International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.9 (2015), pp.31-40
http://dx.doi.org/10.14257/ijmue.2015.10.9.04

Dynamic Generation and Editing System for Wrongly Written Chinese Character Font

Li Xiao[1] and Li Qingsheng[1,2]

[1]*School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, Henan, China*
[2]*Institute of Digital Inscriptions on Bones/Tortoise Shells, Anyang 455002, Henan, China*
joylx@163.com, aiaixiaoya@126.com

Abstract

The uniqueness of Chinese makes the Chinese language have become a hotspot in language learning. In view of the problem of wrongly written character teaching in Chinese language teaching, it provides a simple, convenient and efficient input method of wrongly written characters and realizes a dynamic generation and editing system for wrongly written Chinese character font, which solves the problems of real-time edit, coding and input of wrongly written character in editing process using dynamic editing technology, and provides a convenient input method of wrongly written character in editing, printing, typesetting and the research of digital Chinese language teaching. This method can also be used in ancient variant dynamic editing, generation and processing of ancient variant, Oracle, bronze and folk combined characters.

Keywords: *Chinese character, Wrongly written character input method, Dynamic editing, Digital Chinese language teaching*

1. Introduction

With the increase of economic strength, economic and cultural exchanges between China and the world grow increasingly, with the phenomenon of "Chinese fever" frequently heating up. Chinese attracts worldwide attention with its unique charm and it is precisely the unique characteristics of Chinese that make Chinese learning more difficult. In the last analysis, this certain difficulty is determined by the complex structure of Chinese, in which Chinese characters writing is the most difficult to learn. Beginners are easy to write wrong words, and the writing errors with different Chinese learners have different rules, thus causes a certain degree of difficulty in Chinese characters teaching. The status of the teaching difficult of Chinese characters has restricted the development of the domestic Chinese language teaching and teaching Chinese as a foreign language (TCFL). Although the writing mistakes for Chinese characters can hardly be avoided, but there is no large-scale error statistical analysis results for wrongly written characters which can provide guidance and reference for the Chinese characters teaching. On one hand, there are many difficulties in computer processing of wrongly written characters (*e.g*. editing, coding, input and output, printing and typesetting of wrongly written characters *etc*.). On the other hand, it currently lacks coding scheme of wrongly written characters in line with international standards and simple and effective input method of wrongly written characters in the field. Thus it makes difficulties for computer processing of wrongly written characters.

Therefore, it is very necessary for researching and designing a simple and effective

ISSN: 1975-0080 IJMUE
Copyright ⓒ 2015 SERSC

generating and processing scheme for wrongly written Chinese characters.

2. Application Requirement and Present Situation of Generating System for Wrongly Written Chinese Characters

2.1. Chinese Characters Represented in the Computer

To use Chinese characters in the computer system, the first problem to be solved is how to input Chinese characters into a computer. The prerequisite of processing the information of Chinese characters in the computer is to encode each Chinese character, and these codes are collectively referred to as Chinese characters code. But the characteristics of many words and complex shape of Chinese characters make the Chinese characters have a different encoding rule compared to the ASCII code. Therefore, our country introduced a unified coding standard specially used for Chinese characters information exchange between computer systems: "character set for information interchange of Chinese characters encoding", that is, Chinese characters GB code, also called "interchange code", and all Chinese characters codes should follow the standard.

Chinese characters machine code, also known as the "Chinese characters ASCII code", and the "code" for short, refers to the code composed of 0 and 1 in binary notation for computer internal storage, processing and transmission of the Chinese characters, also it is formed by the highest GB code byte after treatment.

A set of computer keyboard symbols designed for the convenience of Chinese characters input is called "Chinese characters outer code", also called "input code". The external code commonly used includes phonetic code (such as spelling), font code (such as five strokes), water code (such as location code) and sound form code (such as smart ABC) *etc*. The input code in the computer must be converted into machine code, then can be carried on storage and processing [1].

In order to output Chinese characters glyph with using the computer, it usually needs to store the related information of Chinese characters font in the computer, thus the font has been formed. Digital information of Chinese characters glyph stored in the font called Chinese characters font code, a Chinese characters font code corresponding to a unique character code. There are a variety of classification for fonts, based on the different coding standard it can be divided into GB2312-80 font, GBK font, GB18030 font *etc.*; in terms of language it can be divided into Chinese font, foreign language font, graphic symbols *etc.*; according to the format it can be divided into True Type font, PostSript font and OpenType font *etc.*

2.2. Existing Problem and Demand in Generation of Wrongly Written Chinese Characters

It is a very common thing of Chinese characters inputting typesetting and printing using computer currently in the field of office automation and printing. Therefore, the computer font must be used in the treatment of Chinese characters. It often needs to expend setbacks, however, if the Chinese characters input and printed is not in the computer fonts. There are two commonly used methods: one is making Chinese characters not in the computer fonts with character-creation program; the other is using images to substitute the Chinese characters temporarily. As the wrongly written characters are fonts without words, little wrongly written characters can use the above method to generate. But with more and more people learning Chinese characters and the exponentially growing phenomenon of types and number of writing errors, it could not meet the need of digital Chinese language teaching by making

 Copyright ⓒ 2015 SERSC

Chinese characters with character-creation program and editing wrongly written characters images.

Many scholars began to research in word editing and recognition and have made some achievements, typically such as the "wrongly written Chinese characters processing solution based on Unicode" [2-3] written by teacher Li Xiao-qing and Lin min in the Inner Mongolia Normal University, which expressed the wrongly written character code with an orthography as the center and "IVS" of orthography and variant selector based on ideographic variant sequences (IVS) standard from Unicode 5.1, and applied OpenType font technology for input and output.

The word processing method above stores the wrongly written characters using idle area in standard font or infrequently used Chinese characters code region on the basis of the original font, whose biggest deficiency is the occupation of valuable coding space of Chinese characters. And with the expansion of the scale of wrongly written characters, these reserved intervals will soon be exhausted. For example, customized Chinese GBK code are [AAA1-AFFE], [F8A1-FEFE] and [A140-A7A0], just three sections, sum of 1894; customized Unicode code is [E000-F8FF], total of 6400 [1], only for 6400 even though each Chinese character take one wrong word, but the reality is that one Chinese character has far more than one wrong word. So the existing word input and processing method has many defects in the processing of large quantities of wrongly written characters. In addition modern Chinese characters font library is based on the font file as a unit, each font file contains a kind of different encoding Chinese characters, and each Chinese character is described by glyph outline which makes the description of wrongly written characters more trouble because the wide variety reason of wrongly written Chinese characters decided by its generation. Outline font can ensure the output font quality, but is not conducive to edit and dynamically generation of wrong word font [4-6]. The reason that wrongly written Chinese characters generated determines its variety, it will become more trouble using glyph outline for character description.

Therefore, it needs to find an input and editing method for wrongly written Chinese characters based on font description [7-9], so as to open the edit number of wrong words, facilitate user input, better serve the publishing and printing of wrongly written Chinese characters, and provide digital typo editing and printing environment for Chinese teaching especially teaching of Chinese as a foreign language in particular.

3. Dynamic Description Library for Wrongly Written Character Font

According to the requirement above, we propose a method based on font coding of wrongly written Characters, which establishes a dynamic description library for wrongly written characters font (shorted for DDL in the following), and makes an dynamic vector description of wrongly written characters font using stroke segment and stroke unit[10-11], then finds the feature points in the glyph skeleton, and carries on the quantification and storage by feature point, and ultimately realizes font coding of wrongly written Characters. The application of DDL solves the difficulty of font dynamic editing and font transformation caused by using glyph outline description of wrongly written Chinese characters, solves the problem of difficult editing and difficult writing in wrongly written Chinese characters teaching.

3.1. Description of Wrongly Written Characters Font

According to the writing method of modern Chinese characters, we introduce the concepts of directed stroke segment and directed stroke unit to describe glyph skeleton of wrongly

Copyright ⓒ 2015 SERSC

written Characters in DDL. The directed stroke segment is a directed line, which can recognize stroke starting, stroke wielding and the stroke collection in the process of font generation of wrongly written Characters. The coordinates of start point and ending point of each segment are represented as "shi" point and "zhu" point. Let (X_i, Y_i) be "shi" point and (X_j, Y_j) be "zhu" point, so the one-dimensional vector of the directed segment S_{ij} is:

$$S_{ij} = (X_i, Y_i, X_j, Y_j) \quad (1)$$

The stroke unit is a complete stroke structure composed of one or more directed segments, supposing one stroke unit consists of N segments, so this stroke unit can be described as a vector of E_n, $E_n = (S_{i_1j_1}, S_{i_2j_2}, \ldots, S_{i_nj_n})$. For any $K \in \{1,2,\ldots,n\}$ and $S_{i_kj_k}$ is shorted for S_k, the stroke unit above can be recorded as below for short:

$$E_n = (S_1, S_2, ..., S_n) \quad (2)$$

In addition, the "shi" point of the first segment S_1 is called the starting point of E_n, and the "zhu" point of the last segment S_n is called the ending point of E_n.

3.2. Definition of Stroke Unit

In the font description library, boundary point is used to segment each stroke unit. Each stroke unit has the starting point and the ending point, in order to make the starting point and the ending point of different strokes not confused, defined symbols are added before the starting point of each stroke unit so as to define stroke unit, and the defined symbols are called boundary point. Suppose the boundary point is $D = (D_1, D_2)$, then the description vector of E_n is:

$$E_n = (D_1, D_2, S_1, S_2, ..., S_n) \quad (3)$$

3.3. Coding Description of Wrongly Written Characters

A wrongly written Chinese character is a collection of its stroke units. For the convenience of computer recognition, this collection is represented as the arrangement of stroke units, according to Chinese characters written order. Suppose one Chinese character consists of stroke units "m": $E_{n1}, E_{n2}, \ldots, E_{nm}$, thus the description vector of this wrongly written character is:

$$ZX = (E_{n1}, E_{n2}, ..., E_{nm}) \quad (4)$$

The description vector of wrongly written characters are processed into codes in the description library which are stored in a text file, and in order to define different wrongly written character codes, defined symbols "$H=(H_1, H_2)$ and $T=(T_1, T_2)$" are added before the first stroke unit and after the last stroke unit, thus the description vector of this wrongly written character is:

$$ZX = (H_1, H_2, E_{n1}, E_{n2}, ..., E_{nm}, T_1, T_2) \quad (5)$$

3.4. Dynamic Description Algorithm

The main function of dynamic description algorithm is to regulate and store the stroke units information after drawing and adjustment. The steps of the algorithm are as follows:

Step 1: open font description library and initialize variables，including the initialization of boundary point D, starting point H, ending point T, the number of stroke units ele_num and font description library ZXDATA(i).

Step 2: select the type of operation. If the operation is "Ins", then insert the stroke unit; if

 Copyright ⓒ 2015 SERSC

the operation is “Mov”, then move the stroke unit; if the operation is “Del”, then delete the stroke unit; if the operation is “MovDot”, then move the selected point(“shi” point or “zhu” point); if the operation is “Change”, then change the thickness of stroke unit; if the operation is “Copy”, then do transparent copy; if the operation is “NoOper”, then turn to step 3.

Step 3: save the operation and close the font description library.

Inserting stroke units can be achieved through inserting each stroke segment of the stroke unit one by one, and moving the whole stroke unit can be achieved through modifying each point of the stroke unit. In conclusion, the creation process of DDL is shown in Figure 1:

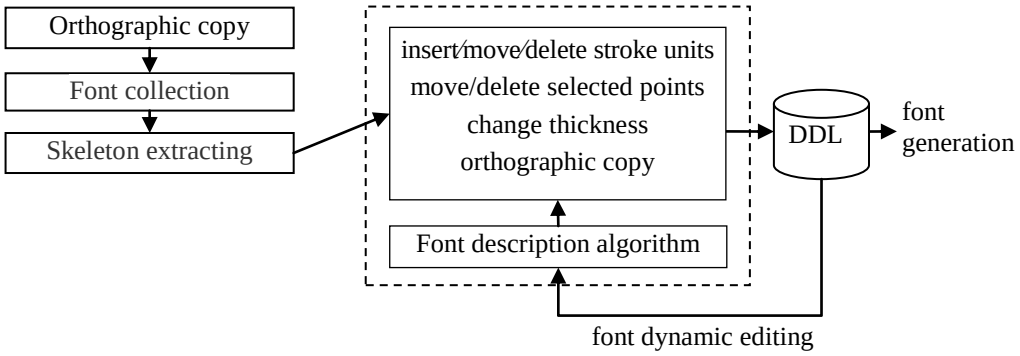


Figure 1. Process of DDL Creation

It can be seen from Figure 1 that the wrongly written characters are dynamically edited from the standard characters, so we establish the connection between the two through the list, the list node structure is as shown in Figure 2, in which identifier domain “Tag” values 0 or 1(“0” means standard characters, and “1” means wrongly written characters), chain domain “Link” stores a pointer to the next node in the same list, coding domain “Code” stores the codes of that Chinese characters.

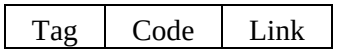


Figure 2. Structure of List Node

When editing the wrong word, first enter the correct word in word document and then depict the skeleton of the orthographic (*i.e.* stroke units) using "transparent copy" in software. The system will record information of the feature points and stored the orthographic codes in head node of the list, then edit the wrong word using operations (such as moving the stroke unit) provided by the software based on orthographic font, and last save the operations, thus the wrong word codes will be stored in the node and inserted into the corresponding list, and so on. While editing a new word, it can be stored in another list. And all head nodes of the lists will be established orthographic index in order to facilitate retrieval. When exiting the system, the system will automatically update all the font codes and generate the recent text file to ensure the smooth implementation of the initialization when opening the description library next time.

3.5. Extraction and Encoding of the Feature Points

According to the above description of dynamic library description, the extraction of feature points relates to the extraction of stroke segments and stroke units of wrong written characters font. The extraction algorithm of stroke units in font can be achieved through searching

Copyright ⓒ 2015 SERSC

boundary points, and the extraction algorithm of stroke segments can be achieved through analyzing "shi" points and "zhu" points of stroke units. So the extraction algorithm of feature points is implemented as follows:

Step 1: open font description library and initialize variables.

Open ZXscript

INT by_num←0, bd_num←0;

POINT D←(m,0), H←(m,0), T←(m, m);

ZXDATA(i)←{m, 0, m, m}

Step 2：compare the types of feature points. If the type is "boundary point", then turn to step 2.1; if it is "shi" point, then turn to step 2.2; if it is "shu" point, then turn to step 2.3; else turn to step 2.4.

Step 2.1: add 1 to the number of stroke units, *i.e.* by_num= by_num+1.

Step 2.2: add 1 to the number of stroke segments, *i.e.* bd_num= bd_num+1.

Step 2.3: save the coordinate of feature points.

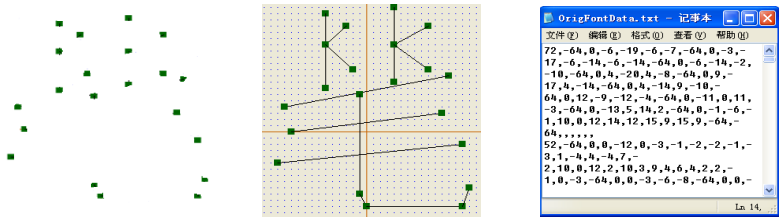
Step 2.4: the extraction of the first feature point is over. Turn to step 2 to continue to extract the next word.

Step 3: save and close font description library.

If the description vector in font description library is $ZX=（H_1，H_2，E_{n_1}，E_{n_2}，...，E_{nm}，T_1，T_2）$, the wrongly written characters font codes of featutre points through extraction algorithm of feature points is: $E_n = (by\_num , bd\_num, S_1, S_2, ..., S_n)$.

For example, the feature points of "毫" through extraction algorithm of feature points is shown in Figure 3 (as shown in Figure 3-a), word recognition program gets the font of "毫" by connection according to these feature point codes (as shown in Figure 3-b). So any wrongly written characters font can be dynamically presented in this system.

The extraction of the characteristics of wrong written characters font has provided the possibility for the coding of wrong written characters font. For example, the font corresponding to "毫" consists of 10 stroke units, 13 segments and 21 feature points, the coding of the feature points is: "72,-64,0,-6,-19,-6,-7,-64,0,-3,-17,-6,-14,-6,-14,-64,0,-6,-14,-2,-10,-64,0,4,-20,4,-8,-64,0,9,-17,4,-14,-64,0,4,-14,9,-10,-64,0,12,-9,-12,-4,-64,0,-11,0,11,-3,-64,0,-13,5,14,2,-64,0,-1,-6,-1,10,0,12,14,12,15,9,15,9,-64,-64,,,,,,"(as shown in Figure 3-c).



3-a 3-b 3-c

Figure 3. Diagram of Connection of Feature Points of Character Font of "毫" (Left is Feature Points, Right is Connection of Feature Points)

4. Dynamic Generation and Editing System for Wrongly Written Character

Combining the above algorithm, this paper creates an input system of wrongly written character for real-time dynamic editing by making a font library for wrongly written characters. The system includes: editor module for wrongly written character font, feature

 Copyright ⓒ 2015 SERSC

extraction module for wrongly written character font, encoding module for wrongly written character font, input module for wrongly written characters and real-time dynamic editing module for wrongly written characters (as shown in Figure 4).

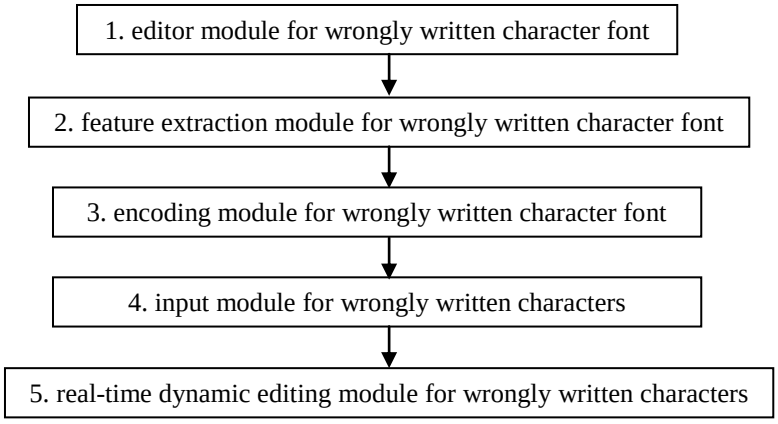


Figure 4. Module Diagram of Input System of Wrongly Written Character and its Realization Method

(1) Editor module for wrongly written character font: edit the wrongly written character that the user needs real-time and dynamically, make visual modification and combination on stroke structures based on the orthography, such as insert, move and delete stroke units, insert, move and delete selected points, transparent copy and change the thickness, and so on. Then the system will transfer the edited structure information of wrongly written character font to feature extraction module for wrongly written character font.

(2) Feature extraction module for wrongly written character font: analyze the structure data of wrongly written character font received, extract the feature points of the wrongly written character using extraction algorithm of feature points, and transfer the feature point data to the encoding module for wrongly written character font.

(3) Encoding module for wrongly written character font: encode and store the feature point data extracted from feature extraction module for wrongly written character font through encoding algorithm for wrongly written character font.

(4) Input module for wrongly written characters: input the corresponding key code through the key board (the system temporarily can only use 26 letter keys and 10 digital key for wrongly written characters input), the program will display the wrong word in the editor on the basis of the codes according to key code in order to edit, typeset and print.

(5) Real-time dynamic editing module for wrongly written characters: receive the wrongly written characters information needed adjusted and edited, call editor module for wrongly written character font to edit real-time and dynamically on the wrong word in document.

5. Example Demonstration for Dynamically Generation of Wrongly Written Characters

Following is the demonstration through this system based on a wrongly written character font of “笔”:

(1) Select orthographic “笔” as the copy object in the edit module for wrongly written character font, change the structure of bamboo prefix to make it a wrongly written character “𥫱” through stroke unit edit, and save the word “𥫱” (as shown in picture 5).

Copyright ⓒ 2015 SERSC

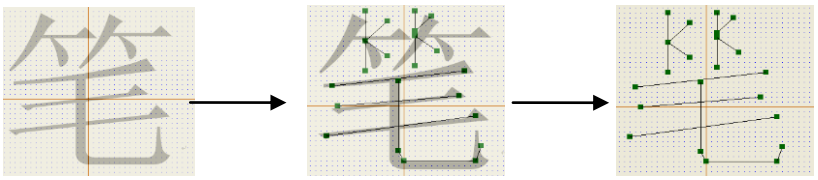


Figure 5. Editing Process of "毿" from the Orthographic "笔"

(2) A sequence composed of several two-dimensional tables (Xi, Yi) indicates every stroke units of "毿" through the extraction algorithm of feature points from feature extraction module (the values of the sequence support dynamic modification), several sequences and an index code corresponding to orthographic "笔" compose the feature code of wrongly written character "毿" (as shown in Figure 6).

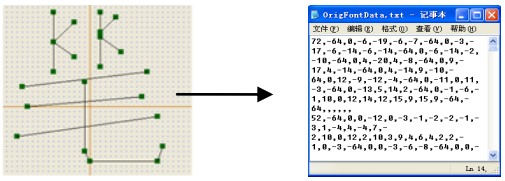


Figure6. Coding and Feature Extraction of Wrongly Written Character "毿"

(3) input the corresponding digital key or letter key of this wrongly written character (*i.e.* key codes) in edit environment "", and the wrongly written character will appear (as shown in Figure 7).

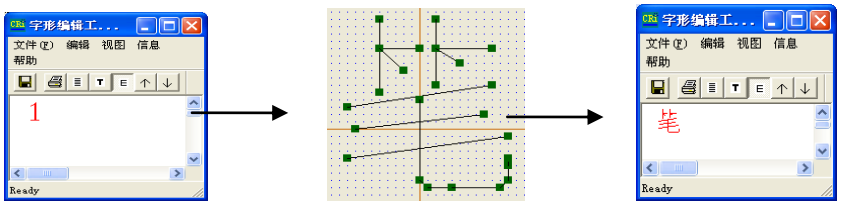


Figure 7. Process of Coding of "毿" using Number "1"

(4) when do dynamic edit to the wrongly written character, first input this wrongly written character through the key board, then right click this character to enter edit mode, modify the character according to the need, and the modified character will be added and stored into the character font list (as shown in picture 8).

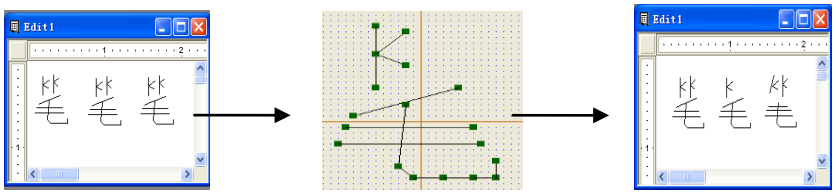


Figure 8. Process of Dynamic Editing of Wrongly Written Character "毿"

Copyright ⓒ 2015 SERSC

6. Conclusion

In view of the problem and the status of wrongly written Chinese characters input in printing and digital Chinese language teaching, this paper studies and designs a real-time and dynamic editing system based on wrongly written characters font for wrongly written characters input and processing; besides it makes full use of the characteristics of changeable structure and complex font of modern Chinese characters to combine the edit and modify of font library of wrongly written characters and Chinese characters copy, ensures dynamic production of various forms of wrongly written character font without changing the original font structure. This system provides a wrongly written character acquisition source for printing, typesetting and digital Chinese language teaching which becomes a simple, convenient and efficient input method of wrongly written characters.

References

[1] Q. M. Zhu, P. F. Li, X. Wu and X. X. Zhu, "Chinese Information Processing Techniques Course", Tsinghua University Press, Beijing, **(2005)**.
[2] X. Q. Li and M. Lin, "Design and implementation of wrongly written Chinese characters processing solution based on Unicode", Computer Engineering and Design, vol. 31, no. 10, **(2010)**, pp.2388-2391.
[3] X. Q. Li, "The Design and Implementation of Wrongly Written Chinese Characters Processing Toolkit Oriented Chinese Characters Teaching", Inner Mongolia Normal University, Hohhot, **(2010)**.
[4] M. Lin and R. Song, "A Stroke-Segment-Mesh (SSM) Glyph Description Method of Chinese Character", Journal of Computer Research and Development, vol. 47, no. 2, **(2010)**, pp.318-327.
[5] M. Lin and R. Song, "Pattern Computing-Oriented Formal Description of Chinese Character Glyph", Journal of Chinese information Processing, vol. 20, no. 3, **(2008)**, pp.115-123.
[6] M. Lin and R. Song, "Stroke-Segment-Mesh Depiction of Chinese Character Glyph and Algorithm for Glyph Comparing", Journal of Computer-Aided Design —Computer Graphics, vol. 21, no. 9, **(2009)**, pp. 1298-1306.
[7] Y. Wang, Y. Huang and F. Y. Zhang, "The Access Technique of True type Font Data on Windows Platform", Journal of Chinese Computer Systems, vol. 18, no. 11, **(1997)**, pp.75-81.
[8] J. Zheng, "Chinese Inputting and Outputting Handling System Design and Implementation for Character Shape Analysis", Inner Mongolia Normal University, Hohhot, **(2009)**.
[9] D. M. Han, "The Study of Chinese Glyph Description Technique", Inner Mongolia Normal University, Hohhot, Hohhot, **(2007)**.
[10] Q. X. Wu and Q. S. Li, "Research of Chinese Character Auto-generation Technology Based on the Dynamic Description Library", Science Technology and Engineering, vol. 8, no. 4, **(2012)**, pp.28-33.
[11] Q. S. Li, Q. X. Wu and Y. X. Yang, "Dynamic Description Library for Jiaguwen Characters and the Research of the Characters Processing", Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 49, no. 1, **(2013)**, pp. 61-67.

Authors

Li Xiao, received her Master's degree in Zhongnan University of Economics and Law in Wuhan, China and currently teaches in the school of computer and information engineering, Anyang Normal University as a lecturer. Her research interest is mainly in the area of security and privacy in wireless sensor networks and she has published several research papers in scholarly journals and international conferences in the above research areas.

Copyright ⓒ 2015 SERSC

Li Qingsheng, received his B.S. degree in Henan Normal University in Xinxiang, Henan, China and his M.S. degrees in Zhengzhou University in Henan, China and Ph.D. degrees in Wuhan University of Technology in Wuhan, China. He visited and exchanged in Bradford University in Britain between 2013 to 2014, and was invited to become an honorary professor of Bradford University. He is currently a professor and master Supervisor in Anyang Normal University, China, and he hold a concurrent post in China Computer Society Chinese Information Technology special Committee member and China Computer Society work Committee member. His research interests include multimedia network, video computing, DNA calculation, and cognitive computing in multimedia intelligent information processing, and he has published over 40 research papers in scholarly journals and international conferences, chaired and completed the national and provincial projects include NSFC more than 10, and has granted several patents in China.

 Copyright ⓒ 2015 SERSC