Structure-Based Web Access Method for Ancient Chinese Characters

Xiaoqing Lu[1], Yingmin Tang[1], Zhi Tang[1], Yujun Gao[2,3], and Jianguo Zhang[2,4]

[1] Institute of Computer Science and Technology, Peking University, Beijing, 100871, China
[2] Beijing Founder Electronics Co., Ltd., Beijing, 100085, China
[3] Center for Chinese Font Design and Research, Beijing, 100871, China
[4] State Key Laboratory of Digital Publishing Technology
(Peking University Founder Group Co., Ltd.), 100871, Beijing, China
{lvxiaoqing,tangyingmin,tangzhi}@pku.edu.cn,
{gao_yujun,zjg}@founder.com

Abstract. How to preserve and make use of ancient Chinese characters is not only a mission to contemporary scientists but is also a technical challenge. This paper proposes a feasible solution to enable character collection, management, and access on the Internet. Its advantage lies in a unified representation for encoded and uncoded characters that provide a visual convenient and efficient retrieval method that does not require new users to have any prior knowledge about ancient Chinese characters. We also design a system suitable for describing the relationships between ancient Chinese characters and contemporary ones. As the implementation result, a website is established for public access to ancient Chinese characters.

Keywords: Ancient Characters, Digital Heritage, Web Access.

1 Background

Ancient Chinese Characters (ACCs) represent an important heritage of Chinese history, which contains rich cultural information and serves as a basis for contemporary research tracing the evolution of modern characters. However, the origin and development of Chinese characters (also referred to as Han characters, Han ideographs, or Hanzis) are not one-dimensional. We see increasing numbers of score marks left on cultural relics of the New Stone Age, as they are unearthed one after another (Fig.1). We come to understand that it has taken a long and complicated process to arrive at the Chinese characters in use today.

The ancient characters studied here date back to at least 3300 year-old oracle-bone inscriptions that have some correlation to modern characters. Researchers have collected more than 4500 different characters from oracle-bone inscriptions, many that are variations of the same character. Other characters such as those of ancient seals are confined in a limited space and lack context for systematic study. The largest number of relics is the newly unearthed Qin and Chu collection of bamboo slips that contain very large quantities of texts related to the Warring States Period.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 72–82, 2013.

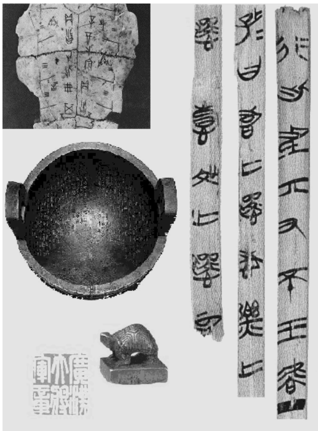

Fig. 1. Types of ACC: oracle-bone inscriptions, bronze inscription, ancient seal, bamboo slip

Despite the abundance of modern computer fonts, input methods, and word processing software, these tools do not suffice to duplicate the ancient characters. There are three principal reasons why it is difficult to decode ancient characters.

First, the research of ACCs involves very large quantities of modern characters. Although the number of ancient characters we have collected to date is limited, most of them represent sources for modern characters. Their relationships are complicated, including one-to-many, many-to-one, and many-to-many modes. To understand the exact meanings of ancient characters and their relationships with modern characters, we necessarily resort to a set of sufficient modern characters. However, the management of modern Chinese characters itself is a great challenge, as most of them are rarely-used. In 2012, Unicode 6.2 had totally encoded 75,215 Han characters [20], including seven main blocks of the Unicode Standard, as shown in Table 1. The term "CJK"—Chinese, Japanese, and Korean—is used in Unicode scripts to describe the languages that currently use Han ideographic characters.

Table 1. Han character encoded in Unicode 6.2

Block	Range	Comment
CJK Unified Ideographs	4E00–9FFF	common
Extension A	3400–4DBF	Rare
Extension B	20000–2A6DF	Rare, historic
Extension C	2A700–2B73F	Rare, historic
Extension D	2B740–2B81F	Uncommon, some in current use
Compatibility	F900–FAFF	Duplicates, unifiable variants, corporate characters
Compatibility Supplement	2F800–2FA1F	Unifiable variants

Lack of software code is a second problem in the research of ACCs. Today's information technology primarily focuses on modern characters, and provides little or no support for ancient characters. Software such as the GB code for China's mainland, the BIG5 code for Hong Kong and Taiwan, or Unicode for international practices, assigns a digital identity to each modern Chinese character so that each character is easily distinguished from another during processing of data streams. Because any coding system is limited by space requirements, none of the above systems is very useful in describing the entire character set of ACCs. The deep-rooted reason causing encoding difficulty is that the glyphs of ACCs vary in structure and stroke styles due to a lack of established rules, so that early ACCs have no fixed form, and one character generally has more than one shape. For instance, each of the characters of the oracle-bone inscriptions, in particular, proves to be precious due to their rarity. To further complicate matters, a single character has various forms (Fig 2). Preservation of the multiple styles used to depict characters adds to the difficulty in digitalizing Ancient Characters.

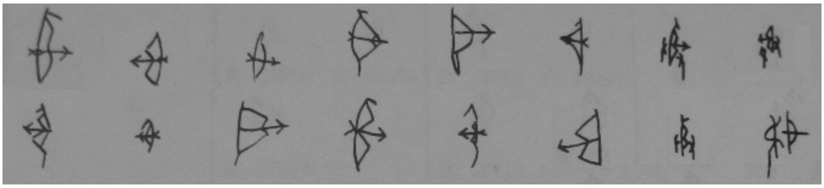

Fig. 2. An oracle-bone character "she (射)" represented by several different glyphs

Without reasonable codes, it is almost impossible to input ACCs directly into a computer, let alone support management and research requiring advanced IT technology. In fact, most contemporary research on ancient characters relies on ambiguous codes corresponding to modern characters.

Last, but not least, traditional IMEs (input method editors) do not have the capability to reproduce ACCs. These IMEs emphasize a high precision rate of character lookup by a short symbol sequence. Most of them require users to have some knowledge regarding a wanted character, such as its pronunciation, shape, or meaning. Most users will not be able to input an ACC using these IMEs, because users are not familiar with ACCs, or encoding schemes cannot guarantee the right relationship between an ACC and its counterparts in many cases. In contrast to IMEs, a practical ACC lookup service should provide users with a higher recall even for rarely used ACCs present in a very large list of candidates.

In recent years, computer technology has shown progress in applications for the study of ancient characters. In 1993, Xusheng Ji completed the electronic version "Index for Individual Characters of Bronze Inscription". In 1994, Ning Li[1] comprehensively presented some general principles for computational research of Chinese writing system. In 1996, Fangzheng Chen of the Institute of Chinese Studies, the Chinese University of Hong Kong, began the set up of a computer database for oracle-bone inscriptions, and carried out adjustment, classification, numbering, and merging of oracle bone inscriptions. Peirong Huang researched into and applied an ancient character font database. The "Statistics and analysis system for structures of Chinese characters" was established by Zaixing Zhang et al.[2], Che Wah Ho's ancient text database in Hong Kong and Derming Juang's Digital Library in Taiwan are all applicable for ancient characters classification. Zhiji Liu[3,4] conducted an investigation of the collation of glyphs of ancient writings. Minghu Jiang[5] presented a constructive method for word-base construction through syntax analysis of oracle-bone inscriptions. Derming Juang et al. [6] proposed an approach consisting of a glyph expression model, a glyph structure database, and supporting tools to resolve uncoded characters. Yi Zhuang et al. [7] proposed an interactive partial-distance map (PDM) - based high-dimensional indexing scheme to speed up the retrieval performance of large Chinese calligraphic character databases. James S. Kirk et al. [8] used self-organizing map methods to address the problem of identifying an unknown Chinese character by its visual features. Furthermore, to input ACCs by handwriting recognition is also feasible. Dan Chen et al. [9] proposed a method for on-line character recognition based on the analysis of ancient character features.

However, there is yet to be a management and search system for ancient characters open for public use in a network environment. Hence the ancient characters system proposed in this article intends to meet the requirements as follows: Design a digital resource pool of ancient characters for network applications; Search for an ancient character form corresponding to a modern character; Search for rare characters such as those beyond the scope of GBK code or even those without a correlative modern character; Search through multiple channels, by font, Unicode, phonetic, or other information.

On the above basis, we can build an academic exchange platform on the Internet that overcomes retrieval time and limited space issues and provides more extensive network services to high-profile designers, scholars studying Chinese heritage, philology research fellows, and amateurs.

2 Formalization of Relationships between ACCs and Modern Characters

To systematically manage ancient characters and provide a network service, we must clearly define and reasonably describe character classification. The latest computer technology can be employed to achieve the above-mentioned objective.

Ancient characters are divided into three categories:

Z1: Recognized characters
This refers to characters that have been studied and interpreted, and are recognized by the academic community. We can find the corresponding relationships of most of these characters with their contemporary Chinese characters. Therefore, contemporary Chinese characters can be used as an index to retrieve the glyphs of corresponding ancient characters.

It must be pointed out that quite a number of recognized glyphs are polysemous characters. In other words, the character pattern, structure, stroke, and shape of the characters are not completely the same, so they might represent different meanings that generally reflect variations of time and location such as different eras and countries.

Z2: Ambiguous characters
This refers to the characters that are provided with multi-conclusions from textual research and are not recognized unanimously by the academic community.

The index of ambiguous characters should be strongly compatible, that is, these characters should be searchable based on different information obtained from textual research. Therefore, when choosing the representative words for ambiguous characters, we must identify and distinguish them in terms of character pattern, usage, and context.

Z3: Unrecognized characters
This refers to characters that have not been defined through textual research. Such ancient characters are numerous, and have no identified correlation with contemporary Chinese characters. Therefore, special codes or symbols are necessary for indexing purposes.

As a result, we briefly state the following definitions:

$$A = \{a_1, a_2, \ldots a_n\} \tag{1}$$

A refers to the collection of existing encoded Chinese characters, a_i refers to a certain Chinese character, and i is the total number of encoded records, $i = 1,2, \ldots n$.

$$B = \{b_1, b_2, \ldots b_m\} \tag{2}$$

B refers to the collection of marks for uncoded Chinese characters, b_j refers to a certain mark, and j is the total number of uncoded records, $j = 1,2, \ldots m$.

The ACCs can be divided into two parts X and Y.

$$X = X_1 \bigcup X_2 \ldots \bigcup X_n \tag{3}$$

X refers to the collection of ACCs bearing corresponding relationships with contemporary encoded characters, where,

$$X_i = \{x_1, x_2, \ldots x_p\}. \tag{4}$$

X_i refers to an ACC set corresponding to a certain contemporary character. $x_k, k = 1,2, \ldots p$ refers to a certain ACC that mainly belongs to recognized characters or ambiguous characters $\langle x_k \in Z_1 | x_k \in Z_2 \rangle$

$$Y = Y_1 \bigcup Y_2 \ldots \bigcup Y_m \tag{5}$$

Y refers to the collection of ACCs bearing no corresponding relationships with the contemporary encoded characters, where,

$$Y_j = \{y_1, y_2, \ldots y_q\}. \tag{6}$$

y_l, $l = 1,2, \ldots q$ refers to a certain ACC that mainly belongs to one unrecognized character($y_l \in Z_3$). Y_j refers to the collection of unrecognized characters.

All ACCs that can be collected and sorted out are expressed by $Z = X \bigcup Y$.

The primary information expected to be used in the ancient character system is the collection of existing encoded Chinese characters and their corresponding ACCs, expressed by,

$$U = \{(a_1, X_1), (a_2, X_2), \ldots (a_n, X_n)\}. \tag{7}$$

As for the uncoded ACCs, the corresponding relationships can be fulfilled by borrowing uncoded Chinese character marks or self-defined codes, so they can be processed together with encoded Chinese characters. This relation can be described as follows:

$$V = \{(b_1, Y_1), (b_2, Y_2), \ldots (b_m, Y_m)\}. \tag{8}$$

Based on this model, the key to the follow-up processing of ACCs is to establish the information base that can store the U and V collections, and simultaneously provide the correct search method based on contemporary Chinese characters a_i or mark b_j.

3 Establishment of Super Large Font

As accessing ACCs relies heavily on sufficient modern characters, we need to establish a super large font to depict modern characters. However, the traditional process of font design is time-consuming and costly, including but not limited to creating basic strokes with the new style, composing radicals, and constructing characters. To speed up font creation, various innovative technologies have been developed to allow creation of new characters based on sample characters [21-26].

We have also focused on the automatic generation of Chinese characters for many years and proposed several methods [27-30]. Take the problem of deformation of stroke thickness and serif for example, as shown in Fig. 3; we adopt a distortionless resizing method for composing Chinese characters based on their components. By using a transformation sequence generating algorithm and a stroke operation algorithm, this method can generate the target glyph by an optimized scaling transformation.

(a) (b)

Fig. 3. Typical problems in recomposing Chinese characters. (a) Adjustment of radicals; (b) Resizing of strokes.

To establish reasonable relationships between ACCs and modern characters, an intensive analysis of their structures is necessary. First, a set of rules regarding glyph structure decomposition is defined. Next, the hierarchical relationship of strokes and radicals is represented by a framework. Generally speaking, most radicals are basic components that will not be decomposed. However, some radicals are compound components, and contain multiple basic components and possibly additional strokes. Consequently, the structural decomposition of a glyph may not be limited to only one possible decomposition. To provide users with more convenience, the redundant expressions of glyph structures are permitted in our system. Furthermore, an algorithm is designed to classify the characters by their multi-level radicals and to calculate the number of corresponding strokes.

4 ACC Database

Based on the in-depth and comprehensive organization of Chinese characters, particularly by considering the varied information on ancient characters, the ACC database is effectively designed.

4.1 Relation Schema

Management of ACCs should integrate the code and related information, so we define the main relation schema in Table 2.

Table 2. Relation schema of ACC database (ACC_RS)

Item	Meaning
Unicode	Contemporary Chinese character Unicode for this ancient character.
Dynasty	Dynasty when this ancient character was used.
Type	Type of this ancient character (e.g. pictographic characters, ideograph, and phonogram)
Classification	Class type of this ancient character (e.g. inscriptions on bones or tortoise shells of the Shang Dynasty, inscriptions on bronze, seal character, etc.)
Place	Contemporary place where this ancient character was unearthed.
Carrier	Carrier of this ancient character (e.g. the name or the number of a certain bronze implement)
Country	Ancient country where this ancient character was used.
SubbaseID	Number of the font database storing this ancient character.
SubID	Code of the ancient character, used in sub-font database.
Filename	File name for the picture of this ancient character.
ID	The unique ID of this ancient character in the font database.

Other relation schemas we used include: Dynasty and Country (DC_RS), Ancient C_Character Classification (ACCC_RS), ACC Type (ACCT_RS), Unicode and Glyph (UG_RS), Radical and Component (RC_RS), Ancient Image (AI_RS), Contemporary Image (CI_RS).

To edit, sort, and manage the information of the ancient characters effectively, all tables are organized properly, and their relationships are shown in Fig. 4.

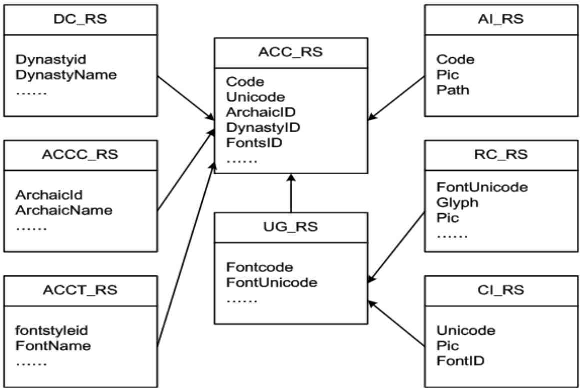

Fig. 4. Relationships of the data tables

4.2 Query and Browse Method

As Fig. 5 shows, a special engine, glyph tree is used to show characters not present in GBK code.

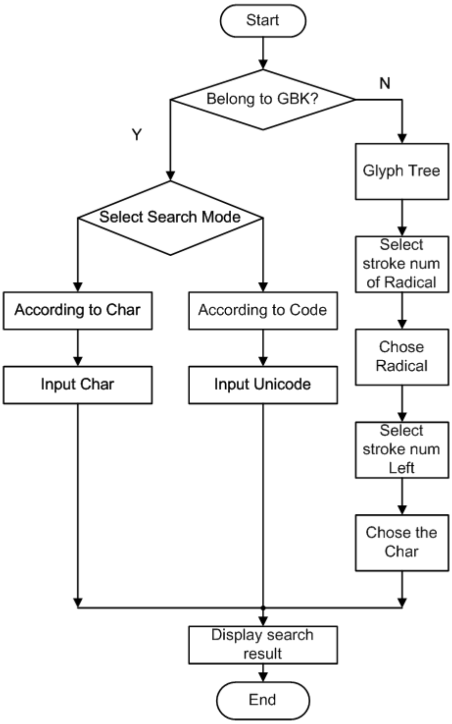

Fig. 5. Flow chart of the search process

Based on the corresponding relationships between ACCs and contemporary characters, the retrieval system consists of two categories, including search of encoded Chinese characters and search of uncoded characters. The encoded Chinese characters, such as within GBK, can be input by common IMEs, while the rare characters and unrecognized characters can be searched by interactive query methods with special glyphs provided by our system.

5 Implementation and Results

Several technologies are adopted to achieve high extensibility, scalability, and maintainability. The development of the software system, collecting, editing, and processing the information of the ACCs took many years to combine into a comprehensive system. The search function is now available, and users can look up the glyphs of old Chinese characters from our website (http://efont.foundertype.com/AgentModel/FontOldQ.aspx). Fig. 6 shows the search results for the Chinese character Ma (马), yielding a number of possible ACCs related to it.

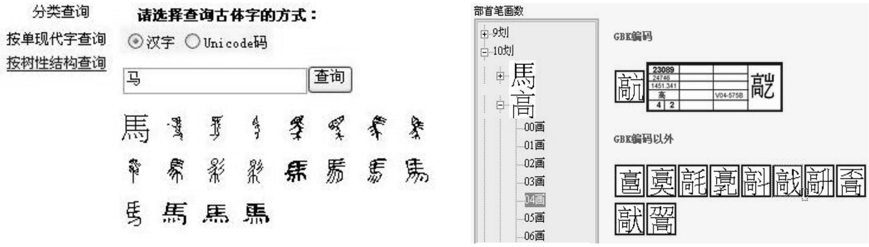

Fig. 6. The search results for the character Ma (马).

6 Further Research

In terms of the ACC system, the most urgent issues so far are how to present the information of ACCs that have lost connection with contemporary Chinese characters (the V collection previously mentioned). As this category of ACCs cannot be backed up by the corresponding contemporary characters, they are rarely displayed in the computer system.

Furthermore, to benefit more people and increase academic interaction, the platform needs to be accessed by more users, experts, and scholars. Any newly discovered ancient characters or useful information can be easily added to the platform, and we can exchange ideas on the source, authenticity, identification, and interpretation of these characters.

With the basic information provided on ancient characters, the public can use the system to make an in-depth study and analysis on the evolution of ancient characters and their connection to character patterns, thus actively enhancing the cognation analysis of ACCs, radical classification and arrangement, as well as automatic analysis of the commonly confused words.

Acknowledgment. This work is supported by Beijing Natural Science Foundation (No. 4132033).

References

1. Li, N.: Computational Research of Chinese Writing System Han4-Zi4. Literary and Linguistic Computing 9(3), 225–234 (1994)
2. Zhang, Z.-X.: On Some Issues of the Establishment of Ancient Chinese Font. Journal of Chinese Information Processing 17(6), 60–66 (2003)
3. Liu, Z.-J.: Investigation into the Collation of Glyphs of Ancient Writings for Computer Processing. Applied Linguistics No 4, 120–123 (2004)
4. Liu, Z.-J.: Encoding Ancient Chinese Characters with Unicode and the Construction of Standard Digital Platform. Journal of Hangzhou Teachers College 29(6), 37–40 (2007)
5. Jiang, M.-H.: Construction on Word-base of Oracle-Bone Inscriptions and its Intelligent Repository. Computer Engineering and Applications 40(4), 45–48 (2004)
6. Juang, D., Wang, J.H., Lai, C.Y., Hsieh, C.C., Chien, L.H., Ho, J.M.: Resolving the Unencoded Character Problem for Chinese Digital Libraries. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2005, pp. 311–319. ACM, Denver (2005)
7. Zhuang, Y., Zhuang, Y.-T., Li, Q., Chen, L.: Interactive High-Dimensional Index for Large Chinese Calligraphic Character Databases. ACM Transactions on Asian Language Information Processing 6(2), 8-es (2007)
8. Kirk, J.S.: Chinese Character Identification by Visual Features Using Self-Organizing Map Sets and Relevance Feedback. In: IEEE International Joint Conference on Neural Networks, pp. 3216–3221 (2008)
9. Chen, D., Li, N., Li, L.: Online recognition of ancient characters. Journal of Beijing Institute of Machinery 23(4), 32–37 (2008)
10. Allen, J.D., Becker, J., et al.: The Unicode Consortium. The Unicode Standard, Version 5.0. Addison-Wesley, Boston (2006)
11. Zhuang, Y.-T., Zhang, X.-F., Wu, J.-Q., Lu, X.-Q.: Retrieval of Chinese Calligraphic Character Image. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 17–24. Springer, Heidelberg (2004)
12. Bishop, T., Cook, R.: A Specification for CDL Character Description Language. In: Glyph and Typesetting Workshop, Kyoto, Japan (2003)
13. Lu, Q.: The Ideographic Composition Scheme and Its Applications in Chinese Text Processing. In: Proc. of the 18th International Unicode Conference, IUC-18 (2001)
14. Juang, D., Hsieh, C.-C., Lin, S.: On Resolving the Missing Character Problem for Full-text Database for Chinese Ancient Texts in Academia Sinica. In: The Second Cross-Strait Symposium on the Rectification of Ancient Texts, pp. 1–8, Beijing (1998)
15. Hsieh, C.-C.: On the Formalization and Search of Glyphs in Chinese Ancient Texts. In: Conference on Rare Book and Information Technology, pp. 1–6, Taipei (1997)
16. Hsieh, C.-C.: A Descriptive Method for Re-engineering Hanzi Information Interchange Codes-On Redesigning Hanzi Interchange Code Part 2. In: International Conference on Hanzi Character Code and Database, pp. 1–9, Kyoto (1996)
17. Hsieh, C.-C.: The Missing Character Problem in Electronic Ancient Texts. In: The First Conference on Chinese Etymology, Tianjin, pp. 1–8. Tianjin (1996)

18. Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access Method for Characters and Rectangles. In: Proceedings of ACM SIGMOD International Conference on Management of Data, ACM SIGMOD 1990, pp. 322–331. ACM, New York (1990)
19. Lin, J.-W., Lin, F.-S.: An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity. In: IEEE The 11th International Symposium on Communications & Information Technologies (ISCIT 2011), pp. 489–492 (2011)
20. The Unicode Standard The Unicode Consortium, version 6.2 (2012), `http://www.unicode.org/versions/Unicode6.2.0/`
21. Xu, S.-H., Jiang, H., Jin, T., Lau, F.C.M., Pan, Y.: Automatic Facsimile of Chinese Calligraphic Writings. Computer Graphics Forum 27(7), 1879–1886 (2008)
22. Xu, S.-H., Jiang, H., Jin, T., Lau, F.C.M., Pan, Y.: Automatic Generation of Chinese Calligraphic Writings with Style Imitation. IEEE Intelligent Systems 24(2), 44–53 (2009)
23. Lai, P.-K., Pong, M.-C., Yeung, D.-Y.: Chinese Glyph Generation Using Character Composition and Beauty Evaluation Metrics. In: International Conference on Computer Processing of Oriental Languages, ICCPOL 1995, Honolulu, Hawaii, pp. 92–99 (1995)
24. Lai, P.-K., Yeung, D.-Y., Pong, M.-C.: A Heuristic Search Approach to Chinese Glyph Generation Using Hierarchical Character Composition. Computer Processing of Oriental Languages 10(3), 307–323 (1996)
25. Wang, P.Y.C., Siu, C.H.: Designing Chinese Typeface using Components. In: Computer Software and Applications Conference, pp. 412–421 (1995)
26. Feng, W.-R., Jin, L.-W.: Hierarchical Chinese character database based on radical reuse. Computer Applications 26(3), 714–716 (2006)
27. Lu, X.-Q.: R&D of Super Font and Related Technologies. In: The Twenty-second International Unicode Conference, IUC22, San Jose, California, September 9–13 (2002), `http://www.unicode.org/iuc/iuc22/a310.html`
28. Tang, Y.-M., Zhang, Y.-X., Lu, X.-Q.: A TrueType Font Compression Method Based on the Structure of Chinese Characters. Microelectronics & Computer 24(06), 52–55 (2007)
29. Sun, H., Tang, Y.-M., Lian, Z.-H., Xiao, J.-G.: Research on Distortionless Resizing Method for Components of Chinese Characters. Application Research of Computers 30 (2013), `http://www.cnki.net/kcms/detail/51.1196.TP.20130603.1459.008.html`
30. Shi, C., Xiao, J., Jia, W., Xu, C.: Automatic Generation of Chinese Character Based on Human Vision and Prior Knowledge of Calligraphy. In: Zhou, M., Zhou, G., Zhao, D., Liu, Q., Zou, L. (eds.) NLPCC 2012. CCIS, vol. 333, pp. 23–33. Springer, Heidelberg (2012)