International Journal of Knowledge and Language Processing
Volume 3, Number 2, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp. 39-**42**

A Survey on Digital Description of Chinese Character Glyph

Xue Yongzeng and Gu Yingying
Department of New Media and Art
Harbin Institute of Technology
P. O. Box 772, No. 92 West Dazhi Street
Harbin, Heilongjiang, 150001, China
{Yongzeng.xue, Yingying.Gu}@gmail.com

Received March 2012; revised April 2012

ABSTRACT. *Description of Chinese character glyph plays an important role in input, encoding, recognition, display and typography of Chinese characters on digital devices. The structure of Chinese character is more complex than that of alphabetic writing systems, which makes its digitalization more difficult. This paper reviews the studies on digital description of Chinese character glyph over the past years and concludes with the advantages and disadvantages of the methods involved. In general, the proposed methods can be divided into two categories: one describes the glyph with fixed offset of stokes or parts; the other describes it with flexible positions. The former relates to applied research while the latter goes closer to ontology of Chinese characters. This paper makes a comparison on different methods with respect to the objectives of various applications and gives discussions on a few promising directions.*

Keywords: Natural Language Processing; Character Description; Chinese Character Glyph

1. **Introduction.** Description of Chinese character glyph plays an important role in input, encoding, recognition, display and typography of Chinese characters on digital devices. The Unicode standard describes Han ideographic characters with a three-dimensional conceptual model, the dimensions of which represent semantic, abstract shape and actual shape of a character, respectively. The semantic dimension describes unique meaning and usage; the abstract shape dimension describes a unique form of a character by a certain semantic attribute; the actual shape dimension describes a unique type design of an abstract shape with a semantic attribute. Therefore, the Unicode standard employs a two-level classification on Chinese characters – abstract shape and actual shape.

In the abstract shape description, the ideograph of a character is analyzed into a combination of components, and then each component is analyzed into inferior components, until all components are analyzed into primitive elements. Thus an ideograph defines an

component tree with the root node the ideograph itself and each leaf node a primitive element. For the sake of abstractness, the positions of components are relative, described with partial positioning. (e.g. left half, top half, etc.)
In the actual shapes of ideographs, those with the same abstract shape, except some unique characters, are unified.
In general, the proposed methods so far can be divided into two categories: one describes the glyph with fixed offset of stokes or parts (absolute positioning); the other describes it with flexible positions (relative positioning). Section 2 and 3 will discuss the two cases respectively. Section 4 will conclude the paper.

2. **Relative Positioning.** Relative positioning refers to the methods which decompose an ideograph of a Chinese character into components with a description by flexible offset of one component from or to each other.
Since graphemic representation of the same Chinese character may differ, many efforts were made to standardize Chinese characters, i.e. distinguish standard glyph from variants. [1] proposed several principles of whether glyphs are the same or not. The motivation of structural theory of Chinese character was discussed on inheritance chains and variations of different structural levels. [2-4] discussed basis and methodology of decomposing Chinese characters into components. Their work results in *Modern Chinese Characters Components List*, consisting of 290 basis components and 94 complex components. [5] discussed the normalization of printed Songti and then pointed out the shortages of current systems and proposed amending methods. [6] also discussed the difficulties in Chinese character component standard of GB13000.1, including the large amount of basic components, the vague definition of components, the confusing rules of definition and the poor rules of composition and decomposition of components.
By defining the component of Chinese character as "the basic unit for constructing a Chinese character, which is either detachable, relatively independent block of strokes or basic stroke", [7] proposed 7 types of relation between strokes. [8] introduced context free grammar (CFG) into description of the structure of compound Chinese characters, and divided the structures into 11 patterns. Figure 1 illustrates the decomposition of "蒜" and the corresponding CFG.

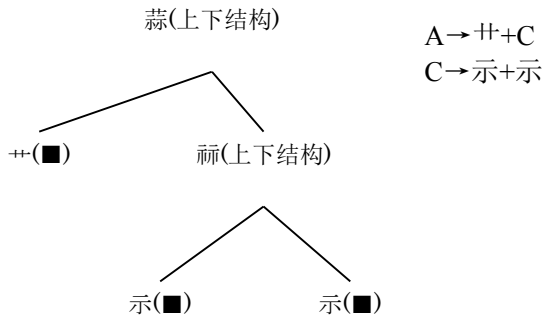

FIGURE 1. Decomposition of "蒜" with CFG.

[9] described a method which decomposed a Chinese character into a one-dimensional (1D) string of strokes with the relationship between consecutive strokes. In that method, the relation of consecutive strokes was divided into classes and combined by different patterns in generation. [10] improved this method by a second-order model, in which the first-order relation was used to build radicals and the second-order to verify their correctness. This model compensates for the lack of distinguish character with similar stroke structures except for the length of one or more strokes, such as "由" and "田". [11] built a hierarchical character database of only basic strokes. Other components and characters were generated through affine transformation, by which the position, shape and size of the same basic stroke differ, thus the radicals can be reused and the size of the database is limited.

3. **Absolute Positioning.** Absolute positioning, on the contrary, refers to the methods which decompose an ideograph of a Chinese character into components with a description by fixed offset of one component from or to each other.

A typical absolute positioning system is Wenlin Institute's Character Description Language (CDL) [12], which is designed to encode CJK characters. CDL itself is encoded in a XML style, with two most important keywords: stroke and comp (abbr. of component). Each "stroke" or "comp" is viewed as a rectangle in shape and has an attribute of "points" to figure the coordinates of the points. Therefore, a CDL description of "蒜" may look like Figure 2.

```
<cdl char='蒜'>
    <comp char='艹' points='0,0,32,8'     />
    <comp char='示' points='0,8,10,33'    />
    <comp char='示' points='10,8,32,33' />
</cdl>
```

FIGURE 2. Decomposition of "蒜" with CDL.

There are many approaches to transform a relatively positioned ideograph of a Chinese character into an absolutely positioned one. Similar to the above-mentioned methods of relative positioning, [13] proposed the HanGlyph language that describes the Han glyph with strokes and operators/relations. Meanwhile, a Chinese Character Synthesis System (CCSS) was developed to translate HanGlyph expressions into METAPOST figures. [14,15] also applied mathematical expressions to composition of Chinese characters. In their method, components are viewed as operands and relations as operators. As the process of calculating in that way, the figure representation of Chinese character is generated gradually.

4. **Conclusions.** This paper discusses present digital methods of describing Chinese Character Glyphs and divides the contemporary methods into two types: absolute positioning which decomposes an ideograph of a Chinese character into components with a

description by fixed offset of one component from or to each other, and relative positioning which decomposes it by flexible offsets. The former relates to applied research while the latter goes closer to ontology of Chinese characters. The paper also introduces the work on structural theory of Chinese character and normalization of components and their structures.

Acknowledgment. This work is partially supported by 2010 Art Planning Fund of Heilongjiang Province (No. 10D002). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] N. Wang, Research on construction of computer database of ancient characters and theory of Chinese characters, *Applied Linguistics*, no.9, pp.54-59, 1994 (in Chinese).

[2] J. Fei, modern Chinese characters components, *Applied Linguistics*, no.2, pp.20-27, 1996 (in Chinese).

[3] N. Wang, Basis of Chinese characters and decomposition of modern Chinese characters, *Language Planning*, no.3, pp.4-9, 1997 (in Chinese).

[4] P. Su, decomposition of Chinese components, *Language Planning*, no.3, pp.10-13, 1997 (in Chinese).

[5] J. Fei and L. Xu, Research of standardization of printed Songti of normative Chinese characters, *Applied Linguistics*, no.3, pp.66-74, 2003 (in Chinese).

[6] X. Zhang, Difficulties in the application of "Chinese character component standard of GB 13000.1 character set for information processing" for Chinese character input, *Journal of Chinese Information Processing*, vol.18, no.4, pp.60-65, 2004 (in Chinese).

[7] D. Pan and Z. Zhan, Research on components of Chinese character, *Chinese Information Processing*, no.3, pp.46-48, 1995 (in Chinese).

[8] Z. Feng, Description of Chinese character structure by context free grammar, *Linguistic Sciences*, vol.5, no.3, pp.14-23, 2006 (in Chinese).

[9] C. Hsieh and H. Lee, Off-line recognition of handwritten Chinese characters by on-line model-guided matching, *Pattern Recognition*, vol.25, no.11, pp.1337-1352, 1992.

[10] H. Lee and H. Hsu, A hierarchical model-guide generation of Chinese characters, *IEEE Proceedings of ICPR'94*, Jerusalem, Israel, pp.256-260, 1994.

[11] W. Feng and L. Lin, Hierarchical Chinese character database based on radical reuse, Computer Applications, vol.26, no.3, pp.714-716, 2006 (in Chinese).

[12] T. Bishop and R. Cook, Wenlin CDL: character description language, *Multilingual*, Vol.18, no.7, pp.62-70, 2007.

[13] C. Yiu and W. Wong, Chinese character synthesis using METAPOST, Proceedings of the 24th Annual Meeting and Conference of the TeX User Group, Hawaii, USA, pp.1001-1009, 2003.

[14] W. Zhang, X. Sun, Z. Zeng, J. Wu, Automatic generation of mathematical expression of Chinese characters, *Journal of Computer Research and Development*, vol.41, no.5, pp.848-852, 2004 (in Chinese).

[15] X. Sun, J. Yin, H. Chen, Q. Wu, X. Jing, On mathematical expression of a Chinese character, *Journal of Computer Research and Development*, vol.39, no.6, pp.707-711, 2002 (in Chinese).