



Generating New Styles of Chinese Stroke Based on Statistical Model*

Miao Xu¹ Jun Dong^{2**}

¹ School of Information Science and Technology, East China Normal University,
Shanghai, 200062, P.R. China.

xumiao_1@yahoo.com.cn

² Software Engineering Institute, East China Normal University,
Shanghai, 200062, P.R. China.

jdong@sei.ecnu.edu.cn

Abstract. Chinese calligraphy is one of the most important arts in Chinese history. It is entertainment form as well as embodiment of imagery thinking. In this paper, a statistical model based approach for generating new styles of Chinese character stroke is proposed. First of all, original calligraphy samples are aligned into the common co-ordinate frame, and a training set consisting of landmarks is generated semi-automatically. After that, several most significant features of the training set are extracted, and a statistical model is built in order to generate strokes with new styles. Then, Bezier curve is used to fit the discrete contour data. Some results are demonstrated finally.

1 Introduction

In calligraphy, Chinese character can have many different fonts, and even the same font often appears in different styles. Figure 1 shows different “Horizontal” strokes in “Li”, one of the most widely used Chinese calligraphy font.

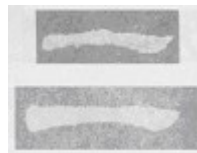


Fig. 1. Different “Horizontal” strokes styles in font “Li”

Because of the complicated structure and appearance, someone being able to master some styles doesn't mean that he can also manage other styles satisfactorily. Moreover, some written characters are not aesthetically well enough. On the other hand, simulating on calligraphy creation also has its importance in cognitive science, where cognitive processes, patterns of imagery thinking and model of intelligence are key issues[1]. In this paper, a statistical model based approach for generating new styles of Chinese stroke is proposed to simulate the basic course calligraphy creation.

* Supported by Shanghai Basic Research Key Project (06JC14058), National Basic Research Program of China (2005CB321904).

** Corresponding author. Tel: +86-21-62232181; Fax: +86-21-62233169.

The rest of the paper is organized as follows. In section 2, some related works are introduced. In section 3, the overview of the approach proposed in this paper is given. In section 4, the statistical model is presented step by step. In section 5, some results are demonstrated. Finally, its potential application and some open questions are discussed in section 6.

2 Related works

In most cases, the automatic generation of Chinese calligraphy can be divided into several phases: (1) shape decomposition, (2) model creation from training set and (3) generation of new artworks. An algorithmic framework for an advanced virtual brush to be used in interactive digital painting has been proposed[2]. Compared with other virtual brushes, this system is designed to present a realistic brush in the sense that the system accurately and stably simulates the complex painting functionality of a running brush. Moreover, an intelligent system using a constraint-based analogous-reasoning process is presented[3]. The system fuses knowledge from multiple sources to support a restricted form of reasoning. The result is pleasant since it can automatically generate Chinese calligraphy that meets visually aesthetic requirements. However, the transformation on each level there depends on the weight of different samples. The efficiency is a bit low since a lot of experience is required to adjust the weight. It would be better if some primary features of a character can be extracted at first. This is the main purpose of our work, i.e., obtaining an approach to extract some critical features from large numbers of samples and then build a model with those extracted features.

3 Overview of the new approach

What we want to get is a statistical model with a parameter. This statement can be represented as:

$$X = M(b),$$

where X is target stroke to be generated with new styles and parameter b is a vector to adjust the stroke features (e.g. long or short of the stroke). Its dimensions represent the number of modes being found. When the parameter is changed with some specific values, a new style will be generated. It's better to find a certain correspondence between the parameter and the stroke style.

Therefore, some samples are selected in order to form a training set. Each of them is marked with several points, including the points on the contour and the control points which will be used by generating the consecutive contour curve later. After that, the training set is aligned by using Generalized Procrustes Analysis (GPA)[4] and the primary features of these samples are extracted by Principal Components Analysis (PCA)[5]. Later on, a few characteristic features in the training set are extracted and represented by the form of several eigenvectors. With the extracted features, we can generate new styles of strokes. Figure 2 shows the steps.

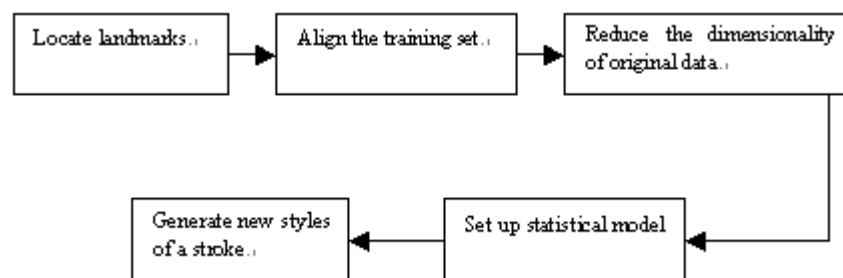


Fig. 2. Five steps for generating new styles of a stroke

4 Construction of the model

4.1 Location of Landmarks

To form a training set, some suitable landmarks (i.e., selected points on each boundary of the image) should be located. Good choices for landmarks are points which can be consistently located from one image to another. In a 2-dimension image, points could be located at corners of object boundaries[6]. In our approach, a semi-automatic method is used.

1) Calculate the angles of all the points on the contour of stroke. The l -angle of a point P_k is defined as the angle between two lines $P_{k-1}P_k$ and P_kP_{k+1} . Usually the calligraphy works are not clear enough, error would appear in the contour data. So we compute ten angles from l -angle to $l0$ -angle and make the mean as the final angle of point P_k .

2) Search the points which are located at a sharp turning of the boundary. To find these points, a threshold is set. If the angle of point P_k is less than the threshold, the point will be inserted into the training set. Some points that are along the boundary between two sharp turning points are also inserted into the training set so that the shape of a stroke can be described clearly.

3) Insert the control points of Bezier curve into the training set.

Making landmarks by hand is also allowed here when the landmarks generated automatically by computer fail to satisfy the requirements. Figure 3 is an example of landmarks which are generated automatically. The dashed is the real contour of the stroke while the solid line is a generated contour by the points marked. Note that the rectangle points are those on the contour and circular points are the calculated control points.

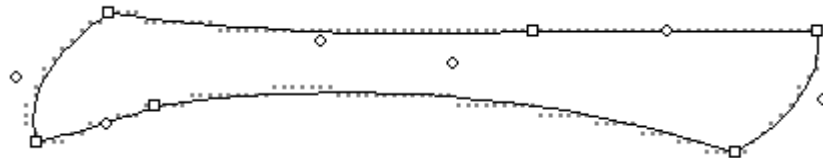


Fig. 3. Landmarks generated automatically. The rectangular points are the points on the contour and the circular points are the control points. The dashed is the real contour of the original image and the solid line is a generated contour by the points marked

A character can be represented by those landmark points. For example, if there are n landmark points for a stroke, $\{(x_i, y_i)\}$ for the i th point's co-ordinate, then a $2n$ elements vector, $X = (x_1, \dots, x_n, y_1, \dots, y_n)^T$, is composed to describe the stroke.

4.2 Align the training set

Before any further step on the training set, the variation that might affect the later analysis should be removed and make sure that all the samples are in the same co-ordinate frame.

One of the most popular algorithms among them is GPA, which is used for comparing shapes of objects. The main idea of this method is moving each shape to minimize the sum of the distances of each shape to the mean.

During the alignment, each shape has three kinds of transformations: translation, scaling and rotation. Because a shape is invariant without considerations of position and size under these Euclidean similarity transformations, the geometrical information remains after the process.

We make all samples center on the origin, with a unit scale and some fixed but arbitrary orientation. Three different approaches to alignment are introduced in [6] and the last operations is used in this paper. This approach is to transform each shape into the tangent space to the mean. The tangent space to a vector \mathbf{x} is the hyperplane normal to \mathbf{x} , passing through \mathbf{x} . The advantage of the tangent space approach is that it keeps the distribution compact and keep any non-linearity to a minimum[6].

4.3 Statistical model

If the distribution of parameter \mathbf{b} can be learnt from the training set, we are able to generate new styles, which are related with but different from the original samples.

Suppose we have s samples and each sample has n points in a 2-dimension plain. Usually n is a large number in our cases (determined by the number of landmarks), so it is not an easy task to analyze such a huge dimension of data. We have to reduce the dimensionality at first.

Reduce the dimensionality. PCA is a useful method to reduce the dimension into a range that is manageable. The basic idea of the method is to describe the variation of a set of multivariate data in terms of uncorrelated (linearly independent) variables, each of which is a particular linear combination of the original variables. The new variables are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data. The objective of this analysis is usually to judge whether the first few components account for most of the variation in the data. If so, it is argued that they can be used to summarize the data with little loss of information while reduce the dimensionality of the data.

PCA is often performed by Singular Value Decomposition (SVD) on covariance matrix of the samples. This process gets two orthogonal matrixes and a diagonal matrix, that is, $\mathbf{X}=\mathbf{U}\mathbf{S}\mathbf{V}^T$. Here, \mathbf{X} is the covariance matrix of samples, \mathbf{S} is a diagonal matrix and both \mathbf{U} and \mathbf{V} are orthogonal matrixes. In \mathbf{S} , the elements (eigenvalue of \mathbf{X}) are arrayed in descending order. Suppose Φ_i is the corresponding eigenvector of λ_i , then we can create a model with Φ , which contains t eigenvectors corresponding to the t largest eigenvalues.

Set up the model. After all the eigenvectors are computed, t most significant eigenvectors are selected to form a matrix. Here, the t eigenvectors could represent the main feature of the stroke. Therefore, the new styles of a stroke can be generated by using the following equation:

$$X = \bar{X} + \Phi \bullet b,$$

where \bar{X} is the mean after the alignment of original sample; Φ is the matrix that composed of the t eigenvectors mentioned above.

Generate new styles. A new style that is closely related with the original training set can be generated if we give a specific \mathbf{b} . In our cases, it proves not to be an easy task to define the range.

4.4 Curve fitting of the discrete contour data

The data generated by the model are some discrete points. To get the consecutive representation of a stroke, curve fitting is necessary. There are several alternative methods. The most widely used method is to represent the contour of a stroke with a few sets of lines and curves. Here, the curves can be quadratic spline, Beta-spline, B-spline or Bezier curves which is not larger than cubic [7]. What we used in our cases is quadratic Bezier curve. It has several advantages in our model:

1) It needs only a few points to generate the consecutive contour of a stroke (see Figure 3). This is very important because the amount of computation has been significantly decreased. It also removes the error among the training set in our cases.

2) It can retain the original style of the font when a stroke is zoomed in or zoomed out, so it has been accepted as industry standard.

When fitting the contour, we first judge whether the points are on a line or not. If they are, Bresenham algorithm is used to fit the contour; if not, the quadratic Bezier curve is applied[8]. The control points which we computed during the period of landmark are used in this process.

The result we get proves that the quadratic Bezier curve is good enough to generate a Chinese stroke with just a few points.

5 Results

We use Visual C++ and Matlab 6.5 to implement the model proposed in this paper and take some experiments on it.

In the first experiment, our training set comes from only one sample, that is, Chinese character “Horizontal” stroke in font “Li” (see Figure 3). This character is stored in Microsoft Word 2000. Figure 4 shows the results. The places marked with circles have been changed significantly with different parameter.

In the second experiment, our training set is also based on Chinese character “Horizontal” in font “Li”, but the original samples come from multiple sources. Parts of them are showed in Figure 5 and the results are given in Figure 6 and Figure 7 respectively.

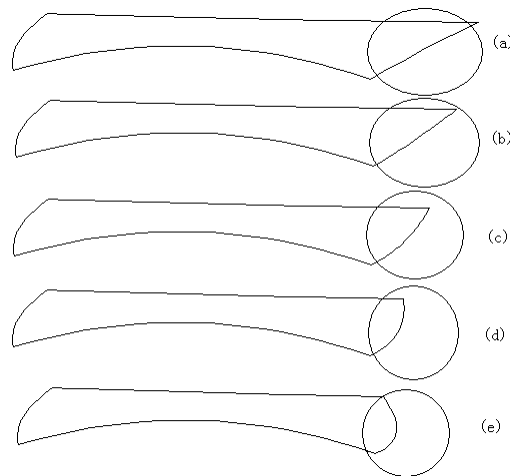


Fig. 4. The training set comes from only one sample. The parameter b has only one dimension: $b = [b_1]$, which is adjusted from -0.10 to 0.10 : (a) $b_1 = -0.10$, (b) $b_1 = -0.05$, (c) $b_1 = 0$, (d) $b_1 = 0.05$, (e) $b_1 = 0.10$.

The results of the second experiment indicate that the dimension of b actually represents different modes.



Fig. 5. Samples in the second experiment

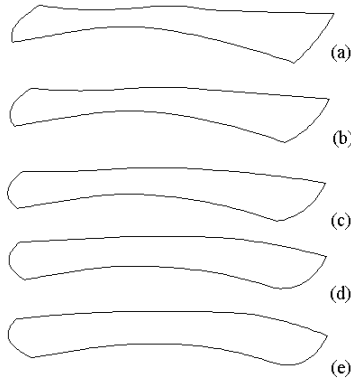


Fig. 6. The training set comes from multiple samples. The parameter b computed by PCA has three dimensions: $b = [b_1, b_2, b_3]$. The first element b_1 in parameter is adjusted from -0.10 to 0.10 : (a) $b_1 = -0.10$, (b) $b_1 = -0.05$, (c) $b_1 = 0$, (d) $b_1 = 0.05$, (e) $b_1 = 0.10$. The other two elements, b_2 and b_3 , remain 0 in this experiment. Note that all of the new generated strokes are similar to the original samples, but each of them has different style. In conclusion, they are derived directly from the statistics of a training set

These modes explain global variation due to different styles. The results also demonstrate that less significant modes cause smaller, more local changes (see Figure 7). For example, in our cases, the mode, which corresponds to b_1 , is more significant than other modes, which correspond to b_2 and b_3 .

Figure 8 gives the results our system obtained using four training samples as the input. The results show that this approach can yield strokes which have different styles and strokes can be reconstructed as a single character.

Note that the characters in Figure 8 are decomposed into several primitive strokes and reconstructed by hands.¹

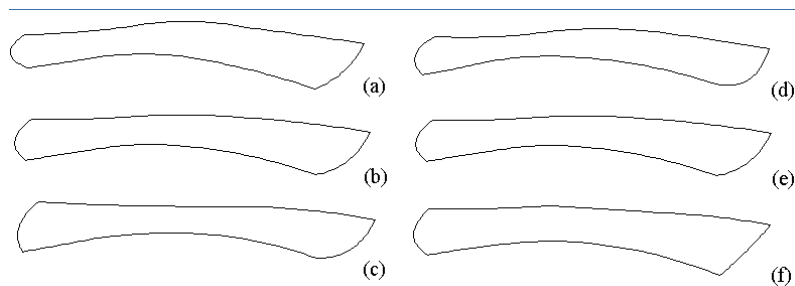


Fig. 7. The training set is the same with the second experiment. But this time other elements in parameter b have been adjusted. The first column corresponds to the adjustment of b_2 and the second column corresponds to the adjustment of b_3 . In the first column, the change mainly occurs at two ends of the stroke (the left end becomes thick and the right becomes thin with the increase of b_2); while in the second column, the change mainly occurs in the middle part of the stroke (the upper boundary in middle part becomes protruding with the increase of b_3)

¹ We selected some generated strokes and constructed a single character by hands. An automatic method is being studied and developed to aid this work. However, this work is beyond the paper here.

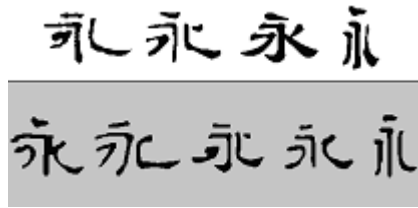


Fig. 8. A single character (means “forever” in Chinese) in different styles. The first row is the training samples, and the second row is reconstructed with strokes automatically generated by our system

6 Applications and discussions

The approach proposed here has several potential applications. First of all, it can be applied in the publishing industry, especially those publications of ancient artworks. For some calligraphy works suffer a lot of abrasion during long history, it is impossible to maintain the original artwork. This approach could be used to generate some similar characters to be selected for publishers.

Another application is to generate personalized fonts according to users’ preference. For example, users provide the system with their handwritings. Based on these knowledge sources, the approach will extract the primary features and generate some plausible styles. Users can adjust and choose one or more styles as their favorite appearance.

The approach has some advantages in the generation of new styles of Chinese character strokes. It does not need any large amount of computation since PCA algorithm is adopted and it reduces the dimensionality of the original samples into a manageable range. We also use quadratic Bezier curve to fit the discrete contour data. This requires only a few points to generate the new styles of a stroke.

However, to make a further investigation, a lot of works are necessary. First of all, if the training set is too small, the effect is unsure. Figure 9 explains the detail about this point of view. This defect can be eliminated by supplying the model with more samples.

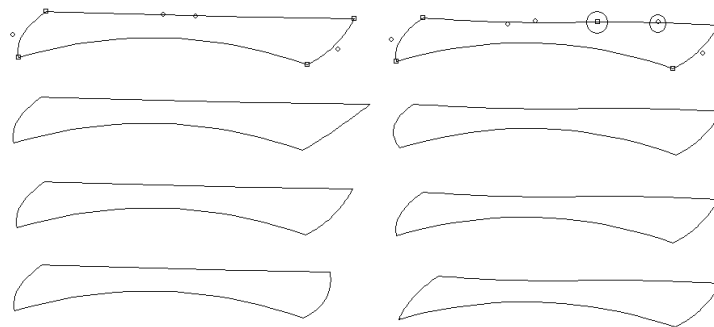


Fig. 9. The training set that is used in the first experiment (only one sample). The difference between two columns is the number and position of landmarks (the two points marked with circles in the second column). The landmarks in first column are what we did in the first experiment, while the landmarks in second column are marked by hand. The result shows that the area changed by the model is related with the number and position of landmarks when samples are not enough

Secondly, up till now, we are not able to make a one-to-one correspondence between the elements in parameter \mathbf{b} and the specific features of style (long or short). If we can find such kinds of relationship, this approach will be more efficient. For example, if element b_1 represents the length and element b_2 represents the width, then

we can simply adjust the length or width without changing other features by only adjusting the related element. Anyway, this is an open question in our experiments.

Thirdly, our model focuses on the transformation of a stroke instead of a whole character, so the construction of a character with strokes is not mentioned here.

7 Acknowledgement

The authors would like to acknowledge Dr. Lao Zhiqiang of Department of Radiology, University of Pennsylvania for providing useful material and discussion.

Reference

1. Dong Jun: *Possible Approaches on Imagery Thinking Simulation*, World Science-Technology Research & Development, 2006, **28**(3): 7-10 (in Chinese).
2. Xu Songhua, Lau Francis C. M., Tang Feng, Pan Yunhe: *Advanced Design for a Realistic Virtual Brush*, Computer Graphics Forum, September 2003, **22**: 533-542.
3. Xu Songhua, Lau Francis C. M., Cheung William K., Pan Yunhe: *Automatic Generation of Artistic Chinese Calligraphy*, IEEE Intelligent Systems, 2005, **20**(3): 32-39.
4. Goodall C.: *Procrustes Methods in the Statistical Analysis of Shape*, Journal of the Royal Statistical Society B, 1991, **53**(2): 285-339.
5. Matthew James Francis Cairns: *An Investigation into the use of 3D Computer Graphics for Forensic Facial Reconstruction (chapter2)*, <http://www.dcs.gla.ac.uk/~mc/1stYearReport/Contents.htm>
6. Cootes T. F., Taylor C. J.: *Statistical Models of Appearance for Computer Vision (chapter 4)*, http://www.isbe.man.ac.uk/~bim/Models/app_models.pdf.
7. Ma Xiaohu, Pan Zhigeng, Zhang Fuyan.: *The Stroke based Chinese Outline Font and its Application*, Chinese Journal of Computers, 1996, **19**(2): 81-88 (in Chinese).
8. Donald Hearn, M. Pauline Baker: *Computer graphics* (2nd edition). Prentice Hall, New Jersey, 1997.