STYLE PRESERVING CHINESE CHARACTER SYNTHESIS BASED ON HIERARCHICAL REPRESENTATION OF CHARACTER

Ying Wang, Haitao Wang

College of Automation Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, 210016, China

Chunhong Pan, Le Fang

Institute of Automation
Chinese Academy of Sciences
Beijing, 100080, China

ABSTRACT

Most of the existing algorithms for character synthesis deal with English, and they are not suitable for Chinese character synthesis. In this paper, we propose an unified approach for modeling and synthesizing Chinese characters. Using a three-level hierarchical representation, each character is decomposed into basic components, which forms the stroke database and radical database. In the synthesis process, we use a wavelet-based approach to select proper strokes and radicals, and some aesthetic constraints are defined based on the relationships between components, then genetic algorithm is employed to search for the optimal results which best match the aesthetic constraints. Experimental results demonstrates the effectiveness of our method.

Index Terms— Character generation, Wavelet transforms, Image processing, Genetic algorithms

1. INTRODUCTION

Personal style character synthesis aims at generating texts in the writer's style from a few samples. A handwriting synthesis has a variety of applications including automatic creation of personalized documents, generation of large quantities of annotated handwritten data for training recognizers [1].

The problem of handwriting modeling and synthesis has been addressed for a long time, and there are many related studies in the literature. Generally speaking, these approaches can be roughly divided into three categories.

The first one is biomechanics approach. In [2] modulation models are used to model and represent the handwriting trajectory, and the handwriting trajectory is analyzed and modeled by velocity or force functions. However, these studies mainly focus on the representation and analysis of real handwriting signal, rather than handwriting synthesis. The second one is shape distortion approach [3]. First the method uses a novel shape matching algorithm to match the two handwriting samples, then performs thin plate spline to generate new samples. The generated samples are just "between" the two training samples, and therefore the variation is very limited. The third category involves shape simulation approaches. It is more practical than methods mentioned above. A straightforward approach is proposed in [4], where handwriting is synthesized from collected handwriting glyphs. In [5], Wang et al. propose a learning-based cursive handwriting synthesis approach. The trajectory is represented by landmark-based spline, and the problem of learning personal handwriting style is transformed to statistically analyze the distribution of landmark points.

All of the methods mentioned above are used for English handwriting synthesis. In [6], Choi et al. present a character generation method based on Bayesian network. Though it can be extended to Chinese character synthesis with personal style, a large amount of handwriting samples have to be collected for each user, since the number of Hangul character is much more than the number of English character. In [7], Jawahar et al. propose a synthesis model for generating handwritten data for Indian languages, using the stroke model and layout model that it can synthesize natural looking words. However it also suffers the same problems in [6].

In this paper, we propose a novel method to synthesize Chinese characters which do not exist in the samples. We exploit wavelet-based approach to select proper components, then employ generic algorithm to search result which best satisfies the aesthetic constraints.

2. CHARACTERISTICS AND REPRESENTATION HIERARCHY OF CHINESE CHARACTER

The difficulties associated with handwriting synthesis are greatly depending upon the nature of character that one wants to synthesize. English contains a small set of symbols and has simple spatial layout, however, Chinese character has a huge set, the number of commonly used characters is more than 3000, and each character has complex spatial layout. Furthermore, Chinese character has various styles and character of different style has different shape of components and spatial layout. All of these make synthesis of Chinese character

This research is sponsored by Natural Science Foundation of China (NSFC No. 60675012).

much more difficult than English character.

For artistic and personal style Chinese character, both the shape of component(radical or stroke) and the relationships between components are important. In order to synthesize artistic and personal style characters, we need to consider both of characteristics from shape and relationships. With selecting the components, we denote the set of all the components in the same style as S. Assume S is in the metric space, then there exists a metric $\rho(s_1, s_2)$, in which $s_1, s_2 \in S$. There are different ways to formulate the metric, when regarding s_1, s_2 as two sets of pixels x_1, x_2, we define the metric as follows:

$$\rho(s_1, s_2) = \sum_{x_1 \in X_1} \min_{x_2 \in X_2} \hat{\rho}(x_1, x_2) + \sum_{x_2 \in X_2} \min_{x_1 \in X_1} \hat{\rho}(x_1, x_2), \quad (1)$$

where $\hat{\rho}$ is Euclidian distance in Euclidean space.

Though Chinese character has a large number of characters, fortunately, the many thousands of Chinese characters can be composed from a relatively small number of basic strokes and radicals. Our approach decomposes Chinese character into the three levels: stroke, radical, and character. Figure 1 shows the hierarchical representation of the Chinese character "liu". The first level is strokes, which are combined to form radicals(second level). By grouping radicals according to certain spatial layouts, the character can be generated(third level).

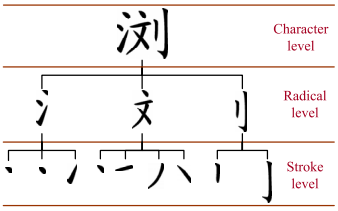


Fig. 1. Three levels representation for Chinese character

For synthesizing personal style Chinese character we need to consider not only the shape of stroke and radical but also the spatial layout between components and thickness of strokes, all of these will be described in detail in the next section.

3. MODELING OF CHINESE CHARACTER

Figure 2 gives the outline of the synthesis steps of our method. The architecture can be divided into two parts. The left part is *offline learning*, which includes two basic models: components model and layout model. The right part is *online synthesizing*, which includes two important parts: components selection and optimal combination of components. In the following subsections we will describe them in detail.

3.1. Spatial Layout Modeling

Layout modeling is very important because spatial layout is capable of capturing the spatial relationships between component classes. Chinese character has complex spatial layout, to model personal style of spatial layout is very complex and difficult. For simplification, here we use two strategies to model the Chinese character spatial layout. The first one is to model the spatial layout by using the bounding box of component. We denote it as $[x, y, w, h]$, where x, y denote the coordinate of top-left corner of the bounding box, and w, h denote its width and height. The second strategy uses the $[x_c, y_c, w, h]$ to model its spatial layout, where x_c, y_c denote the centroid of the component, and w, h denote the width and height of the bounding box.

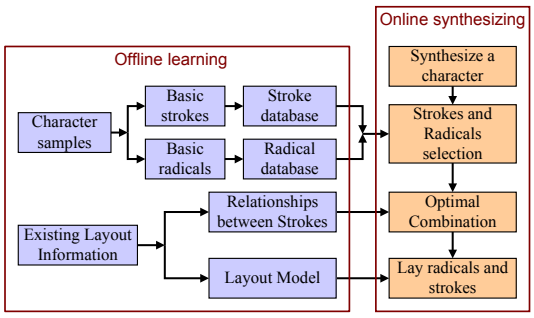


Fig. 2. Architecture of character synthesis

3.2. Modeling of relationships between components

For aesthetic and personal style Chinese character, the relationships between the strokes are as important as the shape of the strokes. Here, we consider a stroke as a set of pixels. We denote any two sets of pixels as A and B, and the number of blocks in a set of pixels as n_b. Denote the relationships between two strokes A and B as $R_l(A, B)$,which usually have four cases as follows:

$$R_l(A,B) = \begin{cases} \textbf{Separated} & A \cap B = \varnothing; \\ \textbf{Included} & A \subset B \textbf{ or } B \subset A; \\ \textbf{Intersected} & n_b(A-B) > n_b(A) \textbf{ and } n_b(B-A) > n_b(B); \\ \textbf{Joined} & n_b(B-A) = n_b(B) \textbf{ or } n_b(B-A) = n_b(B), \end{cases} \quad (2)$$

where $\varnothing$ means empty set. When synthesizing the characters, the control of the relationships of strokes is one of the most significant methods to stylize the character structure. With two given criterion strokes A_s and B_s, when $R_l(A, B) \neq R_l(A_s, B_s)$, the differences of relationships should be considered. Generally, we define the distance of a pair of strokes A, B with standard layout A_s, B_s as:

$$\rho_l\left[(A,B),(A_s,B_s)\right] = \begin{cases} \rho(A \cap B, A_s \cap B_s) + \\ \rho(A \cup B, A_s \cup B_s), & \text{if } R_l(A,B) \\ & = R_l(A_s,B_s); \\ \rho(A \cap B, A_s \cap B_s) + \\ \rho(A \cup B, A_s \cup B_s) + w_l, & \text{otherwise,} \end{cases} \quad (3)$$

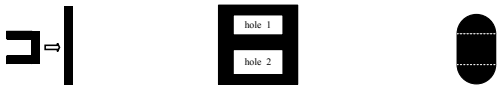
where ρ is the metric of two strokes defined in equation (1), ω_l is a weight for balance the relationships and Euclidean distance.

3.3. Description of stroke thickness

With different style Chinese characters, the thicknesses of strokes may be different. However, in the same character, we should harmonize all the strokes. There are different styles of strokes in Chinese, but strokes in same style are nearly have the same thickness. So we treat any stroke as a spindly polygon or as a rectangle. The area and circumference are not changed during the transform as shown in Figure 3(a). For simply connected stroke, if the thickness is t, and the length is l ($t < l$), we can represent the area $S = t \cdot l$, and the circumference $c = 2(t + l)$. Easy to solve the thickness as

$$t = \frac{c - \sqrt{c^2 - 16S}}{4}. \tag{4}$$

With the stroke which contains holes like Figure 3(b), we can



(a) (b) (c)

Fig. 3. Different model of thickness: (a) The polygon and rectangle model of stroke; (b) Holes in a stroke or a radical;(c) Model for circle-like stroke.

detect the amount of holes n_h using connected components labeling algorithm. The number of blocks n_b can be detected by similar method. Then the thickness can be solved as

$$t = \frac{c - \sqrt{c^2 - 16(n_b - n_h)S}}{4}. \tag{5}$$

If the stroke is similar as a circle, which is a special case like Figure 3(c), we will treat it as a combination of one rectangle and two semicircles. Then the thickness can be found

$$t = \begin{cases} \dfrac{c - \sqrt{c^2 - 4\pi(n_b - n_h)S}}{\pi(n_b - n_h)}, n_b - n_h > 0; \\ \dfrac{c - \sqrt{c^2 - 16(n_b - n_h)S}}{4}, n_b - n_h \leq 0. \end{cases} \tag{6}$$

4. SELECTION OF STROKES AND RADICALS

The next step is to select strokes and radicals to synthesize a character. Here,we employ a wavelet transform on the radical and stroke image to extract wavelet features. Comparing with existing criterion radical and stroke, we get the most likely radicals and strokes for candidates.

4.1. Wavelet Transforms

It is well known that two-dimensional wavelet decomposition of a discrete image $I(m, n)$ represents the image in terms of $3n + 1$ subimages:

$$cA_n, [cH_n, cV_n, cD_n], \cdots, [cH_1, cV_1, cD_1], \tag{7}$$

where cA_j is the approximation of the image at resolution 2^{-j}; cH_j, cV_j and cD_j are the wavelet coefficients containing the image detail at resolution 2^{-j}, respectively, to horizontal high frequencies, vertical high frequencies and high frequencies in both directions. After n-level wavelet decomposition, we get $3n + 1$ subimages from original image.

Coefficients of different resolutions reflect the information of images with different widths; blocks with different sizes in one subimage reflect the information of substructures with different dimensions. To obtain effective features, we utilize blocks with different sizes in one subimage, and extract features from subimages of different resolutions [8].

For one subimage with the size of $N \times M$, we divide it into $K \times K$ non-overlapping blocks, where N and M can be divided exactly by K. We sample each block as follows to get one feature:

$$z = \sum_{(x,y) \in B} |f(x, y)| \cdot w(x, y), \tag{8}$$

where B denotes one block, $f(x, y)$ denotes the wavelet coefficient at (x, y). The $w(x, y)$ is a weighting function defined as follows:

$$w(x, y) = \exp(-\frac{(x - x_{center})^2 + (y - y_{center})^2}{\alpha}), \tag{9}$$

where x_{center} and y_{center} denote the coordinates of the center point of block B, α is a constant which is set as 4.0 in our experiment.

We use 3-level decomposition to get cA_3, $[cH_3, cV_3, cD_3]$, $[cH_2, cV_2, cD_2]$, $[cH_1, cV_1, cD_1]$ wavelet coefficients. Our purpose is to select similar strokes or radicals, so the approximation information is much more important than the detail information. To cH_1, cV_1, cD_1,we choose K=2, and therefore obtain get 3×(2×2)=12 features from these three subimages. To cH_2, cV_2, cD_2 we choose K=2 and K=4, and therefore we get $3 \times (2 \times 2 + 4 \times 4)$ =60 features from these three subimages. To cH_3, cV_3, cD_3 we choose K=4 and K=6, we get 3×(4×4+6×6)=156 features from these three subimages. To cA_3 we choose K=4 and K=6, we get 4×4+6×6=52 features from this subimage. We combine all these extracted features and obtain 12+60+156+52=280 dimension feature vector. Using the above wavelet-based feature, we can get suitable strokes and radicals.

5. OPTIMAL COMBINATION WITH A GENETIC ALGORITHM

Since we have the candidates of strokes and radicals for a certain character, and we also get the spatial layout, then we can synthesize many characters. Utilize the relationships between strokes which described in Section 4.3 for the aesthetic constraints, we can get the optimal results. However the process of achieving the global optimal result is a NP-hard[9]. Genetic Algorithms(GA) offers a class of global search techniques based on the selective adaptation of a population of

individuals [9]. Here we define the fitness function of any possible character $\mathcal{C}$ as follows:

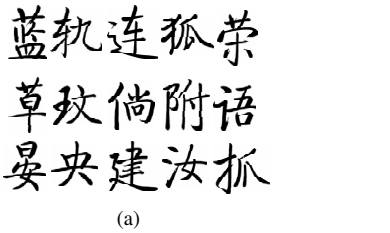
$$E(\mathcal{C}) = \sum_{A,B \in \mathcal{C}} \rho_l \left[(A, B), (A_s, B_s) \right], \tag{10}$$

where ρ_l is defined in equation (3). Let C be the number of components to synthesize a certain character. The optimal result can be found by genetic algorithm as follows:

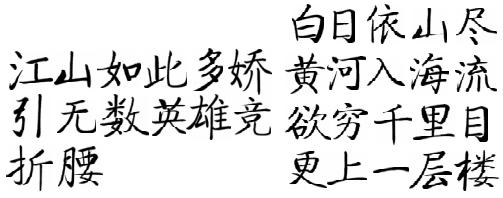
1. Let the population size be $100 \times C$,
2. Set the search range for each components in its candidates,
3. Compute the fitness of the individuals using equation (10), employed Elitism to select 10% of the population which directly copy to next generation ,
4. Let the probability of crossover be 0.8, the probability of mutation be 0.1,
5. Terminate the search when a best result has been the same one for 10 continuous generations.

6. EXPERIMENTS AND RESULTS

We collect 85 characters(each character has several components) with personal style of a certain person, and some samples are shown in Figure 4(a). Then we decompose these characters using the 3-levels model, and get the radical database and stroke database. After choosing the radicals and strokes using the method described in Section 4, genetic algorithm is exploited to get the optimal result. Finally, we adjust the thickness of the strokes using the method described in Section 3.3. Figure 4(b) and Figure 4(c) show some results we have synthesized using the spatial layout of "kai" font, both of the results are acceptable. It is noted that these characters do not exist in sample database.



(a)



(b)

(c)

Fig. 4. Samples and Synthetical results:(a) Selected samples;(b) A poetry of Mao Zedong;(c) A poetry of LiBai

7. CONCLUSIONS

In this paper, we propose a novel method to synthesize Chinese characters with personal style. Since shapes of components and relationships between components are both significantly important to aesthetic criterion and representation of personal style, we consider both of them to synthesize character. We apply wavelet-based method to select proper components and genetic algorithm to search the optimal result. However it also exists several issues. Since Chinese character has complex spatial layout, relationships of strokes could be further enhanced by incorporating more information. Another problem is the representation of the connection between the components, Chinese character has many styles, sometimes strokes or radicals in a character are connected, how to represent their connection is a problem we will consider in the further work .

8. REFERENCES

[1] Tamas Varga and Horst Bunke, "Generation of synthetic training data for an hmm-based handwriting recognition system," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2003, vol. I, pp. 618–622.

[2] Hao Chen, Agazzi.O.E, and Suen.C.Y, "Piecewise linear modulation model of handwriting," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 1997, vol. 1, pp. 363 – 367.

[3] Y.Zheng and D.Doermann, "Handwriting matching and its application to handwriting synthesis," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2005, vol. 2, pp. 861 – 865.

[4] I.Guyou, "Handwriting synthesis from handwritten glyphs," in *Proceedings of International Workshop on Frontiers of Handwriting Recognition*, 1996.

[5] Jue Wang, Chenyu Wu, Ying-Qing Xu, and Heung-Yeung Shum, "Combining shape and physical models for on-line cursive handwriting synthesis," *International Journal on Document Analysis and Recognition*, vol. 7, pp. 219–227, April 2005.

[6] Hyunil Choi, Sung-Jung Cho, and Jin H.Kim, "Generation of handwritten characters with bayesian network based on-line handwriting recognizers," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2003, pp. 995–999.

[7] C.V.Jawahar and A. Balasubramanian, "Synthesis of online handwriting in indian languages," in *Proceedings of International Workshop on Frontiers of Handwriting Recognition*. IEEE, 2006.

[8] Xiaoqing Ding, Li Chen, and Tao Wu, "Character independent font recognition on a single chinese character," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 195–204, February 2007.

[9] David E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley Longman Publishing, 1989.