Stroke Feature Database Construction for Chinese Calligraphy

Chung-Shing Wang
Department of Industrial Design, Tung-Hai University, Taiwan

Ching-Fu Chen
Department of Industrial Design, Tung-Hai University, Taiwan

Chung-Chuan Wang
Department of Multimedia and Game Science, Chung-Chou University, Chung-Hwa, Taiwan

Teng-Ruey Chang
Department of Industrial Engineering and Management, Nan-Kai University of Technology, Taiwan

Abstract

This research aimed at building a platform for Chinese calligraphy handwriting and carried out geometric modeling and analysis to set up a stroke feature database model using Microsoft Access. This has allowed the creation of an independent Chinese character simulation platform for extracting and analyzing Chinese strokes. Users can input through the use of a mouse or digital pad to draw out the strokes of a Chinese character. The constructed system automatically recognize the stroke starting coordinates, angles, area ratio and other parameters for stroke feature analysis and extraction by using Delphi software programming. These data can then be compared with stroke feature information in the database to produce a simulated calligraphy font. This platform also allows the user to output the font simulation result to 3D drawing software such as SolidWorks for further design and applications. The advantage of this research is to develop a calligraphy font simulation system with a user friendly interface to allow traditional Chinese beginners having the opportunity to write out a Chinese calligraphy font on computer. This will in turn serve to stimulate interests in Chinese calligraphy learning. Furthermore, this research enables users to output the results to a computer aided design field for further application whilst promoting and spreading the art of Chinese culture.

Keywords Geometric modeling, Chinese calligraphy, Stroke recognition, Feature extraction, Stroke database.

1. Introduction

The Chinese calligraphy is a very unique and important asset in Chinese culture and its high artistic value is praised by people all over the world. Although currently there are many different kinds of Chinese fonts available for selection for the computer system, the ability for the user to provide modification to the font itself is very limited. For example: we cannot offer a calligraphy font style that caters to personal handwriting habits to create new or composite words for processing and application. This research uses computer geometry modeling to analyze and extract characteristic strokes to produce a handwritten Chinese calligraphy font in order to build a calligraphy character platform. This offers the users an opportunity to further study and understand the writing structure of Chinese calligraphy.

In this study, we analyzed the Kai-font in Microsoft True Type fonts. Through analysis of the characteristic strokes of the Kai-font and carrying out simultaneous processing of strokes from non-traditional writings, we can remove, process and summarize duplicated strokes to create a stroke characteristic database [Chung & Tseng, 1995]. After that, by using the interface for the system platform, users can input handwritten font stokes with input devices such as the mouse or digit pad. The system will then automatically retrieve the corresponding strokes to carry out "Thinning" and "Stroke feature extraction" process for comparison with the characteristic stroke database to find the corresponding stoke to complete font output. The simulated results are then outputted as a JPG file. These files are automatically exported into 3D computer aided design software, SolidWorks for further processing and application.

2. Literature Review

Stroke characteristics are an important feature unique to Chinese calligraphy. Each individual stroke is composed of points, lines and curves. The point corresponds to the black spot; line refers to the linear movement of the tip of the pen and the curve is characterized by curve motion of the pen [Hornby, 1972; Stallings, 1976, 1977]. Strokes can be defined using operation points and parameters. Operation points may include the origin, end point, length, direction… etc. of the stroke. Parameters are then used to define the shape, width, position… etc. for the stroke [Wong & Hsu, 1995]. However, Chinese calligraphy font strokes have a huge amount of variations which result in an extremely complex way for describing strokes. Even for similar a stroke structure, a small change in the shape, angle, position length or proportion can result in a completely different character as shown in Figure 1.

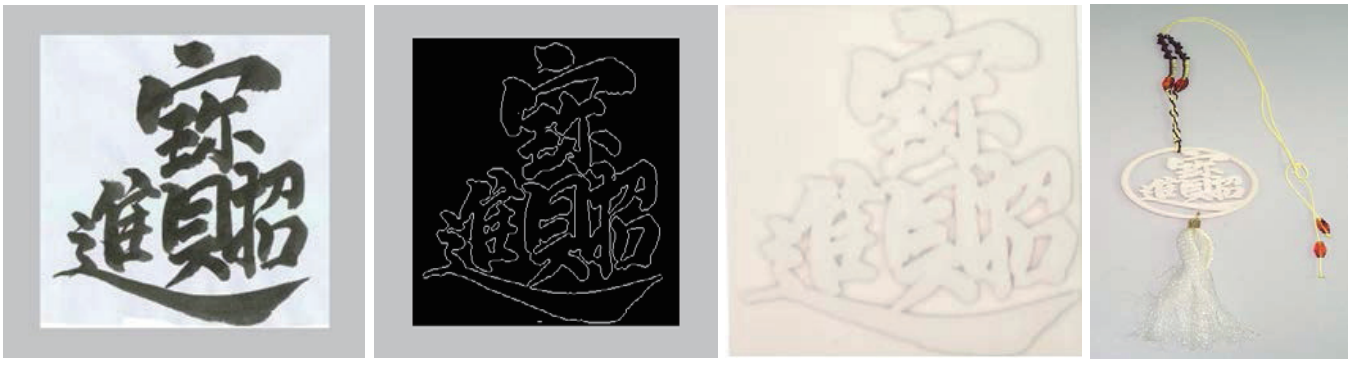
Chinese Strokes	Examples
Different shape	子子孑 戍戍戎 王壬 千干
Different location	玉玊主 犬太 八入
Different length	田由甲申 士土 己已巳 天夫
Different scale	日曰

Figure 1 Various Chinese characters

Calligraphy font recognition, also known as reverse recognition, aims to achieve an effective way to obtain

the point cloud data for the fonts. Lin et al., (2001) used an automatic threshold value setting to identify fonts and create a common words database. The search function of the database was then used to identify the fonts in the document. Detection and identification of common characters allows recognition of handwritten or printed fonts. Romero et al., (1997) used artificial neural network algorithm to enhance the identification rate for similar words along with the use of different search directions to modify the neural parameters to improve recognition speed. Lin et al., (1996) used a "Trend-followed" translation technology to find the contour segment and overall characteristic information for the font. Chuang et al., (1995) used heuristic algorithm to identify printed fonts while utilizing contour features to increase recognition rate. Wu et al., (2013) proposed a mobile Chinese calligraphic training systems using virtual reality technology. A decomposition database for stroke splitting and feature extraction for Chinese characters generated is proposed by Yang et al., (2013).

Previous research by the authors used image processing and reverse engineering (RE) in Chinese calligraphy [Wang, et al., 2006]. This research combined processes of RE, grey prediction theory in pattern processing, geometric modeling for constructing Chinese calligraphy characters and rapid prototyping (RP) for model making. First of all, the written Chinese calligraphy was scanned by scanner. Next, the contours of the Chinese characters were detected with image processing. These contours were then converted to point data, which can be easily processed in any CAD software by using B-Spline curves to fit the points. An example was illustrated below using a Chinese compound word "the fortune and treasure are coming" to show the steps of the process. Finally, a RP model was constructed to show various applications for the products in Chinese calligraphy pattern. With this research, a collaboration between the Chinese calligraphy in handcraft and digital virtual design can be realised; but more importantly, the aesthetic aspects in the characters can be preserved (Figure 2).



(a) Image file of compound character

(b) Grey prediction contour detection

(c) Compound character in a paper weight

(d) A hanging decoration

Figure 2 Reverse design process for a Chinese compound character

3. Methodologies

In this research, we developed a simulation platform for Chinese calligraphy so the user can quickly create simulated calligraphy fonts for further application. The research framework consisted of the following three components:

(1) Stroke database construction
(2) Stroke feature recognition and coding.
(3) Creating a simulation platform and user interface for writing stroke input.

3.1 Stroke database construction

In order to create a client based calligraphy simulation system, we first need to generate a corresponding stroke database. We used the existing built-in Kai-font within Microsoft Windows operating system as basis to obtain stroke information. Kai-fonts are TrueType vector fonts with characters constructed from individual strokes and we can extract its font structure data using Microsoft Development Network (MSDN). With TrueType vector fonts, font data structures are arranged as such that one character can be composed of many stroke contours and every contour can be composed of many curves. Statistics show that the Kai-font has 23,230 characters and 178,196 contour data.

Using the character 「永」 as an example, the creation of the character with Kai-font is separated into six stroke contours, as shown in Figure 3. The meanings of each original stroke contour data are represented below. The mathematical relationship between ❶❷❸❹❺ is:

$$❶ = 16 + (❷ + ❸) * 4 + (❹ + ❺) * 8$$

❶ Size of the contour in the font file in bytes; ❷ Number of straight lines; ❸ Number of curves; ❹ Number of points used by the straight line; ❺ Number of points used by the curve; ❻ Angle of xy; ❼ The actual Chinese character ; ❽ What number stroke it is in the Chinese character. Consider the first stroke in the character 「永」, the "point" stroke, the contour has 0 straight lines, 8 curves, 16 points used by the curves (2 points for 1 curve) and a 39.2 degree angle. The total size is 16 + (0+8)*4 + (0+16)*8 = 176 (Bytes), as shown in Table 1.

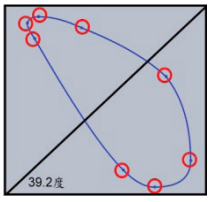


Figure 3 First stroke contour information of the character 「永」

Table 1 Database for the first stroke contour information of the character 「永」

❶	❷	❸	❹	❺	❻	❼	❽
176	0	8	0	16	392	永	1

If we input all the contours of the Kai-font into the contour database, a great amount of time is required for the contour recognition search process, which becomes highly inefficient. Therefore, duplicate contours were removed to reduce number of comparisons required in order to increase operational efficiency. The simplest method was to analyse common contours structures between different characters and removing those contour structures that were the same between them.

This research used the GetGlyphOutline function in

Win32 API to access contour data in selected characters. The following stroke contour data were then analyzed to determine which contours need to be deleted.

1. Size occupied

Because Kai-font characters were created from individual strokes, the strokes with same structures take up the same amount of data space. For example, the first stroke in 「永」 and「汁」 and the second stroke in 「汀」both have the「丶」structure and occupied 176 bytes. However, apart from the size, we also need to consider other attributes such as the angle between xy, to decide whether to remove the selected stroke contour as illustrated in Figure 4.

Figure 4 Stroke contours with same data size

2. Number of lines, curves and points

Each individual stroke may contain many lines or curves and we can distinguish between them by the number of lines, curves and points. For example: the first stroke of 「刀」、「刁」、「力」 all contain the same number of lines, curves and points; Therefore, they all have the same structure (Figure 5).

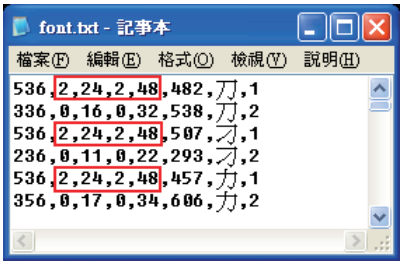


Figure 5 Strokes with same of lines, curves and points

3. Angle of xy

We can determine the structure of the stroke contour by looking at the diagonal angle formed by a box containing the stroke. As shown in Figure 6, the first stroke of the character 「九」 has an angle of 35.2 degrees and the second stroke 62.3 degrees. Calligraphy fonts with same strokes but with different stroke contour height and widths can be distinguished this way.

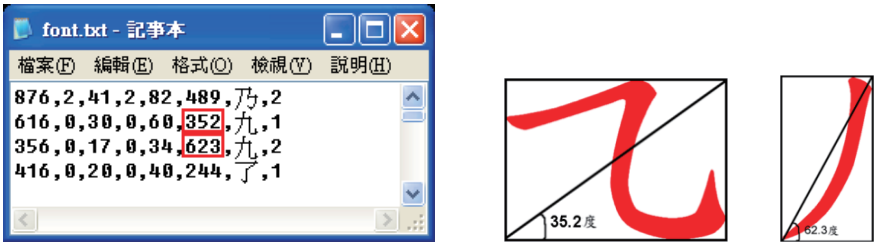


Figure 6 xy angle of the character 「九」

Based on the three criteria mentioned above, same stroke contours can be deleted from the extracted font file. If the three values were the same between different strokes, the one with the largest dimensions was

kept because smaller stroke contours were more likely to have errors when carrying out stroke feature analysis. After analysis and deletion, a total of 10,218 stroke contour data remained. This contour characteristic database was then stored in Microsoft Access as shown in Table 2.

Table 2 Contour coding of contour database

details : 資料表

識別碼	bufsize	allsum	linecount	curvecount	linesize	curvesize	angle	font	token2
1	316	316015030105	0	15	0	30	105	一	211220021200001
2	736	736036072451	0	36	0	72	451	乙	221410021200031
3	316	316015030088	0	15	0	30	88	畫	211220021200001
4	396	396019038736	0	19	0	38	736	何	121320001200012
5	236	236011022131	0	11	0	22	131	七	211220002100001
6	396	396019038533	0	19	0	38	533	七	221410000100001
7	276	276013026588	0	13	0	26	588	乃	222320000100001
8	876	876241282489	2	41	2	82	489	乃	221312112012420
9	616	616030060352	0	30	0	60	352	九	221210021200131
10	356	356017034623	0	17	0	34	623	九	121320000100001
11	416	416020040244	0	20	0	40	244	了	211210002100012
12	368	368117134716	1	17	1	34	716	了	121320001200012
13	316	316015030092	0	15	0	30	92	冊	211220021200001
14	336	336016032561	0	16	0	32	561	人	222320000100001
15	276	276013026400	0	13	0	26	400	人	221410000100001
16	316	316015030635	0	15	0	30	635	犬	121320000100012
17	476	476023046587	0	23	0	46	587	儿	221410001200001
18	276	276013026490	0	13	0	26	490	入	222320000100001
19	296	296014028408	0	14	0	28	408	入	221410000100001
20	276	276013026465	0	13	0	26	465	脊	222320000100001
21	296	296014028421	0	14	0	28	421	八	221410000100001
22	336	336016032655	0	16	0	32	655	凡	121320000100012
23	576	576028056481	0	28	0	56	481	凡	221410001200011
24	536	536224248482	2	24	2	48	482	刀	221310021200012
25	336	336016032538	0	16	0	32	538	刀	222320000100001
26	536	536224248507	2	24	2	48	507	刁	221310021200012
27	236	236011022293	0	11	0	22	293	刁	211220001200001
28	536	536224248457	2	24	2	48	457	力	221410021200012
29	356	356017034606	0	17	0	34	606	力	121320000100001
30	236	236011022129	0	11	0	22	129	麼	211220021200001

記錄: 1 之 10218

3.2 Stroke feature recognition and coding

Stroke feature recognition and coding were the most important aspects of identification and analysis. Normally before feature extraction of writing strokes we needed to carry out a “thinning” (also known as skeletonizing) process. Thinning refers to the process of eliminating points from the outlines of shapes or fonts with different widths until only curves with one pixel width remains. This research used the “Peeling approach” with further corrections and modifications to achieve skinning of font strokes [Zhang & Suen, 1984]. Peeling method, also known as the “Iterative morphological method”, used the relationship between non-zero pixels (which represent the location of line data) and neighboring pixels to determine if it is the edge of the line. If it is, the value is removed and replaced by zero. Using this way similar to peeling method, we can sequentially reduce width from both sides of the line until the width is only one pixel but still retain the connection of the line.

This research proposed a stroke feature and coding method as basis for stroke feature recognition. Each stroke was encoded as a 15 bit unit to generate the stroke characteristic value. Coding in this research was divided into four parts as shown in Table 3. Font stroke distribution characteristic values, which were mainly used as a point of comparison with written stroke characteristic values, can be separated into four categories (Table 4).

4. A Simulation Platform for Chinese Calligraphy Characters

Current research aims to develop a simulation platform for Chinese calligraphy character recognition. Softwares

used for development are shown in Table 5.

Table 5 Software for a simulation platform

Items	Software
Operation system	Windows XP
Programming language	Borland Delphi 7.0
Strokes database	Microsoft Access 2003
3D software	SolidWorks 2010 SP0

The system platform used the “Kai-font” as output for simulated fonts. Figure 7 illustrated the writing simulation process. In addition to outputting as solid fonts, we can also choose to output as font outlines or thin fonts. In addition to providing an output mode on screen, the font simulation platform can utilize the API function in SolidWorks to import simulated fonts as 3D models into software so designers can easily carry out follow-up applications. After importing to SolidWorks, each stroke became an individual unit and can be modified as shown in Figure 8. In Chinese calligraphy, certain auspicious phrases can be written as a compound word. For example:「招財進寶」、「囍」...etc. Figure 9 and Figure 10 showed the compound word,「招財進寶」and the output screen in SolidWorks.

Figure 7 A process for Chinese character「永」

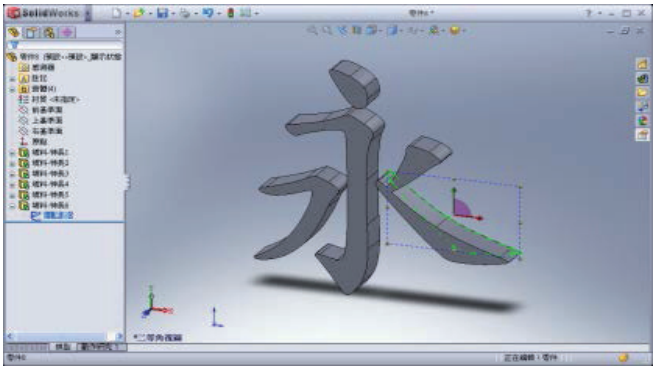


Figure 8 Output「永」to SolidWorks

Figure 9 Compound character writing

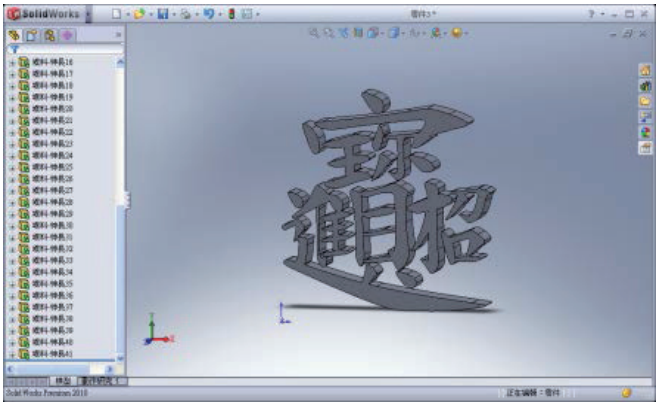


Figure 10 Output to SolidWorks

5. Conclusion

This research aims to develop a platform for Chinese calligraphy font simulation. The results and advantages are as follows:

1. A calligraphy font simulation platform was constructed for the Chinese Kai-font. Beginners can easily write out Chinese character strokes with a mouse, digital pen or other input devices and the system platform will construct a simulated calligraphy font in real-time.

2. The character stroke database for the Kai-font was reconstructed. Through character stroke feature extraction and analysis we combined strokes that were not practical under normal writing habits in order to delete duplicated strokes. This allowed us to reduce the amount of character stroke data for the Kai-font from 178,196 to 10,218, which increased system processing efficiency.

3. A coding system was made for the Kai-font character strokes. Each stroke was encoded as a 15bit unit for comparison with both the index value within the database and the hand-written input strokes.

4. With the use of API method within SolidWorks, we can quickly and efficiently export simulated fonts to 3D drawing software such as SolidWorks. This allowed further processing of Chinese compound words, creation of rarely used words, curve fitting and other additional applications.

Acknowledgement

This research was partially supported by the National Science Council, Taiwan under the Projects (NSC 102-2221-E-029, NSC 102-2632-H-029-001-MY2) and Tung-Hai University GREEnS (Global Research and Education on Environment and Society) Project (No. 4-4-3) for partially supporting this research.

References

1. Chuang, C.T., and Tseng, L.Y. (1995). A Heuristic Algorithm for the Recognition of Printed Chinese Characters. *IEEE Transactions on System, Man and Cybernetics,* 25(4), 710-717.
2. Hornby, A.S. (1972). *The Advanced Learner's Dictionary of Current English*, 2nd Ed., Oxford Press.
3. Lin, J.R., Chen, C.F. (1996). Stroke Extraction for Chinese Characters Using a Trend-Followed Transcribing Technique. *Pattern Recognition,* 29(11), 1789-1805.
4. Lin, C.F., Fang, Y.F., and Juang, Y.T. (2001). Chinese Text Distinction and Font Identification by Recognizing Most Frequently Used Characters. *Image and Vision Computing,* 19, 329-338.
5. Romero, R.D., Touretzky, D.S., and Thibadeau, R.H. (1997). Optical Chinese Character Recognition Using Probabilistic Neural Networks. *Pattern Recognition*, 30(8), 1279-1292.
6. Stallings, W.W. (1976). *Approaches to Chinese Character Recognition, Pattern Recognition*, Pergamon Press, Great

Britain, 8, 87-98.

7. Stallings, W.W. (1977). Chinese Character Recognition, in Fu, K.S., 1977, *Syntactic Pattern Recognition and Application*, Spring-Verlag, New York, 95-123.
8. Wang, C.S., Hsiao, C.Y., Chang, T.R., and Teng, C.K. (2006). Product Development for Chinese Calligraphy Using Reverse Engineering and Rapid Prototyping, *Virtual and Physical Prototyping,* 1(4), 259-269.
9. Wong, P.Y.C & Hsu, S.C. (1995). Design Chinese Typeface Using Components, *IEEE Graphics.* 0730-3157/95, 416-421.
10. Wu, Y., Yuan, Z., Zhou, D., & Cai, Y. (2013). A Mobile Chinese Calligraphic Training System Using Virtual Reality Technology, *AASRI Rrocedia,* 5, 200-208.
11. Yang, G., Liang, H., & Su, Y. (2013). Generating Chinese Characters Based on Stroke Splitting and Feature Extraction, *Displays*, 34, 258-269.
12. Zhang, S., and Fu, K.S. (1984). A Thinning Algorithm for Discrete Binary Images. *Proceedings of the International Conference on Computers and Application*, Beijing, China. 879-886.