Automatic Generation of Personalized Chinese Handwriting Characters

Peng Liu¹,³, Songhua Xu⁴*, Shujin Lin²,³

¹ National Engineering Research Center of Digital Life, State-Province Joint Laboratory of Digital Home & Interactive Applications, School of Information Science & Technology, Sun Yat-sen University, Guangzhou 510006, China
² School of Communication and Design, Sun Yat-sen University, Guangzhou, 510006, China
³ Shenzhen Digital Home Key Technology Engineering Laboratory, Research Institute of Sun Yat-sen University in Shenzhen, Shenzhen 518057, China
⁴ Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, Tennessee, USA, 37830

* Email: xus1@ornl.gov

Abstract — This paper presents a novel algorithmic method for automatically generating personal handwriting styles of Chinese characters through an example-based approach. The method first splits a whole Chinese character into multiple constituent parts, such as strokes, radicals, and frequent character components. The algorithm then analyzes and learns the characteristics of character handwriting styles both defined in the Chinese national font standard and those exhibited in a person’s own handwriting records. In such an analysis process, we adopt a parametric representation of character shapes and also examine the spatial relationships between multiple constituent components of a character. By imitating shapes of individual character components as well as the spatial relationships between them, the proposed method can automatically generate personalized handwritings following an example-based approach. To explore the quality of our automatic generation algorithm, we compare the computer generated results with the authentic human handwriting samples, which appear satisfying for entertainment or mobile applications as agreed by Chinese subjects in our user study.

Keywords— stroke similarity, radical similarity, extraction of Chinese character component; topological structure of Chinese characters

I. INTRODUCTION

In this personalized era, people desire to have their own personal font so that documents can be displayed as if handwritten by their own hands. For certain applications such as texting or emailing, displaying a piece of text in a personal style can add much attraction and intimacy into the conventional written communication means. For example, a manuscript or text message displayed in a user’s personal handwriting style rather than printed using a standard font can draw readers much closer to the writer. To meet this application demand, we propose a new algorithm for automatically generating personalized handwriting style based on a small amount of training samples available for a specific writer. The technology can also be applied for computer-based innovative font design and many other gaming or mobile applications that favor personal text messages.

Imitating personal handwriting style poses an interesting pattern recognition and machine learning challenge. Xu et al. [1] proposed a method that teaches the computer to learn calligraphic handwriting style of a specific calligrapher for automatic calligraphic handwriting generation. Xu et al. [2] also proposed an algorithmic method that enables the computer to generate Chinese calligraphic handwritings that satisfy certain aesthetic requirements automatically harvested from a set of training samples. Chen et al. [3] jointly applied affine transformation and image winding to modify the positions, sizes, and shapes of Chinese character components for personalized handwriting generation. Our work in this paper aims to capture and learn the handwriting style of a given writer for imitation of his/her personal writing style. Unlike the above previous work, this paper proposes a new example-based approach to generate personalized handwriting characters, which includes a new shape parameterization method for Chinese characters and a method for extracting frequent character constituent components from a Chinese character.

The rest of our paper is organized as follows. We first introduce a method based on the improvement of a piece of prior work [9] that calculates stroke similarity according to the feature points of a stroke (Sec. II). We then study several important traits of Chinese characters, such as the spatial and layout structures as well as the relative length distribution of strokes in a Chinese character, according to which we introduce a method for calculating the similarity between two character components (Sec III). In Sec. IV, we design an algorithm to extract frequent character components from a Chinese character, which serves as a fundamental step for decomposing Chinese handwriting character samples. In Sec. V, we discuss how to generate a facsimiled Chinese handwriting result using our new example-based approach. Sec. VI presents some preliminary experimental results.
II. MEASURING STROKE SIMILARITY BASED ON FEATURE POINTS

A fundamental building element in our automatic Chinese handwriting facsimile algorithm is to measure stroke similarity. The measurement results will allow the algorithm to identify most suitable components from all available sample records for the reused purpose. In our method, all strokes decomposed from a Chinese character are classified into 24 kinds. Given two arbitrary strokes, denoted as \( L_g \) and \( L_h \) respectively, we first construct their segment sequences using the method proposed in [9]. Let the construction results be

\[
\begin{align*}
L_g &= \{g_1, g_2, \ldots, g_m\} \\
L_h &= \{h_1, h_2, \ldots, h_n\}
\end{align*}
\]

respectively, where \( g_i \) represents the \( i \)-th segment in \( L_g \) and \( h_j \) represents the \( j \)-th segment in \( L_h \) (see Figure 1).

We then estimate the similarity between \( L_g \) and \( L_h \) as follows. Firstly, we define the similarity between two segment subsequences \( \{g_i, g_{i+1}, \ldots, g_m\} \) and \( \{h_j, h_{j+1}, \ldots, h_n\} \) as:

\[
f(i, j) = \begin{cases} 
\beta(g_i, h_j) & i = 1 \text{ and } j = 1 \\
\min \left\{ f(i-1, j-1), f(i-1, j), f(i, j-1), \beta(g_i, h_j) \right\} & i > 0 \text{ or } j > 0
\end{cases}
\]

(1)

where \( \beta(g_i, h_j) \) represents the similarity between two individual segments \( g_i \) and \( h_j \), as defined in (2) below.

\[
\beta(g_i, h_j) = \min \left\{ \theta_{g_i} - \theta_{h_j}, 2\pi - \theta_{g_i} + \theta_{h_j} \right\}
\]

(2)

where \( \theta_{g_i} \) represents the direction angle of \( g_i \) and \( \theta_{h_j} \) represents the direction angle of \( h_j \). Finally, the similarity between the two strokes \( L_g \) and \( L_h \) is calculated as:

\[
S_m(L_g, L_h) = \frac{f(m, n)}{m+n}
\]

(3)

III. TOPOLOGICAL STRUCTURE OF A CHINESE CHARACTER

To quantitatively represent the topological structure of Chinese character, we introduced the following representation method, which characterizes the spatial relationship between two strokes inside a Chinese character. This representation also parameterizes the character’s topological structure, its layout, as well as the relative length distribution of two strokes. Based on all these aspects of measurements, we can derive the overall similarity between two strokes in a Chinese character.

A. Spatial Relationship Between Two Strokes in a Chinese Character

Given two strokes, A and B in a character \( C_{chs} \), let their minimum bounding boxes be \( S_A \) and \( S_B \) respectively. We first focus on \( S_A \), according to whose boundary we can divide the full image space occupied by the character \( C_{chs} \) into a 3x3 grid structure (see Figure 2). We denote the nine resultant grid cell regions as \( \Delta_{A_i} \) \( (i = 1, 2, \ldots, 9) \). Also, let the area of Stroke B be \( Area_B \) and the area of
stroke B that falls into the grid cell region $\Delta_{Aij}$ be $\text{Area}_{Aij}^B$. Based on these notations, we can easily calculate the proportion of $\text{Area}_{Aij}^B$ in $\text{Area}_A$ as follows:

$$\delta_{Aij}^B = \frac{\text{Area}_{Aij}^B}{\text{Area}_A} \quad (i = 1, 2, \ldots, 9).$$

Now we can characterize the spatial relationship between strokes A and B, denoted $\Psi(A, B)$, as follows:

$$\Psi(A, B) = \left( \begin{array}{ccc} \delta_{A1j}^B & \delta_{A2j}^B & \delta_{A3j}^B \\
\delta_{A4j}^B & \delta_{A5j}^B & \delta_{A6j}^B \\
\delta_{A7j}^B & \delta_{A8j}^B & \delta_{A9j}^B \end{array} \right),$$

where $\sum_{j=1}^{9} \delta_{Aij}^B = 1$.

With the metric $\Psi(A, B)$ we can evaluate the similarity between two pairwise stroke relationships. Assume one pairwise stroke relationship governs strokes A and B and the other relationship governs strokes C and D, we can estimate their similarity $\Xi(A, B, C, D)$ as follows:

$$\Xi(A, B, C, D) = \frac{\sum_{j=1}^{3} \left( \sum_{k=1}^{3} \psi_{j,k}(A, B) - \psi_{j,k}(C, D) \right) + \sum_{j=1}^{3} \left( \sum_{k=1}^{3} \psi_{j,k}(A, B) - \psi_{j,k}(C, D) \right) + \sum_{j=1}^{3} \left( \sum_{k=1}^{3} \psi_{j,k}(A, B) - \psi_{j,k}(C, D) \right)}{3},$$

where $\psi_{j,k}(A, B)$ represents the element on the j-th row and the k-th column in the matrix $\Psi(A, B)$ as defined in formulae (5).

Given a Chinese character $C_{chs}$ that contains N strokes, we can represent it as $L_{chs} = \{L_1, L_2, \ldots, L_N\}$ where $L_i$ ($1 \leq i \leq N$) represents the i-th stroke in the $C_{chs}$. We can then measure the similarity between two stroke sequences $L = \{L_1, L_2, \ldots, L_N\}$ and $\overline{L} = \{\overline{L}_1, \overline{L}_2, \ldots, \overline{L}_N\}$ as follows:

$$\Theta(L, \overline{L}) = \frac{\sum_{k=1}^{N} \left( \sum_{j=1}^{N} \min(\Xi(L_i, L_j; \overline{L}_i, \overline{L}_j), \Xi(L_i, \overline{L}_i; L_j, \overline{L}_j)) \right) 2}{NN - 1}.$$  

B. Layout and Relative Length Distribution between Two Strokes in a Chinese Character

Given a Chinese character $C_{chs}$, which consists of a stroke sequence $L_{chs} = \{L_1, L_2, \ldots, L_N\}$. For two arbitrary strokes $L_i$ and $L_j$ ($i \neq j$) in $L_{chs}$, we denote their minimum bounding boxes as $S_i$ and $S_j$ respectively. We further represent the layout relationship between $L_i$ and $L_j$ as follows:

$$\Pi(L_i, L_j) = \left( \begin{array}{c} \gamma_i(S_i, S_j) \\
\gamma_j(S_i, S_j) \\
\gamma_k(S_i, S_j) \end{array} \right),$$

where

$$\gamma_{i}(S_i, S_j) = \frac{S_{i, x} - S_{i, x}}{S_{i, x} + S_{i, x}} \quad \text{and} \quad \gamma_{j}(S_i, S_j) = \frac{S_{j, y} - S_{j, y}}{S_{j, y} + S_{j, y}} \quad \text{and} \quad \gamma_{k}(S_i, S_j) = \frac{\text{Area}(S_i) \cup \text{Area}(S_j)}{\text{Area}(S_i) \cup \text{Area}(S_j)} \quad (9).$$

In the above formulae, $S_{i, x}$ and $S_{i, y}$ respectively represent the center coordinates of the bounding box $S_i$; $\text{Area}(S_i)$ represents the area of $S_i$. The same notations apply for $S_{j, x}, S_{j, y}$ and $\text{Area}(S_j)$. Given all the above notations, we can now represent the layout relationship of the whole character $C_{chs}$ with which the similarity between two stroke sequences $L = \{L_1, L_2, \ldots, L_N\}$ and $\overline{L} = \{\overline{L}_1, \overline{L}_2, \ldots, \overline{L}_N\}$, can be calculated as follows:

$$\Theta(L, \overline{L}) = \sum_{k=1}^{3} \sum_{j=1}^{N} \left( \begin{array}{c} \gamma_k(L_i, L_j) - \gamma_k(\overline{L}_i, \overline{L}_j) \times \mu_1 \\
\gamma_k(L_i, L_j) - \gamma_k(\overline{L}_i, \overline{L}_j) \times \mu_2 \\
\gamma_k(L_i, L_j) - \gamma_k(\overline{L}_i, \overline{L}_j) \times \mu_3 \end{array} \right),$$

where $\sum_{k=1}^{3} \mu_k = 1$. $\mu_i$ ($i = 1, 2, 3$) are adjustable variables which are empirically assigned to the values of 0.1, 0.1 and 0.8 respectively in our experiments.

Denote the length of the stroke $L_i$ and $\overline{L}_i$ as $\ell_i$ and $\overline{\ell}_i$ respectively. According to their relative length distribution, we can further estimate the similarity between two stroke sequences...
$L = \{L_1, L_2, \ldots, L_N\}$ and $\bar{L} = \{\bar{L}_1, \bar{L}_2, \ldots, \bar{L}_N\}$ as follows:
\[
\Pi(L, \bar{L}) = \sum_{i=1}^{N} \left( \frac{\ell_i}{\sum_{k=1}^{N} \ell_k} - \frac{\bar{\ell}_i}{\sum_{k=1}^{N} \bar{\ell}_k} \right). 
\] (11)

### C. Determining Similarity between Two Stroke Sequences

To measure the topological similarity of two stroke sequences, we first calculate the value of $\Theta(L, \bar{L})$, $\mathcal{H}(L, \bar{L})$ and $\mathcal{R}(L, \bar{L})$, based on which results the topological similarity of two stroke sequences, denoted as $\Pi(L, \bar{L})$, can be estimated as follows:
\[
\Pi(L, \bar{L}) = \lambda_1 \ast \Theta(L, \bar{L}) + \lambda_2 \ast \mathcal{H}(L, \bar{L}) + \lambda_3 \ast \mathcal{R}(L, \bar{L}), 
\] (12)

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1 > \lambda_2 + \lambda_3$. $\lambda_i (i = 1, 2, 3)$ are adjustable variables, which are assigned to the values of 0.65, 0.20 and 0.15 respectively in all our experiments. This configuration empirically maximizes the performance of our algorithm in our experimentation.

### IV. Automatically Extracting Character Components

#### A. Problem Statement

Chinese character can usually be divided into several frequent Chinese character components (See Figure 3). The Chinese national standard bureau defines 514 frequent Chinese character components [5].

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ be the frequent Chinese character component set. Given a Chinese character $C_{chs}$, which consists of a stroke sequence $L = \{L_1, L_2, \ldots, L_N\}$, we aim to divide the stroke sequence into $K$ ($K \geq 1$) groups such that

$\mathcal{G}^1 = \{L_1^1, L_2^1, \ldots, L_h^1\}$,
$\mathcal{G}^2 = \{L_1^2, L_2^2, \ldots, L_h^2\}$,
\[\vdots\]
$\mathcal{G}^K = \{L_1^K, L_2^K, \ldots, L_h^K\}$. 

which collectively meet the following constraints:

- $G \cap G = \emptyset (i \neq j)$, and
- $\sum_{i=1}^{K} h_i = N$, and
- $G \in \Omega, i \in [1, K]$.

#### B. Algorithm for Extraction frequent Chinese Character Components

Assume the stroke sequence $L$ that corresponds to the character $C_{chs}$ is divided into $K$ groups. In the stroke clustering process, we empirically chose the threshold $\tau$ to be 0.2, which works satisfyingly in all our experiments. If the calculated similarity is less than $\tau$, we consider that no frequent Chinese character component is detected. Let $\hat{L}$ be the working stroke sequence during the runtime of the algorithm. The algorithm works as following:

1) $\hat{L} = L$.
2) Find all frequent Chinese character components in $\Omega$, which meets
\[
\mathcal{S} \mathcal{z} \mathcal{e}(L) = \mathcal{S} \mathcal{z} \mathcal{e}(\omega), \omega \in \Omega. 
\] (14)

All the results are assembled to construct a temporary character component set as follows:
\[
\hat{\Omega} = \{\hat{\omega} \mid \mathcal{S} \mathcal{z} \mathcal{e}(\hat{L}) = \mathcal{S} \mathcal{z} \mathcal{e}(\hat{\omega})\}. 
\] (15)

Then, according to Formula (12), we can calculate the topological similarity between each character component $\hat{\omega}$ in $\hat{\Omega}$ and $\hat{L}$.

That is:
\[
\Pi(\hat{\omega}, \hat{L}), \text{ for } \hat{\omega} \in \hat{\Omega}. 
\] (16)

3) If
\[
\min_{\hat{\omega} \in \hat{\Omega}} \Pi(\hat{\omega}, \hat{L}) \leq \tau,
\]
we then extract \( \hat{L} \) as a frequent Chinese character component.

After that, we update the content of \( L \) as 
\[
L = L - \hat{L}.
\]
If \( L = \emptyset \), then stop; otherwise, go back to Step 1.

4) If \( \min_{\theta} \Pi(\theta, \hat{L}) > \tau \), update the content of \( \hat{L} \), such that
\[
\hat{L} = \hat{L} - \left\{ \hat{L}_{\text{sel}(\hat{L})} \right\},
\]
where \( \hat{L}_{\text{sel}(\hat{L})} \) represent the last character component in \( \hat{L} \).

Go back to Step 2.

Using the above algorithm, we can extract frequent Chinese character components for the majority of the input Chinese character samples.

V. AUTOMATIC HANDWRITING FACSIMILE

Figure 4 shows the overall flow chart of our system for automatically facsimileing Chinese handwriting characters. The system has three main parts: Its first part decomposes sample Chinese character into multiple constituent components; the second part then extracts and learns characteristics of personal handwriting samples by a specific writer following the algorithmic steps introduced in the above. Finally, the third part generates the personalized handwriting facsimileing results for the given writer.

For a specific writer W and a Chinese character \( C_{\text{chs}} \), which W has not written or whose handwriting sample is not available for our algorithm, the algorithm can generate the shape of \( C_{\text{chs}} \) in W’s personal handwriting style as follows.

First the character \( C_{\text{chs}} \) is decomposed into multiple constituent components \( R = \{R_1, R_2, \ldots, R_k\} \) using the method introduced in Sec. IV. After that, the following steps are executed in sequence:

1) If the character \( C_{\text{chs}} \) has been written by W, the candidate handwritten character set of \( C_{\text{chs}} \) can be constructed as
\[
\Omega_{C_{\text{chs}}} = \{C_{\text{chs}}^{\text{hw}}, C_{\text{chs}}^{\text{hw}}, \ldots, C_{\text{chs}}^{\text{hw}}\},
\]
where each \( C_{\text{chs}}^{\text{hw}} \) is the handwriting character written by W during the training phase.

In this case, we go to Step 7.

2) For each component \( R_j \), \( j \in [1, K] \), if it has been previously written by the writer W, the candidate component set for \( R_j \) can be constructed as
\[
\Omega_{R_j} = \{R_j^{\text{hw}}, R_j^{\text{hw}}, \ldots, R_j^{\text{hw}}\},
\]
where each \( R_j^{\text{hw}} \) is the handwriting frequent Chinese character component written by W during the training phase.

In this case, we go to Step 5.

3) For the component \( R_j \), its stroke sequence set can be defined as
\[
R_j^{\text{st}} = \{L_1^{\text{st}}, L_2^{\text{st}}, \ldots, L_p^{\text{st}}\},
\]
where each \( L_i^{\text{st}} \) is one stroke of \( R_j \).

For each \( L_i^{\text{st}} \), we first select up to ten candidate strokes that have been previously written by the writer W that are of the same stroke type as the stroke \( L_i^{\text{st}} \). Next, we construct a new set with these stroke samples as follows:
\[
\Omega_{L_i^{\text{st}}} = \{L_1^{\text{hw}}, L_2^{\text{hw}}, \ldots, L_m^{\text{hw}}\},
\]
Lastly, we randomly select a handwritten stroke, denoted \( L_r^{\text{hw}} \), from the set \( \Omega_{L_i^{\text{st}}} \) with the following probability distribution:
\[
P_{L_r^{\text{hw}}} = \frac{m_r(L_r^{\text{hw}}, L_r^{\text{hw}})}{\sum_{r=1}^{m} m_r(L_r^{\text{hw}}, L_r^{\text{hw}})}. \]
In a similar way, a new handwritten component, denoted \( R_j \), can be constructed. Including \( R_j \) into the set \( \Omega_{R_j} \), we have:
\[
\Omega_{R_j} = R_j \cup \Omega_{R_j}.
\]

4) Repeat Step 3 five times.

5) For the candidate component set \( \Omega_{R_j} \), we follow the following probability distribution:
\[
P_{R_k^{\text{hw}}} = \frac{m(R_k^{\text{hw}}, R_j)}{\sum_{j=1}^{K} m(R_k^{\text{hw}}, R_j)}.
\]
when selecting a handwritten component \( R_k^{\text{hw}} \) from \( \Omega_{R_j} \) randomly. In the same way, a new handwritten
character \( \overline{C}_{\text{chs}} \) can be constructed and included into the set \( \Omega_{C_{\text{chs}}} \), that is
\[
\Omega_{C_{\text{chs}}} = \overline{C}_{\text{chs}} \cup \Omega_{Q_{\text{chs}}}.
\] 
6) Repeat the algorithm from Step 2 to Step 5 five times.
7) Select a handwritten character \( C_{\text{hw}}^k \) from the set \( \Omega_{C_{\text{chs}}} \) randomly with the probability distribution of
\[
P_{C_{\text{hw}}}^k = \frac{\Theta(C_{\text{hw}}^k, R_j)}{\sum_{n=1}^{\text{Size}(\Omega_{C_{\text{chs}}})} \Theta(C_n^{\text{hw}}, R_j)}.
\]

The selected handwritten character \( C_{\text{hw}}^k \) becomes our algorithm’s end output.

VI. EXPERIMENTAL RESULTS

To verify the performance of our algorithm, in our experiments, we first divided the whole collection of available handwritten character samples into two parts—the first part for training our system to capture the personal handwriting style of a given writer, the second part used as testing samples for evaluating the quality of handwriting facsimile results automatically generated by the trained algorithm. For each target writer involved in our experiments, we invited the person to write 83 Chinese characters as training samples. Figure 5 shows the samples written by one writer. The same writer was also invited to write another 48 Chinese characters to be used as groundtruth for the evaluation purpose. Figure 6 shows the facsimile results for one writer in our experiments. For the ease of comparison, we organized these 48 testing cases into four groups. For every group, Figure 6 displays two rows where the first row shows the facsimiled results automatically produced by our algorithm and the second row displays the original handwriting samples created by the target writer. Even though our facsimiled results do not appear identical to the authentically written ones, for most of the characters, the facsimiled results appear visually acceptable as for the original human handwriting samples in use scenarios such as online texting and many other gaming or mobile applications.

VII. CONCLUSION

This paper presents a new method for automatically generating facsimiles of a writer’s Chinese handwriting through analyzing and learning handwriting features of both standard Chinese characters as defined by the Chinese national standards and the handwritten characters personally created by the writer. In our method, we adopt a parametric character shape representation method for characterizing the shape difference between components in a standard Chinese character and their counterparts in a personal handwritten character sample. As verified by our experiment results, the proposed handwriting facsimile algorithm can generate handwritings visually similar to the authentic human handwritten samples. The new algorithm can be applied to creative computer font design as well as many online applications where handwritten text messages are more appealing than standard font displays.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) grant No. 60903132, as well as NSFC (61103162), NSFC-Guangdong Joint Fund (U0935004, U1135003), the National Key Technology R&D Program (2011BAH27B01, 2011BHA16B08), and the Industry-academy-research Project of Guangdong (2011A091000032). Xu performed this research as a Eugene P. Wigner Fellow and staff member at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

REFERENCES

Figure 4. Flow chart of the proposed automatic handwriting facsimile algorithm

Figure 5. Training set of 83 Chinese characters handwritten by a target writer
Figure 6. Four groups of facsimileing results in comparison with the authentic handwriting samples. For each group, the first row shows the facsimiled results while the second row displays the authentic handwritings produced by the target writer.