems are not expressed in physical dimensions nor in fixed grid formats such as 40 columns x 24 rows, but rather in normalized units of 0 to 1. Such a unit screen approach ensures that pictures will be displayed independent of the hardware configuration and objects within pictures will remain in the same relative position with respect to each other, even though the resolution of the physical display may be different.

2.2 Geometric Primitives

There are 5 geometric primitives, namely, POINT, LINE, ARC, RECTANGLE and POLYGON. Enclosed objects such as arcs, rectangles and polygons may be either filled or in outline form.

With these primitives one can draw very simple pictures such as graphs, pie-charts, etc., or very complex pictures. A fundamental design concept is that of overlay, that is to say, a geometric primitive will draw an object overlaying or covering what was drawn previously. Hence, one can create images as a painter or an artist would. For example, one can create a blue sky background on the whole screen by using the FILLED RECTANGLE primitive and defining the BLUE color attribute. Then one can draw a green lawn using the FILLED POLYGON and the GREEN color attribute. The green polygon will overlay on part of the previously drawn blue screen and the total picture now is a green lawn with the blue sky. One can go on and draw a house, etc. What in fact is taking place is that a set of drawing instructions is sent to the terminal in a natural way to compose the picture. That is why we call these Picture Description Instructions (PDI's).

The INCREMENTAL POINT, INCREMENTAL LINE and INCREMENTAL POLYLGON geometric primitives included in the standard, allow efficient coding of certain types of drawings. The INCREMENTAL POINT function allows photographic type pictures to be drawn.

A very important feature of NAPLPS is that by using either the LINE or the INCREMENTAL LINE primitive, Chinese Hanji characters of any complexity can be drawn and displayed on the screen. Further the drawn character can be stored as DRCS.

2.3 DRCS

Dynamically Redefinable Character Set (DRCS) is a character coding technique in which the shapes of the character symbols are not pre-defined, but has to be first down-loaded from the data base to the terminal and then used by the terminal.

The DRCS scheme included in NAPLPS is independent of the hardware implementation of the terminal. This is done by a down-loading process which uses the unit screen for the description of the symbol or character shape, which is then scaled down to the storage character size as implemented within each terminal. A very efficient way of drawing characters (such as Hanji) on the unit screen is by means of PDI primitives. Only the shape of the DRCS character is stored. Attributes such as color, display character sizes, etc., apply only when the character is displayed, just like a normal alphabet character.

2.4 Mosaics

Mosaics are character blocks of different shapes used to compose pictures. Mosaic pictures generally have poor resolution. A mosaic code table, which is the union of the French and British mosaic standards is included in NAPLPS, primarily to facilitate inter-working with European systems. Mosaics may be used to fill patterns for textured filled polygons.

2.5 Macro

Macro instructions are a string of coded information of any sort down-loaded by the data base into a terminal and identified by a macro name. Later, when that macro name is called or invoked, that string of information will be processed or in the case of a special macro, called the transmit macro, the coded information is sent back to the data base. Macros are therefore very efficient ways of coding and transmitting information to a terminal which uses that same information several times. In the case of the transmit macro, it provides the user terminal with what may be called programmable function keys. That is to say, the functions of these keys depend on the macro instructions being sent back to the data base computer.

2.6 Alphanumeric Text

The coding of alphnumeric text in NAPLPS is based on the CCITT and ISO standards using the composition method. That is to say, accented Latin based characters are coded by a combination of a non-accented character in the Primary Set together with a non-spacing accent symbol in the Supplementary Set. In NAPLPS, the Primary Set is chosen to be the 7-bit ASCII code table which is the de facto world standard for ordinary computer terminals. The objective is to ensure that an NAPLPS videotex terminal, if equipped with a keyboard, can be used as a normal computer terminal thereby much enhancing the utility of NAPLPS terminals. The composition method of using the Primary Set and Supplementary Set caters to essentially all Latin based languages of the world.

2.7 Attribute Controls

These are of prime importance to the display of images on a terminal. The major attribute controls are the following. The list is not exhaustive.

2.7.1 DOMAIN

The control determines the accuracy to which the coordinates of the drawing point is specified. DOMAIN also specifies the size of the logical picture element.

2.7.2 Color

Three color modes are specified:

Color Mode "0": Colors are directly defined in terms of Green, Red and Blue (GRB) components.

Color Mode "1": Colors are indirectly defined via a color map. A range of colors may be chosen from the color palette of several hundreds or thousands of colors

Color Mode "2": This provides both a foreground as well as a background color.I uses the same color mapping technique as in Color Mode 1".

2.7.3 Blink

This feature allows specified colors to change in a time sequence, giving rise to color animation effects.

2.7.4 Texture Filling

Filled drawings of arc, rectangle or polygon can be by texture masks of 8 kinds, 4 of which are pre-defined in the standard and 4 can be down-loaded from the data base.

2.7.5 Character Size Scaling/Rotation/Scrolling

Characters can be displayed, can be continuously scaled to any size, and can be rotated in the four major directions. Text can be displayed as 40 columns by 24 rows, 40 column by 20 rows or any other format. Scrolling of text is provided.

2.7.6 WAIT

The WAIT command causes the execution of instructions and presentation of drawings or tex to be delayed at increments of 1/10th of a second. The 1/10th second increment is chosen t be compatible with both 50Hz and 60Hz interlaced TV scanning systems.

3. INCLUSION OF CHINESE CHARACTER SET

A fundamental approach of the NAPLPS is its adherence to the principles and concepts of the ISO 2022 Code Extension Standard. Using these principles, it is easy to expand NAPLPS to cater to any additional language needs. Expansion of NAPLPS to C-NAPLPS, using the international standard of ISO 2022, ensures that the C-NAPLPS is in full compliance with Data Syntax III, CCITT Recommendation T.101 (1988). For international videotex interchange this is a important consideration.

Table 1 presents the complete set of Escape Sequences for designation of the Control and G-sets of the C-NAPLPS.

The ISO 2022 scheme allows the simultaneous designation of four G-sets, namely G0, G1, G2 and G3. Further, the designation of any of the G-sets can be changed at any time without any restrictions.

4. THE HANZI KEYBOARD INPUT

Hanzi is a set of some 60,000 ideographic characters. Of these, the most frequently used group of 3000 - 7000 are sufficient for everyday

use. Hanzi is composed of radicals, which in turn are made up of strokes. There are several basic strokes, including the point, horizontal line, vertical line, and the stroke to the right and to the left. The well known Kang Xi dictionary identifies 214 distinct radicals. Chinese is a tonal language, with about 1280 syllable-tone combinations. Frequently there is an unavoidable many-to-one mapping between the Hanzi character and its pronunciation.

Because of the complexity of Chinese ideographic characters, keyboard entry presents a major challenge. Early Hanzi keyboards contained the full set of about 7000 keys, one for each of the commonly used characters. The use of this device was restricted to a relatively few highly trained operators. On the other hand the standard western keyboard, widely used in both typewriters and microcomputers, is not difficult to master. Its use for Chinese entry requires a suitable encoding technique. About 500 such methods have been proposed, and over ten of them have been adopted by commercial Chinese computing systems. However, no single input method can cater to all the needs universally.

Existing input techniques may be summed up under three categories: shape-based, pronunciation-based and a combination of the two. In view of the range of input systems, a Chinese keyboard should cater to a number of alternatives. The principal pronunciation-based method is Pinyin. Based on the Latin alphabet, it has served as the standard sound annotation system for China since 1958. It is attractive mainly because of its simplicity, but it has several drawbacks, especially: (a) confusion associated with differences in dialect; (b) hard-to-pronounce characters; and (c) high duplication rate. On the other hand, shape-based methods are more easily acceptable because Chinese characters are ideographic by nature. Consequently methods based on radicals and strokes have become the main ones, while the Pinyin method is a supplement. The future trend is probably towards procedures based on a combination of shape and phonetic attributes.

In addition to keyboard input methods, Hanzi Phonetic Recognition and Optical Character Recognition (OCR) techniques have developed rapidly, and on-line recognition of handwritten Hanzi input is becoming a practical supplementary method.

4.1 Input Methods

The previous paragraphs examined three generic forms of keyboard entry. It is proposed that specific examples of these be adopted for the creation and editing of Chinese text files for videotex:

(a) KIM, an intelligent radical input method developed in conjunction with Kaihin Research Inc. of Toronto[4]. Users of KIM enter Hanzi characters as a set of radical strings just as they would normally write them. The built-in intelligence allows the system to accept out-of-sequence radical strings, and to tolerate to some extent input errors such as missing and/or superfluous strokes. It is not necessary to memorize an input coding technique. The method is thus easy to learn for the non-professional. The KIM keyboard layout is given in Fig.1.

(b) Augmented Pinyin. When the user enters the Pinyin code, the homonyms (or a subset of them, at least) appear on the bottom of the screen. The user can often select the desired character without entering all of its constituent radicals.

(c) PCA, the Phonetic Chinese Alphabet.[5] This combination of pronunciation-based and shape-based methods employs a newly created tone-based 85 element alphabet set, as shown in Fig 2. It permits simultaneous phonetic-tone spelling and relies on radical-styled letters to provide rational homonym resolution.

(d) Matrix table lookup according to the rows and columns of the code table of the ISO Registed Set Number 58, i.e. Chinese Standard GB 2312-80. This is also the internal code system for Chinese computers.

Any of these four input methods may be selected by function keys, each one being selectable at any time during the typing process. Active memory space can be saved through the use of ROM chip GB 5199A, containing the complete 8192 Hanzi character set in the GB 2312-80 Standard. Input methods can make use of software or additional ROM chips.

5. CHINESE/ENGLISH VIDEOTEX NETWORK

In the establishment of a Chinese videotex network, advantage can be taken of links to other networks containing Latin alphabet text information, typically in English. The major components of such a network are as follows:

* Information provider system (IPS) for Chinese/English text file creation and editing, Chinese/English text and graphics integration;

* Videotex central processing and database facilities for Chinese/English database setup and management, Chinese/English frame file transaction, conversion, reception, display, printing, etc.;

* Communication network for connecting videotex center and user terminals;

* User terminals for Chinese/English frame file reception and display.

An example of such a possibility would be based on the new ALEX system, introduced recently by Bell Canada. Named after Alexander Graham Bell, the inventor of the telephone, the ALEX system is aimed at a wide range of applications, including education, home banking, entertainment and restaurant information, home shopping, travel/transportation, news, financial services, sports, games, messaging and real estate information. Since this communications service is NAPLPS based, its subsystems frame creation, transmission, reception and display could be augmented by the addition of the Hanji registered character set and the Chinese keyboard entry system described earlier in this paper. Such a videotex system could be termed a C-NAPLPS videotex system. Further, research has been successfully completed which demonstrated the ability of a NAPLPS terminal to receive the Japanese CAPTAIN videotex transmission, which contains the Japanese Kanji character set. A similar technique could easily be developed to convert C-NAPLPS whereby an ALEX terminal could receive and display C-NAPLPS information. In the reverse direction, a C-NAPLPS terminal should be able to receive information from an ALEX database. Indeed, the proposal outlined in this paper could lead to an international interworking between ALEX and C-NAPLPS services.

6. CONCLUSICN

This paper has examined the feasibility of extending the NAPLPS videotex standard to the Chinese language. The existing GB 2312-80 standard with its 8192 Hanzi characters can be realized in a ROM implementation as part of a NAPLPS system. However, keyboard entry remains a key issue, regardless of the standard. It is recommended that Pinyin and conventional shape-based approaches be enhanced by the use of software intelligence to reduce user input effort. A recently developed 85-element phonetic Chinese "alphabet" shows particular promise in allowing the use of a standard keyboard as an effective entry tool.

It is recommended that both videotex and broadcast teletext use the same C-NAPLPS for encoding the data base information content. This will enable easy interchange of information content, and promote larger scale production of common user equipment and components.

Implementation of C-NAPLPS as proposed here fully conforms to the CCITT and ISO standards and enable international interconnection of Chinese and western world videotex services which should yield commercial, educational and social benefits.

REFERENCES

[1] CCITT Recommendation T.101, Annex D, Data Syntax III For International Interactive Videotex Service, Red Book, Fascicle VII.3 - ANNEX III.

[2] CCIR Recommendation 653, Teletext Systems, (1986), Volume XI - Part 1, Broadcasting Service (Television).

[3] Videotex/Teletext Presentation Level Protocol Syntax, ANSI X3.110 - 1983 and CSA T500 - 1983.

[4] A. Chiu, F. Wong, H. Cai "An Intelligent Knowledge-Based Chinese Input System" International Conference on Chinese Information Processing, Beijing, August 1987.

[5] Victor C. Yeh "The Key to Chinese Language Modernization and Computerization - The Phonetic - Tonal sort," Proc. of 1987 International Conference on Chinese Information Processing, Vol. 1, 487-497, Beijing. August, 1987.

[6] Hengsheng Cai, C. Douglas O'Brien, and J. Spruce Riordon, "A Proposed Scheme for Chinese Videotex Standard", Computer Processing of Chinese & Oriental Languages (ISSN 0715-9048), Vol. 2, No. 4, 181-197, October 1986.

[7] Hengsheng Cai, J. Spruce Riordon, and C. Douglas O'Brien, "A Proposed Chinese Language videotex and Teletext Standard". International Conference on Communication Technology, Nanjing, China, November 1987.

[8] Zhu Wei-le, Gu De-ren. "The Chinese Character Capability Extension to NAPLPS and Teletext Standardization in China." International Broadcast Information, Vol. 1, No. 4.

TABLE 1

ESCAPE SEQUENCE FOR DESIGNATION OF C- AND G-SETS OF C-NAPLPS

Escape Sequence	Set to be Designated
Control Sets	
ESC 2/1 4/11	C0 Set
ESC 2/2 4/6	C1 Set
94 Character Sets	
ESC I 4/2	Primary Character Set (ASCII)
ESC I 7/12	Supplementary Character Set
where I is 2/8, 2/9, 2/10, 2/11 for G0, G1, G2, G3, respectively.	
96 Character Sets	
ESC I 5/7	PDI Set
ESC I 7/13	Mosaic Set
ESC I 2/0 7/10	Macro Set
ESC I 2/0 7/11	DRCS Set
where I is 2/13, 2/14, 2/15 for G1, G2, G3, respectively.	
Two-Byte Sets	
ESC 2/4 4/1 (For G0 set)	Chinese Hanji
ESC 2/4 I 4/1	Chinese Hanji
ESC 2/4 I 2/0 4/0	Two-Byte DRCS Set
where I is 2/8, 2/9, 2/10, 2/11 for G0, G1, G2, G3, respectively. (except for G0 set for Hanji)	

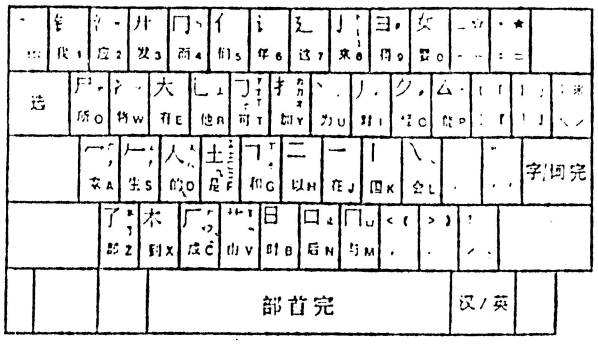

Figure 1 --- The KIM keyboard layout

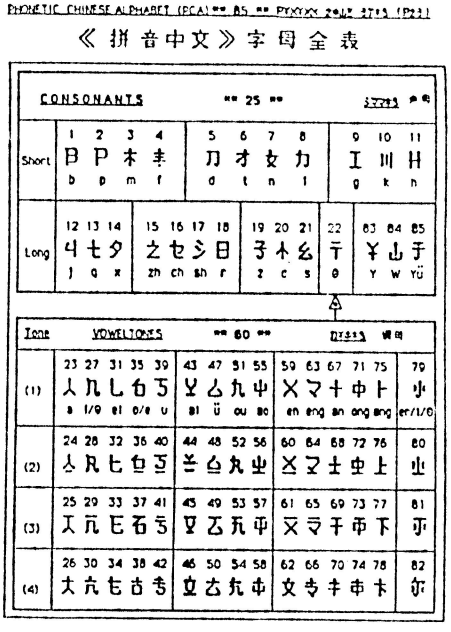

Fig2. Phonetic Chinese Alphabet