Journal of Electronic Imaging

JElectronicImaging.org

Computational method for calligraphic style representation and classification

Xiafen Zhang
George Nagy

SPIE. IS&T imaging.org

Computational method for calligraphic style representation and classification

Xiafen Zhang[a,*] **and George Nagy**[b]
[a]Shanghai Maritime University, College of Information Engineering, Shanghai 201306, China
[b]Rensselaer Polytechnic Institute, Electrical, Computer, and Systems Engineering, Troy, New York 12180, United States

Abstract. A large collection of reproductions of calligraphy on paper was scanned into images to enable web access for both the academic community and the public. Calligraphic paper digitization technology is mature, but technology for segmentation, character coding, style classification, and identification of calligraphy are lacking. Therefore, computational tools for classification and quantification of calligraphic style are proposed and demonstrated on a statistically characterized corpus. A subset of 259 historical page images is segmented into 8719 individual character images. Calligraphic style is revealed and quantified by visual attributes (i.e., appearance features) of character images sampled from historical works. A style space is defined with the features of five main classical styles as basis vectors. Cross-validated error rates of 10% to 40% are reported on conventional and conservative sampling into training/test sets and on same-work voting with a range of voter participation. Beyond its immediate applicability to education and scholarship, this research lays the foundation for style-based calligraphic forgery detection and for discovery of latent calligraphic groups induced by mentor-student relationships. [DOI: 10.1117/1.JEI.24.5.053003]

Keywords: calligraphic style; style representation; style classification; style quantification.

Paper 15353P received May 8, 2015; accepted for publication Jul. 21, 2015; published online Sep. 3, 2015.

1 Introduction

Long before the introduction of wood-block printing in China and before Gutenberg's 1450 AD invention of the printing press and rediscovery of cast type, talented scribes (artists and artisans) began compiling our cultural heritage by creating visual representations of spoken language on any available medium (stone, wood, bamboo, papyrus, silk and, eventually, paper). Outstanding examples of handwritten records are called "calligraphy." Different eras have different styles, some of which are no longer used. Preserved in libraries and museums, they are now being digitized to accelerate the dissemination of their form and content.

The largest collections of calligraphy are of Chinese, Indic, Persian, Arabic, and Latin scripts. The earliest preserved samples of Chinese writing are nearly 5000 years old, and calligraphy still occupies an important place in Chinese culture. Though printed text is now ubiquitous, the ministry of Chinese education has ordered all students to learn how to write conventional script as well as how to type. The customary way of learning calligraphic writing is to copy examples of calligraphic master works of various traditional styles. To learn to write properly, students must also learn to recognize these styles. Our goal here is to develop automated aids for style discrimination rather than for identifying the origin and date of given works. We, therefore, consider calligraphic style as the intrinsic signature of homogenous groups of writers and seek critical indicators of the difference between intragroup and intergroup similarity of semantically identical characters. Although the notion of style extends to fonts and typefaces in printed matter, it will be seen that its manifestations in calligraphy are substantially different and, because of the variability of handwriting, more difficult to detect.

We introduce improved methods for segmenting page images into character images and storing the labeled character images in a database, and we propose a model for identifying calligraphic characters. The calligraphy images are scanned in the on-going China Academic Digital Associative Library (CADAL) project,[1] which is an important part of the Universal Digital Library.[2] As our contribution, we

1. segment 259 page images into 8719 individual characters and label individual characters;
2. extract visual style features and build a database of class and style annotated images; and
3. propose style representation, quantification, and classification based on visual features.

The remainder of this article is organized as follows: related research on calligraphy segmentation and style representation is discussed in Sec. 2. Section 3 describes image segmentation. Section 4 is an overview of our graphic-user-interface for labeling characters. Section 5 develops the style representation model. In Sec. 6, we present and evaluate our experimental results on style classification and quantification. Section 7 summarizes our research and proposes potential applications.

2 Related Work

Scanning calligraphy works and books into images to enable web access is well established but computing calligraphic style and offering image-based character retrieval is still under research. Researchers are exploring relationships between calligraphic image features and visual cognition of writing

*Address all correspondence to: Xiafen Zhang, E-mail: xfzhang@shmtu.edu.cn

style. Srihari[3] has been investigating handwriting individuality for decades and proposed computational methods to determine, using statistical models, whether two English writing samples originate from the same person. Panagopoulous et al.[4] also used statistical criteria to determine whether two Greek inscriptions were carved by the same writer. Writer identification shares some problems and solutions with style identification because calligraphy is a kind of handwriting. But, they differ in the following aspects:

1. Variability: writing with brush and ink leads to far more thickness variation, even within the same stroke, than writing with a pen or pencil.
2. Deformation: calligraphers often manifested their personality through special artistic effects of purposeful deformation or quick writing with "dry strokes."
3. Deterioration: works created hundred or thousand years ago exhibit noise due to paper and ink degradation.

Azmi et al.[5] propose a way of classifying Arabic calligraphy based on features extracted from scalene triangles. The major procedures of data collection, preprocessing, feature extraction, and testing are the same as in calligraphy style identification. But triangle features do not work for Chinese characters because each Arabic word is composed of consecutive left to right strokes, and its letters can be split into triangle blocks, while Chinese characters are composed of left to right, top to bottom, outside to inside, and mixed stroke components that cannot be split in the same way. Bar-Yosef et al.[6] represent historical Hebrew calligraphy using feature vectors extracted from selected letters and identify the writers by comparing questioned handwriting with samples of known handwriting. Their methods can be used for Chinese characters, but the features and style-matching we propose are better suited to the higher stroke complexity of Chinese characters.

Like CADAL, the National Digital Library of China[7] also digitizes and posts on the web historical Chinese documents. But, due to the lack of corresponding image understanding technology, they too only provide "metadata-based search," not "content-based retrieval." There are only a few papers about Chinese calligraphy segmentation. Peng's group proposes segmentation by local projection and stroke extraction.[8] It works well for old printed characters but not for handwritten characters with deformed and overlapping strokes. In our early CADAL work, we segmented 483 page images into 13,351 individual character images.[9] Here, we show an almost tenfold increase in the size of the database due to the improvement of our techniques described below.

Previous research on style includes that of Jinhui et al., who synthesize cursive calligraphy by brush texture patches collected from artworks,[10] and of Yu et al.[11], who synthesize style-consistent text using characters written by the same calligrapher. The synthesis of style-similar patches, strokes or characters does not, however, explicitly classify style similarities. Xu et al.[12] also imitate the style of given examples to generate new calligraphic characters. Their method is interesting and useful, yet it does not quantitatively represent particular styles. Lu et al.[13] construct latent style models and present experiments including data from our former database. The idea of computing the similarity between different styles is inspiring, but they do not specify style features and style elements.

We presented preliminary work in Ref. 14 based on a smaller database, different features and different classifiers on pairs of characters. We were most successful when comparing pairs with the same GB label. In the current work, we do not depend on the GB label.

3 Segmentation

Scanned page images are first segmented into individual characters in reading order. Then, individual characters are labeled by the standard GB 2312 Chinese character set code. Each character image is located by the coordinates of its minimum bounding box and indexed by its character identifier, page number, book number, and work identifier.

3.1 *Source of Calligraphy*

The calligraphy books are scanned page by page at 600 dots per inch (dpi) or 23.6 lines per inch (lpi) into 24-bit RGB and retained in both DjVu and TIFF formats by the CADAL scanning center. Lossless TIFF format is used for page segmentation, and the more compact DjVu format is used for display. Figures 1(a)–1(d) show examples of pages with vertical lines, stamps, low contrast and salt-and-pepper noise. Figure 1(e) shows a page without noise that can be segmented according to the background swaths between columns and rows.

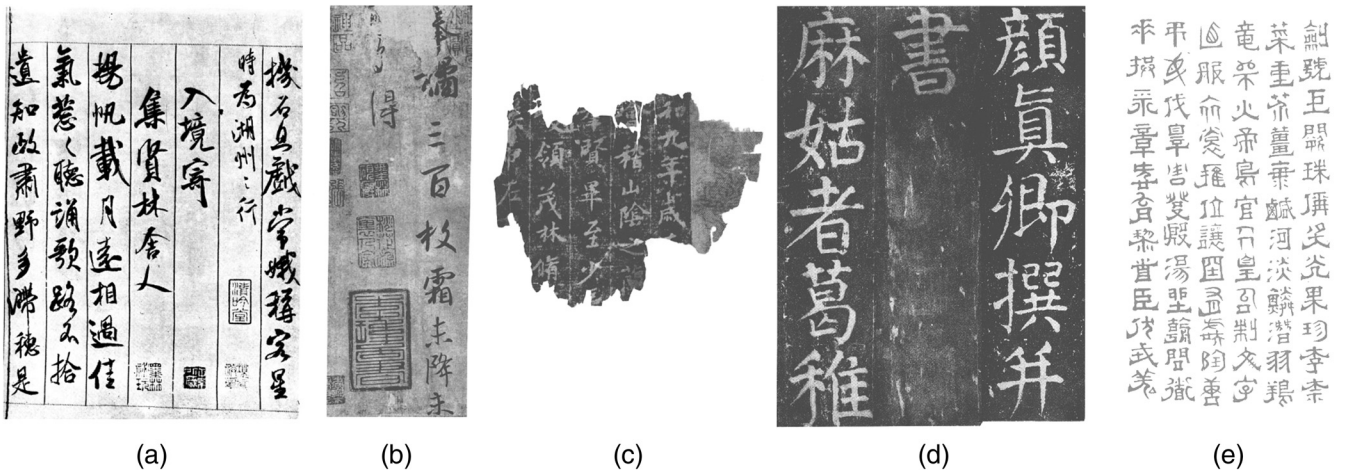
3.2 *Binarization and Segmentation*

Seals affixed by successive owners, such as those shown in Fig. 1(b), are detected by a click on the seal border before the RGB images are binarized. Page borders are computed from pixel color amplitudes and locations. Pixel amplitudes are globally thresholded as proposed in Ref. 15. Vertical rulings like those in Fig. 1(a) are eliminated by analyzing small eigenvalues in a local window as in Ref. 16. White noise patterns, such as shown as Fig. 1(d), are partially eliminated by open-and-close operations of mathematical morphology as proposed in Ref. 17.

We extract characters by computing their minimum-bounding box from the binary image, as shown in Fig. 2. The traditional reading order is right-to-left by column and top-down by character; therefore pages are first cut into columns and then each column is cut into individual characters. Overlapping characters in calligraphic works are far rarer than in contemporary handwriting. The few we found were segmented by dragging and dropping lines and boxes.

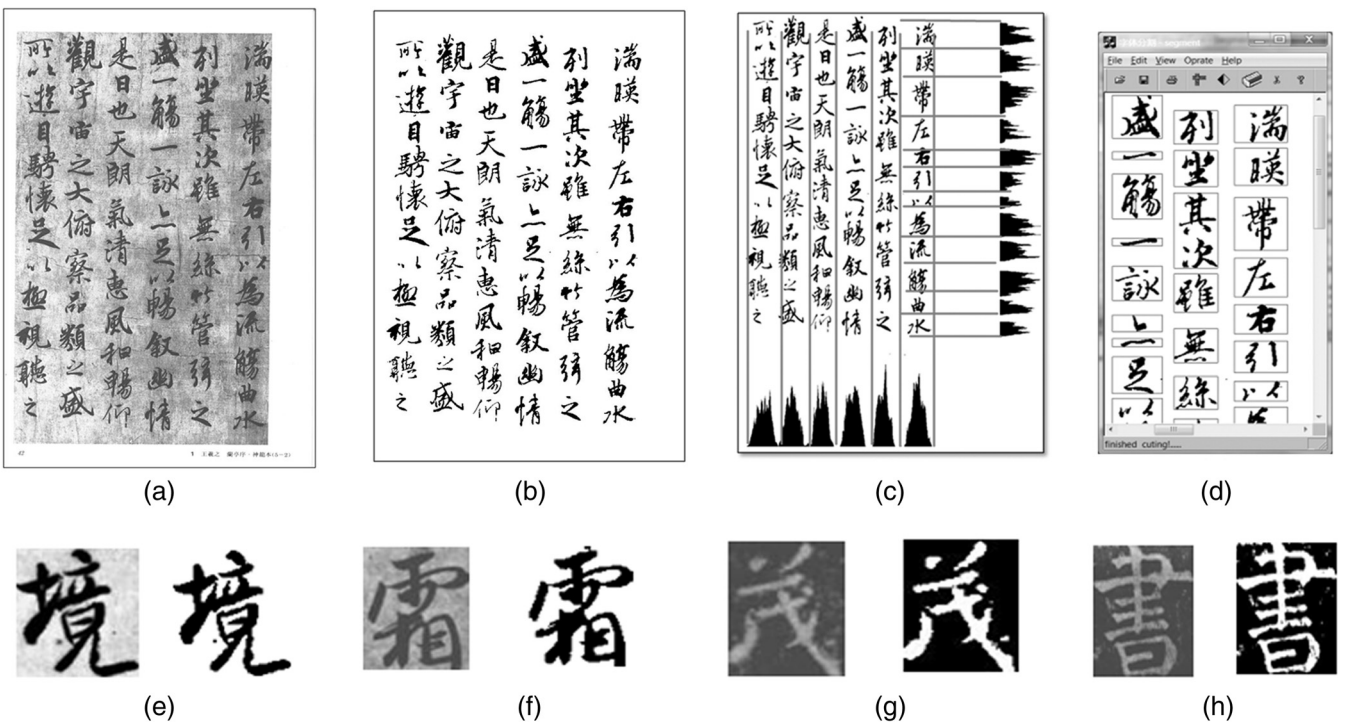
3.3 *Labeling and Metadata*

Character labels are necessary for training data and for reporting style identification accuracy. Therefore, all 8719 character images were manually labeled with 16-bit GB 2312 codes. The 32-bit Chinese National Standard GB 18030 is a superset of both ASCII and GB 2312 (GB 2312 codes are the 16 low-order bits of the corresponding GB 18030 codes). Unrecognized characters were labeled with null codes. Figure 3 shows our interactive labeling interface. Labeling in context is more accurate than labeling in isolation, especially for deformed character images and confusing obsolete characters.



(a) (b) (c) (d) (e)

Fig. 1 Samples of source pages: (a) vertical lines must be eliminated; (b) stamps interfere with segmentation; (c) the minimum bounding box of several characters must be inserted manually; (d) white noise patterns should be omitted; and (e) clean page that can be segmented automatically.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2 Segmentation: (a) source page image with a footer of printed page number and work title; (b) after layout analysis and page binarization; (c) page cut into columns and columns cut into characters; (d) segmentation interface showing some of the corresponding minimum bounding boxes; (e) and (f) are regular script segmented from the page shown in Figs. 1(a) and 1(b); (g) and (h) are from the stone rubbing shown in Figs. 1(c) and 1(d), which will be flipped for consistency with the black foreground of regular script.

Figure 4 shows our calligraphy database organization. The metadata includes catalog entries for the source page, title of the book, publication date, and author as well as the minimum bounding box, which preserves the character's original position. The latter is stored in the character table as top_x, top_y, bottom_x, and bottom_y.

The data is stored in an SQL Server 2008 database. Table 1 is a screenshot of the metadata for 20 characters. The first three columns on the left show the source of each character image. The next three columns are the manually labeled metadata. The four columns on the right are the top-left and bottom-right coordinates of the minimum bounding box, which indicates where the original character is located in the page. The column "file_path" is the file name of the individual character image.

4 Style Features

Handwriting is learned by copying formulary examples and stylized handwriting is called "calligraphy" regardless of origin. In literary and aesthetic studies, Chinese calligraphic

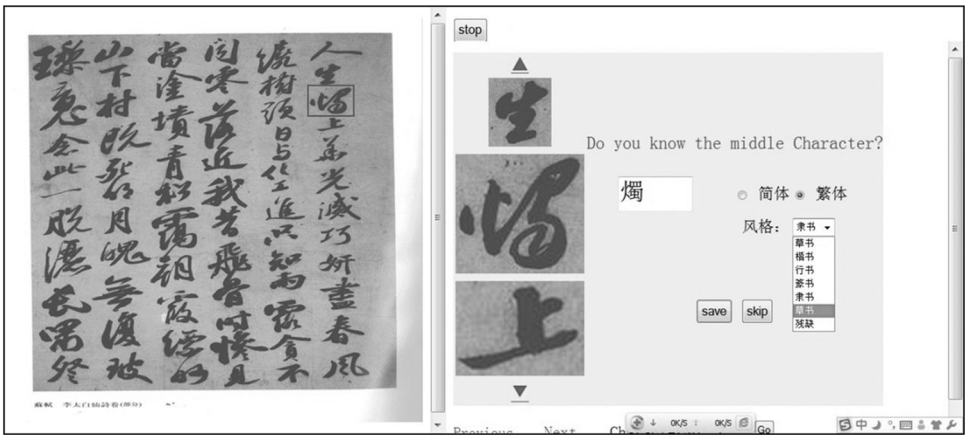


Fig. 3 Interactive character labeling interface: the central character image on the right is about to be labeled. The box in the page shown on the left indicates its original position in context.

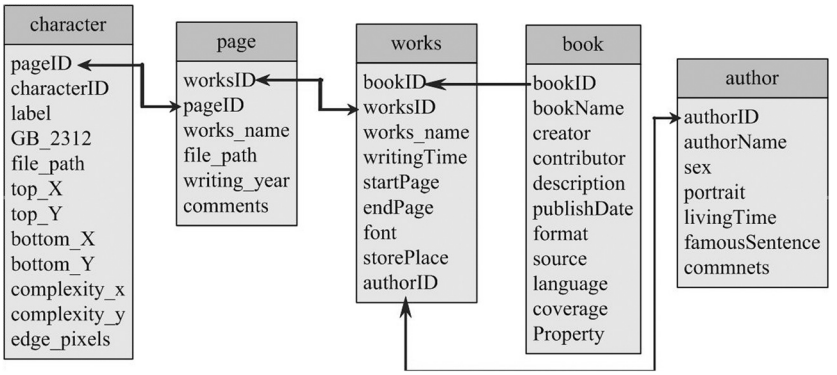


Fig. 4 The calligraphy data is organized into five tables: "book," "works," "page," "character," and "author." A book contains many works, which may have consecutive pages. These five tables are related by the primary keys: "bookID," "worksID," "pageID," "characterID," and "authorID."

Table 1 An extraction from the calligraphic database.

bookID	pageID_8	characterID	label	type	file_path	top_*x*	top_*y*	bottom_*x*	bottom_*y*
06100018	00000114	244398	及	1	00000114(709,125,833,309)	709	125	833	309
06100018	00000114	244399	青	1	00000114(731,322,821,497)	731	322	821	497
06100018	00000114	244400	阳	2	00000114(716,497,827,671)	716	497	827	671
06100018	00000114	244401	之	1	00000114(745,706,812,810)	745	706	812	810
06100018	00000114	244402	意	1	00000114(725,838,842,999)	725	838	842	999
06100018	00000114	244403	不	1	00000114(724,1025,841,1132)	724	1025	841	1132
06100018	00000114	244404	可	1	00000114(513,134,632,275)	513	134	632	275
06100018	00000114	244405	及	1	00000114(529,294,625,409)	529	294	625	409
06100018	00000114	244406	矣	1	00000114(525,416,629,540)	525	416	629	540
06100018	00000114	244407	其	1	00000114(525,677,638,866)	525	677	638	866
06100018	00000114	244408	一	1	00000114(524,920,626,959)	524	920	626	959
06100018	00000114	244409	往	1	00000114(297,72,417,224)	297	72	417	224

styles are divided into five main categories: "seal," "clerical," "standard (regular)," "running," and "cursive scripts." There are overlaps between these five different styles, with fuzzy borders like those between colors. These five styles can be combined to form a new style, much like three monochrome light beams can be added together to form a new color. We, therefore, define typical samples of the above five main calligraphic styles as the basis of a style space, where an arbitrary calligraphic style can be represented by a weighted combination of basis components. Figure 5 shows five characters, which have the same label but different main styles.

A person's writing style typically originates from a particular copybook with a certain style that may change somewhat over time and become more personalized. For example, one may write a character resembling both cursive script and running script. In these terms, calligraphic style means a combination of copybook systems. The copybooks are the classical calligraphy works in our database, with known styles assigned by specialized scholars.

Following the suggestion of Ref. 15, character images are mapped for display comparability to a unified size, of 45×45 pixels. We found, however, that both linear and nonlinear image normalization lose some information. Therefore, the features are extracted from the original image and then the features are normalized before similarity matching. Normalizing features are also faster than normalizing images.

The next three subsections describe the extraction of three kinds of style features from the horizontal strokes, the vertical strokes, and the characters. We evaluated 390 features of projections, stroke transect, left slant stroke, and right slant stroke. Most were discarded because they characterize GB-labels rather than styles.[9] We finally selected 24 features for style classification.

4.1 *Character Skeleton and Stroke Extraction*

The shapes and configurations of the strokes of a character define both its meaning and its style. The strokes are extracted based on the character skeleton to obtain stroke-level features.

4.1.1 *Skeleton character representation*

Hundreds of thinning and skeletonizing algorithms have been published because the accepted topological and geometric criteria are self-contradictory.[18,19] We chose the procedure based on the medial axis-transformation of Ref. 20. Individual pixels of a stroke skeleton bear little information about its style, but together they convey a stroke's length and direction. The chain code[21] is used to represent the stroke's trend. In order to facilitate tracing, the original skeletons are made one-pixel wide by deleting all but one of horizontally or vertically connected pixels in the eight-connected neighborhood of each skeleton pixel. This expedient provides a unique path to follow before meeting a fork, which is our main requirement.

Let $\text{seg}_i = \{\text{B}_i, \text{E}_i, \text{Snake}_i\}$ represent i'th skeleton stroke, where B_i and E_i are its beginning and end points, and Snake_i is its chain code. $\text{Snake}_i = c_1, c_2, \ldots c_3, \ldots, c_n$ of the U-shaped stroke of Fig. 6(a) is

```
6656665666666666656100010000010700000000-
70070700722223222232222222222
```

It begins with code 6 and ends with code 2. That is the trend: first down and eventually up. The 5's among the many 6's are not noise but twists in the writing process. They are important style attributes that reflect an individual calligrapher's personality and preference.

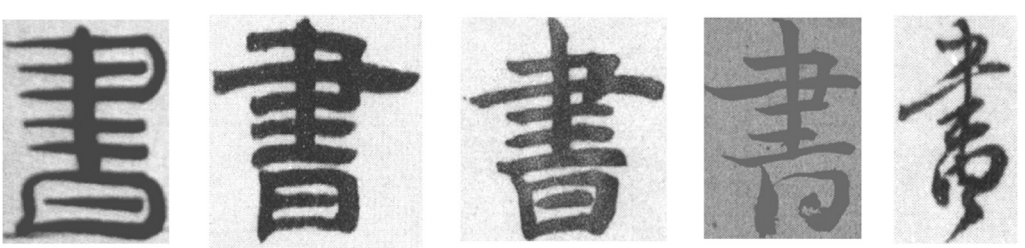
4.1.2 *Stroke extraction*

Skeleton pixels are classified into three types: "end pixel," "fork pixel," and "common pixel." An end pixel is a pixel connected only to one other pixel. A common pixel has two connected pixels, and a fork pixel is connected to at least three pixels. Figure 7 shows the three types of pixels.

An individual stroke starts from one end-pixel and ends at another end-pixel or a fork pixel. The operational question is which way to track a stroke when meeting a fork. Two observations resolve the quandary: (1) the character is traditionally written from left to right, from top to bottom, from outside to inside; (2) for ease of writing, directional continuity is preserved whenever possible. According to (1), the top left end pixel is chosen as the start. According to (2), when meeting a fork like the one in Fig. 7(c), $P_{\text{out},2}$ is chosen as the exit because it maintains the direction of the early part of the stroke, which enters from the left, without twisting the wrist, stops in the point P_{in} in Fig. 7(c). This is, in fact, how this character is written. Stroke order and direction are observed far more strictly in Chinese writing than in Western scripts.

4.2 *Stroke Direction Features*

The directional stroke features are listed in Tables 2 and 3 along with their denotation as feature vector components x_i. Horizontal and vertical strokes are the two simplest



(a) (b) (c) (d) (e)

Fig. 5 Five style representatives of the character that means "book:" (a) seal, (b) clerical, (c) standard, (d) runing, and (e) cursive scripts and represented by style #1, style #2, style #3, style #4, and style #5, respectively.

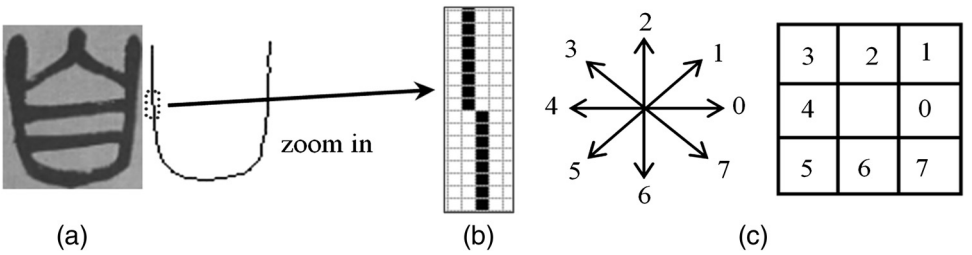


Fig. 6 Stroke chain code: (a) original character and the corresponding U-shaped stroke skeleton; (b) a magnified segment of the stroke skeleton; (c) orientation code and the corresponding map of the eight-neighborhood of a pixel.

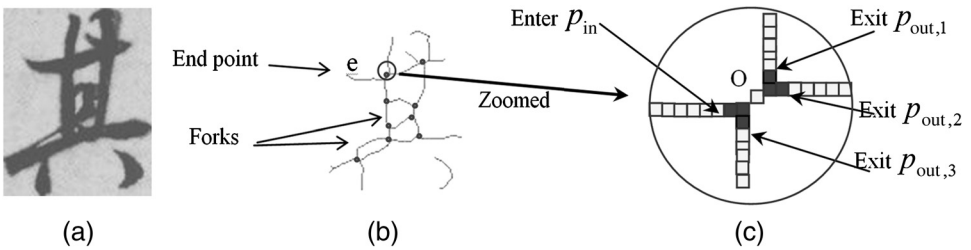


Fig. 7 Stroke extraction (a) original character image; (b) the corresponding skeleton, with forks marked with red dots; (c) a zoomed fork: pixels in light dark are common pixels, pixels in dark are forks that interfere with stroke tracking, six fork points with the connecting common point united as O.

strokes that are discriminative characteristics of all handwriting. Figure 8 shows two horizontal strokes, which are not completely horizontal and therefore do not have all-zero chain codes. They are the first strokes of two samples from style # 4 and style #1. The histograms of the chain-code show clearly, which way each stroke, is twisted.

The extracted skeletons of touching or overlapping strokes do not always represent exactly their printed version. The more cursive a character is, the more difficult it is to detect its horizontal and vertical strokes. The features in Table 2 indicate some style-related characteristics of the horizontal strokes. Feature f_{h_count} is the number of detected horizontal strokes, f_{h_slope} is their average gradient, f_{h_ratio} is the proportion of chain codes belonging to near-horizontal strokes, and f_{h2} and f_{h3} are sensitive measures of left and right slants. For example, a horizontal stroke of style #1, such as in the second row of Fig. 8 is more horizontal than those from style #4 in the first row. The last pair of features, f_{h_head} and f_{h_tail}, help to detect serif-like protrusions at the beginnings and the ends of horizontal strokes. Analogous features for vertical strokes (except for stroke-head and stroke-tail) are listed in Table 3.

Table 2 Bag of seven visual style words based on horizontal strokes.

Feature	Symbol	Component
Number of detected horizontal strokes	f_{h_count}	x_1
Average gradient of the horizontal stroke	f_{h_slope}	x_2
Average ratio of horizontal chain-code to total chain-code	f_{h_ratio}	x_3
Average ratio of left slant chain-code to total chain-code	f_{h2}	x_4
Average ratio of right slant chain-code to total chain-code	f_{h3}	x_5
Average gradient of top 1/8 chain-code	f_{h_head}	x_6
Average gradient of bottom 1/8 chain-code	f_{h_tail}	x_7

4.3 *Stroke Width Features*

Calligraphy is written with a soft brush. Pressing on the brush thickens or widens the stroke; therefore, variation in width tends to reflect pen pressure. Although $A/(A-Q)$, where A is the count of the of all the foreground pixels and Q is the count of 2×2 squares in the image, is a good approximation of average stroke width,[22] for style characterization, some measure of its variability is also required. Figure 9 illustrates our computation of stroke width at each pixel of the skeleton. A disk is centered on the skeleton pixel

Table 3 Bag of five visual style words on vertical strokes.

Feature	Symbol	Component
Number of detected vertical strokes	f_{v_count}	x_8
Average gradient of the vertical stroke	f_{v_slope}	x_9
Average ratio of vertical chain-code to total chain-code	f_{v_ratio}	x_{10}
Average ratio of left slant chain-code to total chain-code	f_{v2}	x_{11}
Average ratio of right slant chain-code to total chain-code	f_{v3}	x_{12}

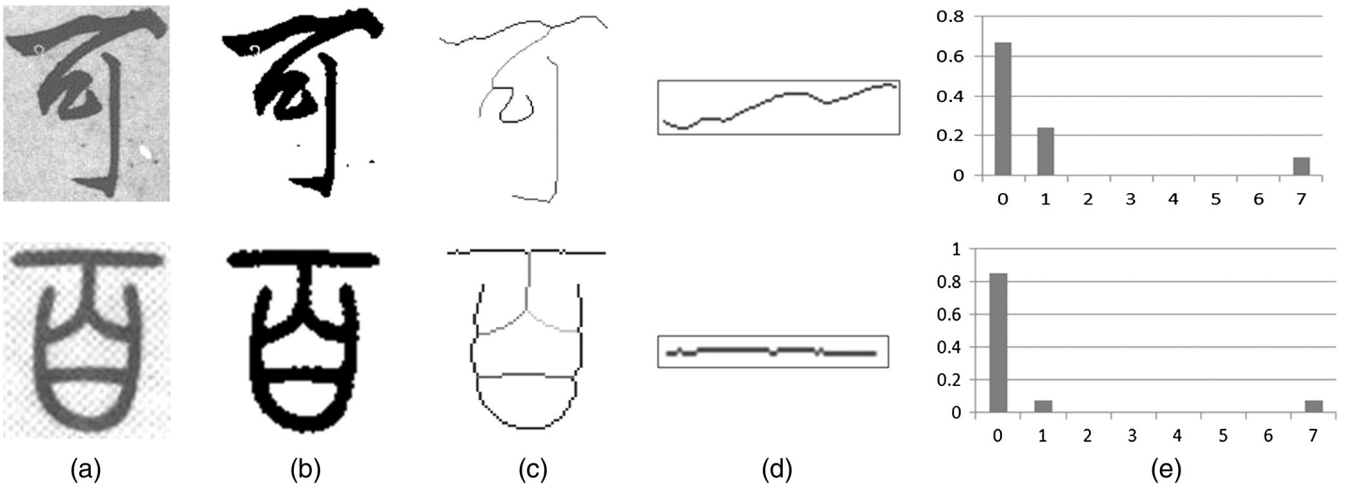


Fig. 8 Stroke extraction and stroke characterization. (a) Original character images; (b) binary images; (c) extracted skeletons with different strokes marked with different shades of gray; (d) isolated first horizontal strokes; (e) histograms of codes of horizontal strokes with the *x*-axis representing the chain-code direction.

with an initial radius of 1 pixel, then its radius is increased until the ratio of foreground to background area covered by the disk in the original image drops below the empirically determined $\theta = 0.8$. The average and standard deviation of stroke width within character are selected as style features. They can be calculated as

$$f_{\text{ave\_width}} = \frac{1}{n}\sum_{i=1}^{n} d_i, \tag{1}$$

$$f_{\text{sig\_width}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - f_{\text{ave\_width}})^2}, \tag{2}$$

where n is the total number of pixel on the character skeleton and d_i is the radius for pixel i.

Characters of seal style (style #1), and sometimes also of clerical style (style #2), are written slowly and have uniform stroke width, as shown in Fig. 10(a). But running and cursive characters are written with emotion, thus some strokes touch and even overlap, resulting in very large stroke width; while other strokes have a threadlike ending because the brush is lifted and barely touches the paper, as shown in Fig. 10(b). These characteristics are revealed by the visual style words $f_{\text{max\_width}}$ and $f_{\text{min\_width}}$ in Table 4.

4.4 *Mass Distribution Attributes*

The remaining features listed in Table 4 do not depend on skeletonization or stroke extraction. The ratio of foreground to background area $f_{\text{area\_ratio}}$ of characters is a measure of visual weight in calligraphy just as in Western typeface design. The aspect ratio (height-to-width) $f_{\text{aspect\_ratio}}$ is also valuable because some calligraphers consistently favor tall characters while others like fat characters.

Fig. 9 Stroke thickness. Red pixels are the skeleton and the gray pixels are the foreground of a stroke. *d* is the radius of the maximal foreground disk, centered on a skeleton pixel marked with a darker dot.

$$f_{\text{area\_ratio}} = \text{foreground/background}, \tag{3}$$

$$f_{\text{aspect\_ratio}} = \text{height/width}. \tag{4}$$

Geometric moments are used for measuring the centroid and the stroke extension. Let M and N be the width and the height of a calligraphy character image $f(x, y)$. Then, a character's (p, q)th order moment is defined as

$$m_{pq} = \sum_{x=0}^{x=M-1}\sum_{y=0}^{y=N-1} x^p y^q f(x, y). \tag{5}$$

The centroid $(\bar{x}, \bar{y})$ is

$$f_{\bar{x}} = \frac{m_{10}}{m_{00}}, \qquad f_{\bar{y}} = \frac{m_{01}}{m_{00}}. \tag{6}$$

The central moments are

$$u_{pq} = \sum_{x=0}^{x=M-1}\sum_{y=0}^{y=N-1} (x - \bar{x})^p (y - \bar{y})^q f(x, y). \tag{7}$$

Values of u_{pq} *with* odd values of p or q indicate asymmetries of foreground distribution that may differentiate styles. We divide u_{30} (horizontal skewness) into left-of-center (negative) part ${u_{30}}^-$ and right-of-center (positive) part ${u_{30}}^+$. Then, $|{u_{30}}^-| > |{u_{30}}^+|$ indicates that the calligrapher pressed the brush harder on the left than on the right of the character. The normalized horizontal and vertical stress variations $f_{\text{stress\_}x}$ and $f_{\text{stress\_}y}$ can thus be defined as

$$f_{\text{stress\_}x} = \frac{{u_{30}}^+}{{u_{30}}^+ + {u_{30}}^-}, \qquad f_{\text{stress\_}y} = \frac{{u_{03}}^+}{{u_{03}}^+ + {u_{03}}^-}. \tag{8}$$

(a) (b)

Fig. 10 Examples of different thickness variation: (a) uniform thickness and (b) variable thickness.

Table 4 Bag of 12 visual style words on character.

Feature	Symbol and equation	Component
Average of stroke width	f_{ave_width} [Eq. (1)]	x_{13}
Standard deviation of stroke width	f_{sig_width} [Eq. (2)]	x_{14}
Average top 1/15 maximum and bottom 1/15 minimum width	f_{max_width} and f_{min_width}	x_{15}, x_{16}
Binary ratio	f_{area_ratio} [Eq. (3)]	x_{17}
Aspect ratio h/w	f_{aspect_ratio} (Eq. (4)]	x_{18}
Centroid	$f_{\bar{x}}$ and $f_{\bar{y}}$ [Eq. (6)]	x_{19}, x_{20}
Stress variation	f_{stress_x} and f_{stress_y} [Eq. (8)]	x_{21}, x_{22}
Slant balance	f_{slant_x} and f_{slant_y} [Eq. (9)]	x_{23}, x_{24}

Two additional features are analogously based on the mixed central moments u_{21} and u_{12}:

$$f_{\text{slant\_}x} = \frac{{u_{21}}^+}{{u_{21}}^+ + {u_{21}}^-}, \qquad f_{\text{slant\_}y} = \frac{{u_{12}}^+}{{u_{12}}^+ + {u_{12}}^-}. \tag{9}$$

The style model has $7 + 5 + 12 = 24$ features, shown as in Tables 2–4. Each feature is treated as a visual word, with a value and an origin position.

5 Style Classification and Quantification

After the bag of visual words (histograms of feature values) is extracted, the next step is to learn its statistical distribution and construct the style model. For style classification, the visual words are assigned to components of a feature vector. The probabilities of a feature vector belonging to each of the five styles constitute a measure of style quantification.

5.1 *Style Model*

The five main styles (seal, clerical, standard, running and cursive script) are denoted as $\omega 1$, $\omega 2$, $\omega 3$, $\omega 4$, and $\omega 5$. The training sample consists of M_k samples of each style. The notations used are

Samples $c_{j(k)}$, $j(k) = 1,2,\ldots,M_k$
Styles ω_k, $k = 1,2,\ldots,K$ ($K = 5$ here)
A sample $c_{j(k)}$ of style ω_k has features $x_{i,j(k)}$, $i = 1,2,\ldots D$ ($D = 24$ here)
The training feature vectors of style k are $\mathbf{x}_{j(k)}$, $j(k) = 1,,2,\ldots,M_k$

Averages over training samples of each feature:

$$\mu_{i,k} = \frac{1}{M_k}\sum_{j(k)=1}^{M_k} x_{i,j(k)}, \qquad \boldsymbol{\mu_\kappa} = [\mu_{1,k}, \mu_{2,k}, \mu_{3,k}, \mu_{4,k}, \mu_{5,k}]. \tag{10}$$

Average feature values over style classes:

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K}\mu_{i,k}, \qquad \sigma_i^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_{i,k}^2. \tag{11}$$

Covariance matrix of style k:

$$C_k = \sum_{j(k)=1}^{M_k}(\mathbf{x}_{j(k)} - \boldsymbol{\mu}_k)(\mathbf{x}_{j(k)} - \boldsymbol{\mu}_k)'. \tag{12}$$

Style features will not work for GB label classification because the goal of style classification is to find the differences between character images regardless of GB label, while the goal of GB label classification is to find differences regardless of style. Thus, style features are designed to be sensitive to exactly the kind of variation that is considered noise in optical character recognition-type classification. A writer's character-to-character variations matter far more in transcription than in style classification because we never train a classifier on same-GB-label characters.

5.2 *Style Classification*

A character of unknown style is assigned to one of the five main styles with a linear classifier trained on samples of characters with style labels. The GB labels are retained only for further study of the experimental results.

We use a simple linear classifier based on the usual assumption of Gaussian feature distributions. Since the covariance between pairs of features depends far more on the designated features than on the characters' class and style, we use the same covariance matrix for each style class, i.e., the pooled covariance of the training set. Then, the conditional probability of a character with feature vector x given style k is

$$p_k(\mathbf{x}|\omega_k) = \mathbf{w}_k'\mathbf{x}, \tag{13}$$

where the style weight vectors $\mathbf{w}_k = C^{-1}\boldsymbol{\mu}_k$ are computed from the inverse covariance matrix C^{-1} and the average

style feature vectors $\boldsymbol{\mu}_k$ of the training set. The sample character is assigned to the style $\hat{k}$, where $\hat{k} = \text{argmax}\, P_k p_k(\mathbf{x}|\omega_k)$. The empirical prior style probability P_k, required for maximum *a posteriori* (MAP) classification, is based on the style frequencies in the training set.

The accuracy of the classification depends on how well the feature vectors and the style frequencies of the training set correspond to those of the test set. Therefore, the experimental results depend on how the entire data is divided into training and test sets. We investigate training based on three different sampling schemes.

Split 1: The customary sampling method is random partitioning of the data into training and test sets. We adopt this as our baseline and call it Split_1. The results of N_{expt} random partitions are averaged for statistically significance. Each partition is generated by taking for the training set the samples corresponding to the first N_{train} integers of a pseudorandom permutation of N_{char}, the total number of samples, and the remaining $N_{\text{test}} = N_{\text{char}} - N_{\text{train}}$ samples as the test set.

Split 2: Our data consists of many works, each by a single author in a single style. Therefore, we would expect that partitioning each work into training and test samples would raise the accuracy. This is our Split_2, obtained by generating random permutations to sample each work. Here, again, we average N_{expt} partitions with different seeds for the pseudorandom number generator.

Split 3: In the applications we have in mind, there is no guarantee that a new character belongs to one of the works in the training set. Therefore, neither Split_1 nor Split_2 is realistic for predicting how well the style of a previously unseen character can be predicted. With Split_3, every work is randomly assigned to either the training set or the test set. Therefore, characters are tested under realistic conditions, without any samples of their own work included in the training set. This means, however, that in contrast to Spit_1 and Split_2, different random partitions will yield different sizes of training and test sets (because assigning many large works to the training set will decrease the size of the test set). This effect is significant in our data because of the highly nonuniform distribution of the number and size of works of different styles.

The size of the training set is controlled by the T_{fraction} parameter. For Split_1, T_{fraction} is used to split random permutations of $1,2,\ldots,N_{\text{char}}$. For Split_2, it splits the permutation vectors corresponding to the number of characters in each work. For Split_3, T_{fraction} splits permutations of $1,2,\ldots N_{\text{works}}$, the number of works in the data. We report results with $T_{\text{fraction}} = 1/3$ and $T_{\text{fraction}} = 2/3$. The classification results from the three methods of splitting the data into training and test sets are cross-validated in the same way. Each selection of samples for training, testing and voting is replicated 10 times with different random permutations.

Instead of trying to find out the style of a single isolated character, often the style of a fragment of a work, or perhaps that of an entire work, is of interest. In that case, we can exploit the fact that all the available characters must be assigned the same style label. The simplest way to do this is to classify each of N_{vote} participating characters and then to select the "mode" (peak) of the resulting vote distribution over styles. This is equivalent to majority voting with ties broken in favor of the lower style number. In Sec. 6, we report voting results with values of $N_{\text{vote}} = 1$, 3, and 7. If the classifier gives a ranking of style classes instead of only the class of the highest posterior probability, then the Borda Count can be used instead of majority voting.

5.3 *Style Quantification*

An unknown calligraphic character may have one distinguished style or represent a mixture of several styles. Therefore, identifying a composite style requires computing the similarity between the unknown character and each of five main calligraphy categories.

In the section above, the classification is "crisp:" every character is assigned to a unique style. We can, instead, provide a five-element style vector for every character. The unity-length character-style vector is

$$\boldsymbol{\gamma} = [P_1 p_1(x|\omega_1), P_2 p_2(x|\omega_2), P_3 p_3(x|\omega_3), P_4 p_4(x|\omega_4), P_5 p_5(x|\omega_5)]. \quad (14)$$

If the empirical probabilities are not expected to represent the prior probabilities of the characters in the test set, then we set $P_k = 1/5$ for $k = 1,\ldots,5$. The style weights can be obtained with any of our three sampling schemes.

A measure of the style cohesiveness of any set of characters is given by the eigenvalues of the covariance matrix of their feature vectors. We defer experiments on composite style quantification to future work.

6 Experiments

6.1 *Data Set*

Two calligraphy books were scanned page by page at 600 dpi (23.6 lpi) and kept in both DjVu format and TIFF format by CADAL[1] scanning center. Ignoring the printed covers and the printed illustration pages without calligraphy, 259 calligraphy page images were segmented into 8279 individual characters. Figure 11 gives number of characters per page.

6.2 *Size and Label*

The width and height of original scanned calligraphy page images and of segmented individual character images are shown in Table 5. Since they were scanned at 600 dpi, the

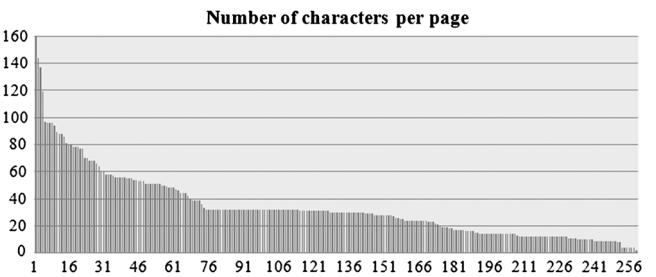


Fig. 11 Calligraphy character density per page.

Table 5 Size of original calligraphy image.

Measurement	Statistics	Page image	Character image
Height	Average	1280 pixels	91 pixels
	Max	1289 pixels	481 pixels
	Min	873 pixels	11 pixels
Width	Average	892 pixels	90 pixels
	Max	1284 pixels	331 pixels
	Min	868 pixels	16 pixels

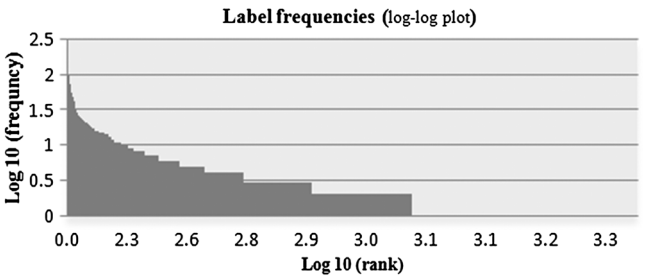


Fig. 12 Frequency distribution of character labels.

average height of each page is about 27 cm and its width is about 19 cm.

There are 1908 labels that occur in the corpus. The most frequent label is "zhi" (之), which has 183 instances, but 755 labels occur only once. Figure 12 shows a log-log plot. Zip's law for English words (that we compare with Chinese character statistics) states that in any long English document, half the words appear only once. In our collection, the 755 singletons are fewer than half of the labels because they exclude many characters, which occur only once. (71 ancient characters have been seen only once in historical works, are never used in modern literature, and we cannot recognize and label them).

According to Zipf's Law, the frequency of any English word is inversely proportional to its rank in the frequency table. Our label frequencies fall off initially more rapidly than $1/n$ because they cover works from the Qing dynasty (221 BC to 206 BC) to the Ming Guo period (1912 AD to 1949 AD). Characters that meant the same changed from dynasty to dynasty, which forces less repetition than seen in contemporary text.

6.3 *Style Classification*

Several hundred images, mainly of running style, were added to the corpus for the style classification and quantification experiments. While labeling individual character images, we found that 8285 instances clearly belong to a single style and we annotated them accordingly. The style populations are nonuniform, ranging from 650 characters of style #4 to 2867 characters of style #2.

Our $3 \times 2 \times 3 \times 10$ experimental design compares the three methods of splitting the data into training and test sets, two ratios of ½ and 2 of the size of the training set to the size of the test set, voting with 1, 3, and 7 participants, and tenfold cross-validation. The 24 style features used were described in Sec. 4 and the classifier in Sec. 5. The averaged classification accuracy and the standard deviations of these experiments are shown in Table 6, where T_{fraction} is the split ratio, N_{vote} is the number of voters, N_{train} is the number of training characters, N_{test} is the number of test characters, Single rec (%) is the percentage of correctly recognized test characters with standard deviation STD_s among the ten random permutations, and Vote rec% is the correct recognition under the voting scheme with standard deviation STD_v.

From Table 6, we may observe that

Table 6 Classification results.

	T_{fraction}	N_{vote}	N_{train}	N_{test}	Single rec %	STD_s	Vote rec%	STD_v
Split_1 split data	0.33	N/A	2734	5551	69.1	0.52	N/A	N/A
	0.67	N/A	5551	2734	70.3	0.61	N/A	N/A
Split_2 split each work	0.33	3	2696	5589	71.3	0.70	79.2	0.90
		7	same	same	same	same	88.2	0.80
	0.67	3	5452	2833	71.7	0.60	80.0	1.0
		7	same	same	same	same	88.5	1.1
Split_3 split by work	0.33	3	N/A	N/A	69.6	1.8	77.4	1.7
		7	N/A.	N/A	same	same	78.5	2.0
	0.67	3	N/A	N/A	70.5	1.3	80.6	1.4
		7	N/A	N/A	same	same	87.3	2.0

Table 7 Confusion table. Rows and columns are true and assigned styles.

	Style #1	Style #2	Style #3	Style #4	Style #5	Total
Style #1	528	0	1	0	0	529
Style #2	1	883	34	0	0	918
Style #3	8	28	452	3	2	493
Style #4	0	1	2	155	53	211
Style #5	20	3	30	118	318	489
Total	557	915	519	276	373	2640

Table 8 Style-specific performance.

Style	Precision	Recall	*F*-ratio
Style #1	0.95	1.00	0.97
Style #2	0.97	0.96	0.96
Style #3	0.87	0.92	0.89
Style #4	0.56	0.73	0.64
Style #5	0.85	0.65	0.74

1. Even with a random split, most works are sufficiently represented in the training set. Split 1, a random partition of the data, yields barely lower classification accuracy (69.1% and 70.3%) than Split 2 (71.3% and 71.7%), which splits every work into training and test samples.
2. Voting samples of the same work, which is of course possible only if the works are identified, yields a significant improvement: 70.5% to 80.6% with three voters, and 70.5% to 87.3% with seven voters (with Split 3 and 67% of the samples used for training).
3. Because there are so many samples per class, a larger training sample has little effect on the accuracy. Without voting, only with Split 1 does the accuracy increase by more than 1% when the size of the training set is doubled. Training sets much smaller than 1/3 of the data do increase the error, as expected.

Each run produces a confusion table. Table 7 is the result from first random iteration for Split 3, with two-thirds of the samples assigned to the training set, and 7 voters. We see, for example, that 118 of the 489 test samples of Style #5 were erroneously assigned to Style #4, and 53 samples of Style #4 were assigned to Style #5. Confusions between running and cursive styles account for 171 of the 304 errors. The binary classification performance measures of precision, recall and *F*-ratio can be computed directly from the confusion table and are shown in Table 8.

The top 12 visual characters that represent the five main styles are shown in Fig. 13, ranked by their MAP probability [Eq. (14)]. The number below each character image is the corresponding characterID. With our calligraphy data organization, the page image of any character can be easily displayed for visual verification of its style. Figure 14(a) shows the popup window when the last character in the first row of Fig. 13 is clicked and Fig. 14(b) shows the pop up window when the third character in the fourth row is clicked. This allows inspecting many other characters of the same style as the query. If the corresponding characterID, such as the ninth in the first row is clicked, then the values of the style components will be shown as in Fig. 15(a).

6.4 *Style Quantification*

The five posterior probabilities in Eq. (14) computed during classification are regarded as component values like RGB components for an arbitrary color. Figure 15 shows two examples of how an individual character of arbitrary style is described by five style components. The character shown in Fig. 15(a) is predominantly seal style (style #1). The character in Fig. 15(c) is a hybrid mainly of standard and running styles (#3 and #4). An art expert could tell that the character

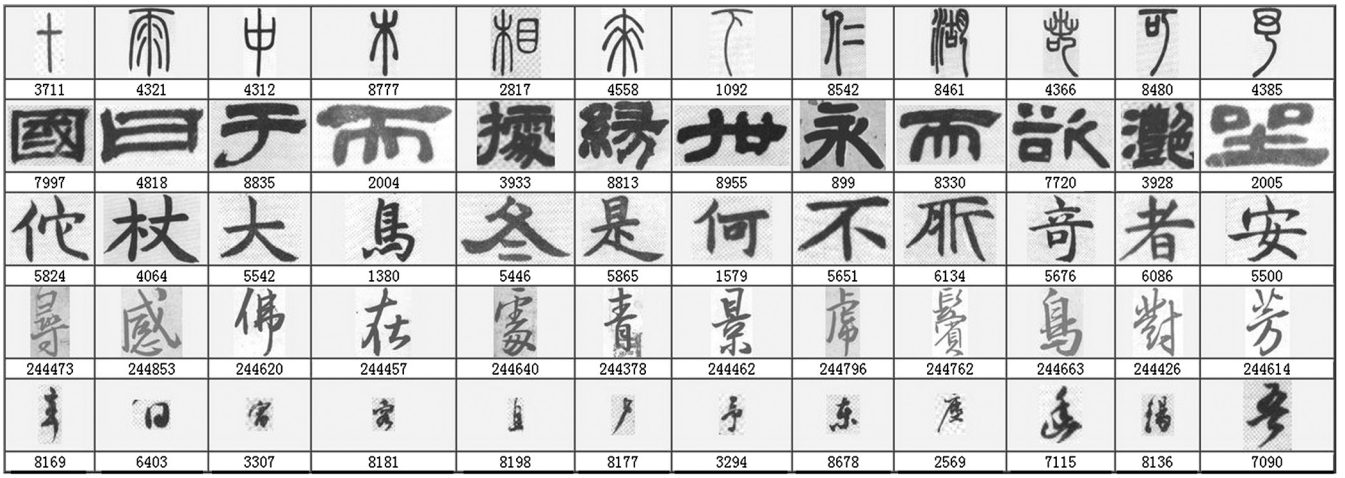
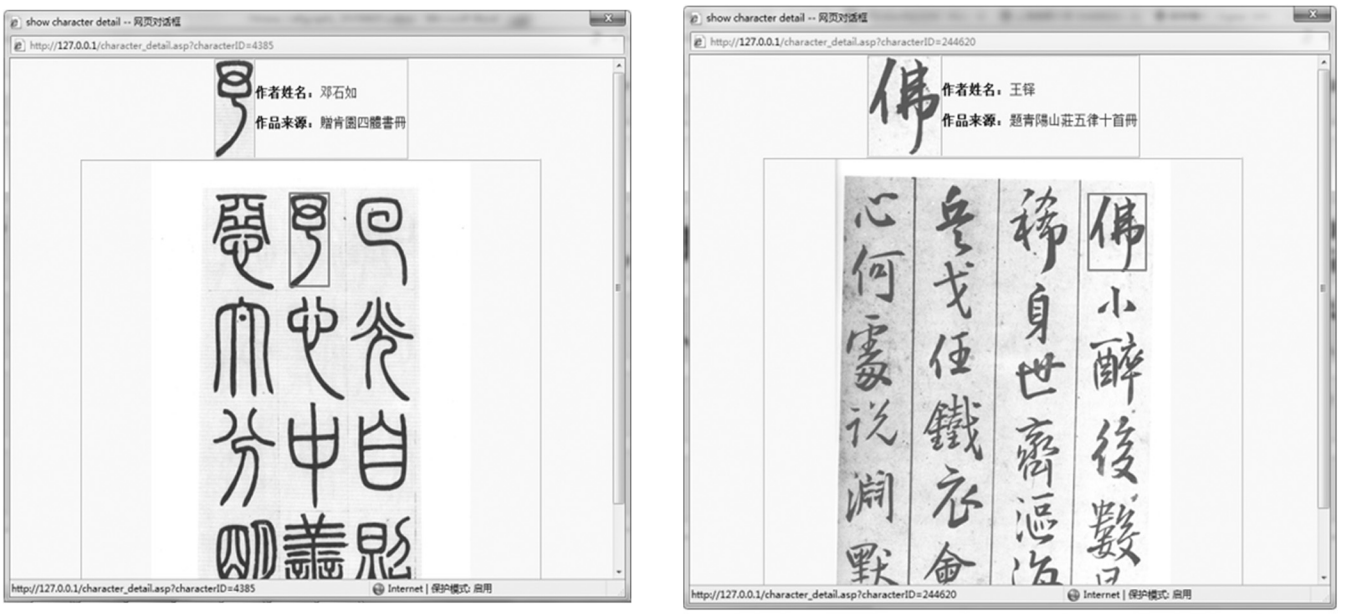
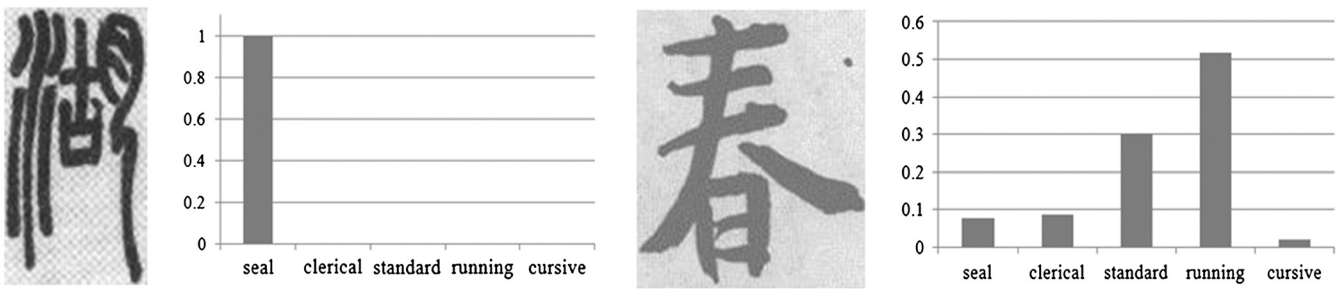


Fig. 13 Screenshot of top 12 character instances for five main styles from first to fifth row: "seal script," "clerical script," "standard script," "running script," and "cursive script."



(a) (b)

Fig. 14 Screenshot of a character query, including the (a) printed metadata of name of the author and the title of the work, followed by the (b) page image with a minimum bounding box showing the original character.



(a) CID=8461 (b) (c) CID=244850 (d)

Fig. 15 Style quantification: (a) and (c) are arbitrary style characters; (b) style components values (0.995, 0.003, 0.001, 0, 0) of character shown in (a); (d) style components (0.076, 0.085, 0.303, 0.518, 0.019) of (c).

of Fig. 15(a) was written in seal style, yet would find it difficult to tell the style of the other character because it is a mixture.

7 Conclusion and Future Work

The current set of 8791 character images (8285 with style labels in addition to GB codes) is available now on the CADAL web site.[1] The reported experiments explore computational aspects of calligraphic style. The proposed classification and quantification rely on specialized local and global visual style features, but the sampling and voting methods may be useful in other multilabel object recognition tasks as well. With good style features and a representative training set including characters from the same work, the style of a single arbitrary character can be recognized with over 71% accuracy. The single-voter error is about three times lower for seal, standard, and clerical scripts than for running and cursive scripts. The accuracy can be raised considerably by voting the styles assigned to several characters from the same work.

In many applications, however, there is no guarantee that the work from which the unknown character originates was represented in the training set of the classifier. In that case, voting is essential. Even a sample of three characters of the same work helps, though seven samples is much better. Experiments not reported here show that the accuracy keeps rising when more voters are added.

Hybrid styles can be quantified by their posterior probabilities. This is useful because the evolution of calligraphic styles is a continuum rather than a series of steps from one style to another. While this project addresses only historical styles, works in new styles are presented by artists at yearly expositions. Potential applications of style classification and quantification include:

1. Interactive style recognition: recognition of calligraphic images (possibly snapped by a phone camera

from a scroll or other artifact) for cultural purposes, with an interface similar to that shown in Fig. 14.

2. Duplicate detection: Duplicate check can be performed on metadata in a calligraphic database before deciding priorities for scanning new troves of works on paper.
3. Style relationship discovery: Skill in calligraphy was always acquired and improved by copying available works. Therefore, styles of related calligraphers exhibit similarity even though they may change gradually over time. Discovering style relationships may help to identify the author and date of historical calligraphy.
4. Forgery detection: Famous calligraphy works are frequently counterfeited because of their high price. Our system can be further developed to identify the particular style of an individual author. Thus, when given suspected calligraphy, one can compare it to calligraphy from the putative author that is already in the database. The system can give a probability of how likely the style of the suspected work belongs to the putative author.

Acknowledgments

This work was supported by the Science & Technology Program of Shanghai Maritime University (Grant No. 20130467), and by the National Nature Science Foundation of China (Grant No. 61303100). We also gratefully acknowledge the support of CADAL for scanning calligraphy books.

References

1. Chinese Calligraphy Service of CADAL, [EB/OL], http://www.cadal.zju.edu.cn/NewCalligraphy (2015).
2. Universal Digital Library Web Site, [EB/OL], http://www.ulib.org (2015).
3. S. N. Srihari, "Computational Methods for Handwritten Questioned Document Examination," Final Report, Award Number: 2004-IJ-CX-K050, U.S. Department of Justice, http://www.ncjrs.gov/pdffiles1/nij/grants/232745.pdf (2010).
4. M. Panagopoulous et al., "Automatic writer identification of ancient Greek inscriptions," *IEEE Trans. Pattern Recognit. Mach. Intell.* **31**(8), 1404–1414 (2009).
5. M. S. Azmi et al., "Arabic calligraphy identification for Digital Jawi Paleography using triangle blocks," in *International Conf. on Electrical Engineering and Informatics*, Malaysia, pp. 1–5 (2011).
6. I. Bar-Yosef et al., "Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents," *Int. J. Doc. Anal. Recognit.* **9**(2–4), 89–99 (2007).
7. National Digital Library of China-Ancient books, [EB/OL], http://www.nlc.gov.cn/newen/atbooks (2015).
8. L. Yang and L. Peng, "Local projection-based character segmentation method for historical Chinese documents," *Proc. SPIE* **8658**, 86580O (2014).
9. X. Zhang and Y. Zhuang, "Dynamic time warping for Chinese calligraphic character matching and recognizing," *Pattern Recognit. Lett.* **33**, 2262–2269 (2012).
10. J. H. Yu and Q. S. Peng, "Realistic synthesis of cao shu of Chinese calligraphy," *Comput. Graphics* **29**(1), 145–153 (2005).
11. K. Yu, J. Wu, and Y. Zhuang, "Style-consistency calligraphy synthesis system in digital library," in *Proc. of 9th Annual Int. ACM/IEEE Joint Conf. on Digital Libraries*, pp. 145–152 (2009).
12. S. Xu et al., "Automatic generation of Chinese calligraphic writings with style imitation," *IEEE Intell. Syst.* **24**(2), 44–53 (2009).
13. W. Lu, Y. Zhuang, and J. Wu, "Latent style model: discovering writing styles for calligraphy works," *J. Vis. Commun. Image Represent.* **20**(2), 84–96 (2009).
14. X. Zhang and G Nagy, "Style comparison in calligraphy," *Proc. SPIE* **8297**, 82970O (2012).
15. J. Zhang et al., "Denoising of Chinese calligraphy tablet images based on run-length statistics and structure characteristic of character strokes," *J. Zhejiang Univ. Sci. A* **7**(7), 1178–1186 (2006).
16. I. T. Phillips and A. K. Chhabra, "Empirical performance evaluation of graphics recognition systems," *Pattern Anal. Mach. Intell.* **21**(9), 849–870 (1999).
17. P. R. A. Peters, "A new algorithm for image noise reduction using mathematical morphology," *IEEE Trans. Image Process.* **4**(5), 554–568 (1995).
18. U. Eckhart and G. Maderlechner, "Invariant thinning," *Int. J. Pattern Recognit. Artif. Intell.* **7**(5), 115–144 (1993).
19. P. Tarabek, "Performance measurements of thinning algorithms," *J. Inf. Control Manage. Syst.* **6**(2), 125–132 (2008).
20. K. B. Kegel and A. Krzyzak, "Piecewise linear skeletonization using principal curves," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(1), 59–74 (2002).
21. H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Trans. Electron. Comput.* **EC-10**(2), 260–268 (1961).
22. M. R. Bartz, "The IBM 1975 optical page reader part II: video thresholding system," *IBM J. Res. Dev.* **12**(5), 354–363 (1968).

Xiafen Zhang, received her PhD at Zhejiang University in 2006 and she is a lecturer in the College of Information Engineering, Shanghai Maritime University (http://cie.shmtu.edu.cn/archives/category/faculty/lecturer). Since she joined the China-America Digital Academic Library project in 2004, she has been interested in and focused on document image analysis and pattern recognition.

George Nagy received his PhD at Cornell University in 1962 and, co-authored with Dick Casey, published the first Chinese character recognition paper in the world. He has been working on document analysis and pattern recognition for dozens of years (http://www.ecse.rpi.edu/~nagy).