Procs. Document Recognition & Retrieval, SPIE IS&T Electronic Imaging Conference, San Francisco, February 2016

Training a calligraphy style classifier on a non-representative training set

George Nagy ***Rensselaer Polytechnic Institute, Troy NY USA and Xiafen Zhang, Shanghai Maritime University, P.R.China***

Abstract

Calligraphy collections are being scanned into document images for preservation and accessibility. The digitization technology is mature and calligraphy character recognition is well underway, but automatic calligraphy style classification is lagging. Special style features are developed to measure style similarity of calligraphy character images of different stroke configurations and GB (or Unicode) labels. Recognizing the five main styles is easiest when a style-labeled sample of the same character (i.e., same GB code) from the same work and scribe is available. Even samples of characters with different GB codes from same work help. Style classification is most difficult when the training data has no comparable characters from the same work. These distinctions are quantified by distance statistics between the underlying feature distributions. Style classification is more accurate when several character samples from the same work are available. In adverse practical scenarios, when labeled versions of unknown works are not available for training the classifier, Borda Count voting and adaptive classification of style-sensitive feature vectors seven-character from the same work raises the ~70% single-sample baseline accuracy to ~90%.

1. Introduction

In countries with recorded cultural histories of thousands of years, there is much popular and scholarly interest in calligraphy. Applications of calligraphic style recognition are more cultural, artistic and educational than those of font and writer recognition. Our source books of calligraphy are scanned at 600 dpi and 24-bit color by the China Academic Digital Associative Library (CADAL) project [1], which is an active partner part of the Universal Digital Library (UDL) [2].

Style classification has obvious similarities to font and writer recognition. Conventional OCR is normally concerned only with the alphanumeric label of each character image and of its neighbors. In font and writer identification both the character label and the font or writer label play a part. In our calligraphic database, each pattern has three labels

(1) a GB-label selected from the 16-bit Chinese character subset of the Unicode equivalent to the Western ASCII codes; (2) a WorkID assigned arbitrarily to a printed copy of each original archival work created by a single scribe; and (3) a StyleID ranging from 1 to 5 that corresponds to the top level of the traditional calligraphic style taxonomy: *seal, clerical, standard (regular), running,* and *cursive*.

The top level of an analogous taxonomy consists of the *serif, sans-serif, cursive, fantasy,* and *monospace* font families. This traditional printing and publishing terminology was adopted by the World Wide Web Consortium (W3C) for the Cascading Style Sheet (CCS) specifications. Within each font there are size, weight, slant, aspect ratio and kerning variants. Cursive handwriting (as opposed to block lettering) lacks such well-defined divisions, but experts recognize *Copperplate* (or *English round hand*, emulated in the original printed version of the United States Declaration of Independence), *Chancery, Italic* (not necessarily slanted), *Spencerian*, *Library hand*, and *Palmer Method* scripts. Taxonomies based on Graphology [3], designed to reveal personality traits, have been largely discredited.

We report research on: (1) Alternative feature representations for style discrimination. (2) Methods of alleviating the effects of lack of representative training data and unequal sample sizes on a set of 250 digitized calligraphic works of historical and artistic importance. (3) Interpretation of the results of comprehensive experiments on calligraphic style classification in terms of the statistical distributions of various elements of an archival database. (4) An experimental design and statistical data analysis that provide a realistic approximation of the expected applications of style quantification and also bear on aspects of classification that are often lacking in synthetic test data.

After a brief review of prior relevant work, we describe our recently developed style features and their statistical behavior on the database in Section 3. In Section 4 we examine alternatives for splitting the data into training and test sets for style recognition by means of a standard Bayesian classifier with empirical priors. Section 5 presents experiments on enhanced style classification using Borda-count voting and a proven scheme of classifier adaptation. We speculate on the causes and possible cures for, egregious misclassifications. In the final section we consider the gap between where we are and the contemplated wide-scale applications.

2. Prior Work

Research on font and writer recognition started long before automated calligraphic style classification, but all three aim at source identification rather than content extraction. There are many works on multi-character classification and classifier adaptation, so we mention only a few key references. Relevant background material can be found in the recent handbook edited by Doermann and Tombre [4], the still useful handbook of Bunke and Wang [5], the smaller DIA/OCR collections of Chaudhuri and Parui [6], Marinai and Fujisawa [7], Cheriet et al. []8, and in Ferilli's instructive monograph [9]. Instead of the many excellent treatises on pattern recognition, machine learning and image processing we cite only a work on data complexity edited by Basu and Ho [10].

2.1 Style, Font, and Writer Recognition

Fortunately a highly respected work on the development of Chinese calligraphy is available in a scholarly English translation [11]. We have not found much research on algorithmic recognition of the style of hand-handwritten Hanzi characters aside from the style models of Zhuang, Lu and Wu [12]. A calligraphic tutoring system based on interactively extracted strokes and automated calligraphy generation was developed by Xu et al. [13]. Their system attempts to automatically evaluate aesthetically pleasing spatial configurations [14]. The classification of Hebrew and Arabic calligraphic handwriting styles has little bearing on Chinese calligraphy because the character structures are entirely different [15].

A comprehensive set of experiments on calligraphic fonts was recently conducted by Bozkurt, Duygulu, and Cetin [16]. This paper also includes a thorough review of previous work on font features and font recognition, including the pioneering contributions of Ingold and Zramdini [17] The value of font recognition for text processing systems was demonstrated by Shi and Pavlidis [18]. A recognized authority on the forensic aspects of writer verification and forgery detection is S.N. Srihari [19,20] An indication of contemporary interest in font design was the year-long Helvetica exhibition at the Museum of Modern Art in New York in 2007-2008.

2.2 Multi-pattern Classification and Adaptation

The way Anne writes a '*c*' suggests how she might write an '*e*'. Only ransom notes and font catalogs mix arbitrary typefaces. The availability of multiple characters from the same source helps to recognize both the characters and their source, as has been repeatedly demonstrated for hand-printing, handwriting, and printed material. Studies of adaptive systems date back at least to the work of Lucky [21] and Tsypkin [22] and later proposals often incorporate variations of Expectation Maximization [23]. We make use here only of the simple bootstrapping scheme presented at DR&R in 1993 [24]. The advantages of the Borda Count were discussed at the same time in the context of decision combination in multiple classifier systems by Ho, Hull and Srihari [25]. An article on analytical results published in 2007 on style-constrained Bayesian classification references several earlier experiments on characters from same-source fields [26].

2.3 Our Earlier Work on Calligraphy Recognition

Initial experiments on forgery detection in Chinese calligraphy were conducted by Zhang and Zhuang in 2007.[27] We described an early version of a graphic interface for verifying or correcting the GB-labels of automatically classified character images at ICDAR 2011 [28] and presented the organization of the database for storing the bibliographic information and the high-resolution color character images at HIP 2011 [29]. The organization of our database has not changed since then, but we are still adding images to it. Our first attempts at style comparisons drew useful suggestions at DR&R in 2012 [30]. At that time we had not yet developed style features. We reported the results of majority vote classification with new style features in a recent JEI article [31]. The current paper summarizes only aspects of our previous research necessary for stand-alone reading and adds (1) statistical analysis of the distance distributions resulting from various partitions of the data set, (2) classification results obtained with training sets realizable under actual operating conditions, and (3) classifier enhancements (adaptation and Borda Count) that are necessary under realistic conditions for labeling styles accurately enough for the proposed applications.

3. Style Features vs. Classification Features

All of our results are based on characters segmented from pages of books of calligraphy scanned at 600 dpi. Characters merged or broken by the initial vertical and horizontal projections (first into columns then into characters) are corrected and labeled using a graphic user interface developed for this purpose. The process is time-consuming because some of the pages are badly degraded or in an unusual layout (Fig.1) and even for experts the assignment of GB-labels and styles to rare and ancient characters often requires consulting scholarly works. Zero labels are entered for characters with unrecognizable meaning or style. All the segmented images are added to our ACCESS database with their bibliographic source, page number, and bounding box coordinates as primary keys.

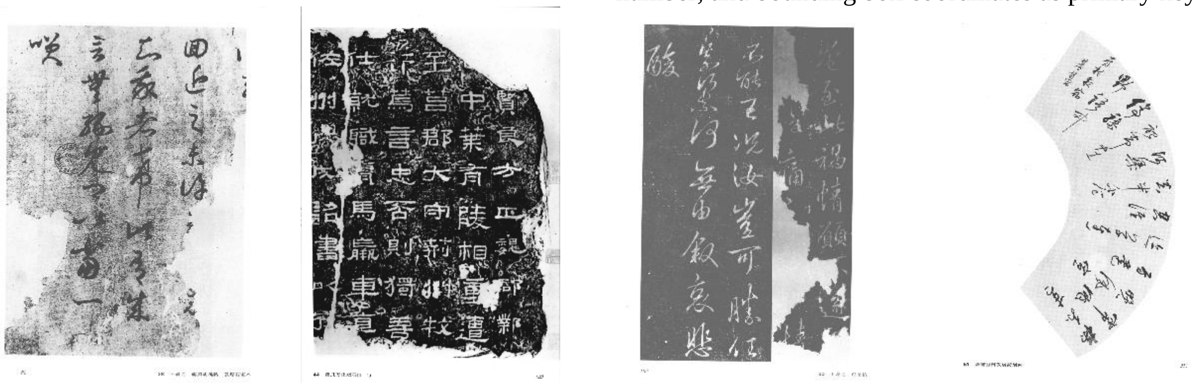


Figure. 1. Examples of difficult to segment page images due to degraded or unusual layout.

3.1 Style Features

Twenty-four style features are extracted from 8719 isolated bi-level character images and their skeletons obtained by medial image transforms. There are twelve stroke level features, including seven features characterize slope variations in near vertical strokes, and five are slope descriptors of the less discriminating near-horizontal features. The remaining twelve features target stroke widths and mass features derived from 2-D central moments.

Because writing with a brush induces style differences conveyed by horizontal and vertical strokes, they are characterized separately. Average slope and changes in slope along selected segments of the stroke are determined directly from the skeleton segments. These measures depend on overall slant and on the presence or absence of serif-like manifestations, which are sensitive style attributes. Average and variability in stroke width, which are related to stress or pressure on the brush, require examination of the binary image surrounding the skeleton segments. The ratio of foreground to background pixels (density), the aspect ratio of the character bounding box, and the location of the foreground centroid with respect to the bounding box are global character features that do not require reference to the skeleton. Additional indices of stress variation and slant balance are obtained from the values of the third central moments M_{03}, M_{30}, M_{21}, and M_{12} in the top/bottom and left/right halves of the image. Altogether there are 7 vertical stroke features, 5 horizontal stroke features, and 12 global character-level features.

Segmenting character into strokes is hampered by deformed crossings and T-junctions. The skeleton stroke segments are therefore traced according to the traditional Chinese character writing rules: left-to-right, top-to-bottom, and outside-in. Directional continuity is preserved when possible to emulate calligraphers. The stroke width variability is computed based on the width of each skeleton pixel. After the strokes are extracted, no further use is made of their putative order.

The feature vectors are entered into an array along with their unique sample identifier CharID, their work identifier WorkID, and the StyleID and GB-Label assigned by the calligraphic data entry expert. Since we cannot use here samples without style labels, 834 characters with unrecognized style (i.e., with StyleID=0) are deleted. However, 63 samples with unknown meaning (with GB-label =0) are kept because their style is the same as that of neighboring characters in the same work. The labeled feature arrays are imported into Matlab for further analysis.

3.2 Style-feature Distance Distributions

None of us ignore the dominant role of features in classification. It is a commonplace that good features exhibit large differences between patterns of the different class and small differences between patterns of the same class. We therefore study the distributions of feature distances between the classes or groupings relevant to calligraphic character images. Below we report the average value and standard deviations of distances between characters of: (1) the same and different works; (2) the same and different styles (3) the same and different GB-labels. Since these three variables condition every distribution, its distances, populations, means and standard deviations under eight conditions are computed. Table 1 shows the mean and standard deviation of the distance distributions of all three variables. The individual components of the Euclidian distances are standardized to unit variance to prevent any feature from dominating the others. Since pairs of characters from the same Work cannot be of different style, the corresponding entries are null.

Table 1. Average and STD of distance between characters with the same or different GB_labels, StyleID, and WorkID.

Grouping		Same GB-label			Different GB-label		
		Average	STD	Count	Average	STD	Count
Same Style	Same Work	3.89	3.5	1882	5.79	3.1	206153
	Different Work	4.42	3.2	22688	5.80	2.9	7892606
Different Style	Same Work	-	-	0	-	-	0
	Different Work	6.37	3.1	63144	6.43	2.9	26129997

Statistics averaged over same or different variables can then be obtained from these values without further distance calculations. The number of characters ("Count") in broader groupings is just the sum of their constituents. The statistics for combining groupings can be derived from Table 1 using the standard formulas and notation, with n_1 and n_2 standing for the appropriate Counts:

$$\mu = (n_1\mu_1 + n_2\mu_2)/(n_1 + n_2)$$

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\mu_1 - \mu_2)^2}$$

Combining the distances over the three variables leads to three possible comparisons:

Table 2. Distances grouped over Styles

Grouping	Same Work		Different Work	
	Average	Count	Average	Count
Same GB-label	3. 89	1882	5.866	85832
Different GB-label	5.79	206153	6.28	34022603

Table 3. Distances grouped over GB-labels

Grouping	Same Style		Different Style	
	Average	Count	Average	Count
Same Work	5.77	208035	--	0
Different Work	6.79	7915294	6.42	26193141

The three possible final grouping are shown in Table 5.

Table 4. Distances grouped over works

Grouping	Same Style		Different Style	
	Average	Count	Average	Count
Same GB	4.38	24570	6.37	63144
Different GB	5.79	8098759	6.43	26129997

Table 5. Averages for each grouping, averaged over the other two groupings.

	GB-labels		Works		Styles	
	Average	Count	Average	Count	Average	Count
Same	5.816	87714	5.771	208035	5.791	8123329
Different	6.275	34228756	6.278	34108435	6.424	26193141

The total Count is 34,316,470, which is equal to Nchars×(Nchars-1)/2, where Nchars = 8285 is the number of characters with style labels. The Grand Average of the distances is 6.27 with a Standard Deviation of 2.94. As expected, characters with the same GB-code are much more similar than corresponding pairs from different works. (Table 2). We can also see from Table 4 that regardless of their source, it is much easier to tell whether two characters from the same style if they have the same GB label. There are, however, far fewer same-GB characters than different-GB characters. Finding such pairs in most applications is far more difficult than finding same-letter comparisons in alphabetic scripts. Regardless of the level of grouping, it is clear that our features do indeed discriminate between styles. Table 5 shows that the overall difference between different and same styles (6.42 – 5.79) exceeds the difference between different and same GB labels (6.27 – 5.81). Are they better than "ordinary" classification features?

To answer this question, we computed the same statistics over the same data set for a set of 16 features designed for Chinese character recognition (i.e., finding the GB-label for an unknown image). We show only the final groupings in Table 6. The actual values of the means for different numbers of features cannot be compared directly. The increase in the distance between patterns is a complicated function of the feature dimensionality because in high dimensions all the patterns tend to be confined to a thin spherical shell [32]. It is clear, however, that for these features the difference between same and different GB labels is much greater than that between same and different styles. The contrast between the two feature sets is even more striking when we consider the difference of the means normalized by variance. The standard deviation of the 16 character recognition features is about 50% smaller in every grouping than that of the 24 style features.

Table 6. Averages for each grouping for features designed for GB-label rather than style discrimination. The standard deviations of the various groupings range from 1.4 to 1.8.

	GB-labels		Works		Styles	
	Average	Count	Average	Count	Average	Count
Same	3.826	87714	4.530	208035	4.934	208035
Different	5.369	34228756	5.370	34108435	5.499	34228756

3.3 Style Vectors

The five 24-dimensional style vectors are the feature vectors averaged over each style. The Standardized Euclidean distance matrix of Table 7 indicates the separation of the style vectors in feature space. Style #5 (*cursive*) is clearly very different from the other four styles. The distance matrix does not show that it is also far more dispersed in feature space.

Table 7. Distances between style vectors.

	Style #1	Style #2	Style #3	Style #4	Style #5
Style #1	0	6.4039	5.9483	5.6301	8.1491
Style #2		0	3.8138	5.0519	9.4464
Style #3			0	3.5119	9.0102
Style #4				0	9.0931
Style #5					0

4. Effects of Training Set Selection and Size

The distribution of character samples in our corpus is highly non-uniform with respect to GB-labels, sizes of works, number of works, and sample populations of each style. This increases the variance of the style classification results when works are randomly assigned to either the training or the test set.

The distributions of the GB-labels, of the sizes of works, and of the number of samples of each style can be computed directly from the database (Fig. 2). The five style populations are 1664, 2867, 1437, 650, and 1667.

The very gradual reduction in the error with the increase in the size of the training set can be seen in Table 8. The table shows results averaged over twenty-fold cross-validation. Each run generates a random permutation of the list of unique CharIDs. The first N characters of the permutated list are assigned to the training set, with N chosen according to the specified fraction. Therefore larger training sets are necessarily evaluated on smaller test sets. None of the experiments (training and testing on 8285 samples, with 20-fold cross-validation) take more than 15 second on a 2GHz computer running Matlab R2013a under Windows 7.

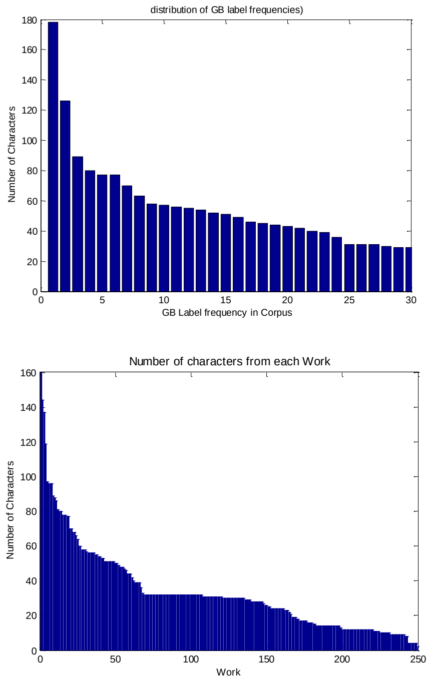


Figure. 2. Above: frequency of GB-labels. The most common character occurs 176 times, and 6606 characters occur 29 times or less. Below: Distribution of work sizes. The largest work has 160 characters.

When there is no training set (0%) the style with the largest number of characters, for example StyleID = 2, is assigned to every sample. The last row of the table shows the results of training on the test set. The residual error rate is high because with the current features, no linear boundary can separate the styles. Quadratic boundaries reduce the training-on-test error rate by only 0.4%. Nevertheless the error rate is lower than might be expected from the distance distributions of Section 3.2 We use pooled rather than individual covariance only because correlations between features are generally due far more to feature design than to classification. For example one might expect both stroke thickness and the number of strokes to be positively correlated with foreground density regardless of style. The linear classifier aligns the hyper planes according to the pooled covariance matrix of all the training samples.

Anticipating the experiments on multi-character recognition in the next section, Table 8 reports results on three different methods of splitting the entire data into mutually exclusive and totally exhaustive training and test sets.

Method 1 is the conventional random split without regard for the source of the characters (i.e., of their WorkIDs).

Method 2 strives for a more representative training set by assigning a given fraction of each work to the training set.

Method 3 is the most realistic evaluation. Each work is randomly assigned to either the training or the test set.

As expected, the highest classification accuracy is obtained by assigning part of each work to the training set. The roughly three-fold increase in variability of Method 3 is due to the large effect of entire works being randomly assigned to either the training set or the test set. The error rate with Method 3 is uniformly higher than with Method 2.

5. Classifier Enhancements

Multi-character classification with voting is equivalent to recognizing the style of a work from which several random samples are available. In character recognition, we often assume that a string of unknown characters is from a single source such as a printed or handwritten phrase or paragraph. The advantage of using linguistic context in character recognition is maximized by selecting adjacent characters. For style recognition a random sample may be more effective because it will be less affected by local page deterioration and segmentation errors.

5.1 Majority and Borda Count voting

The Borda Count is an effective alternative to majority voting because the classifier score (posterior probability) of each voter is taken into account by ranking. Hence a character which has Style #3 as the second highest score contributes more to assigning the unknown to Style #3 than a character which has Style #3 ranked fourth or fifth.

Consider the (real) example below. Each row represents the classifier output for one character that participates in the vote (the actual unknown is the last row.) Digits in the first row are the votes for the first character. From right to left, the classifier assigns Style #4 as the most likely class for the first character, followed by Style #5, with Style #2 least likely. Looking only at the rightmost column, there are two votes for each of styles #3 and #5, and one for each of #2 and #4. Therefore majority voting results in a tie between styles #3 and #5, Borda Count voting uses additional rank information. Style #5 has rankings of 4, 2, 3, 5, 5, and 4 (the contribution decrease from right to left, and each row is ranked independently). Therefore its Borda Count is 4+2+3+5+5+4 = 23. The Borda Count for Style #3 is 22. Therefore the assigned class is #5, which is correct here.

$$\begin{bmatrix} 2 & 1 & 3 & 5 & 4 \\ 1 & 5 & 2 & 4 & 3 \\ 1 & 4 & 5 & 3 & 2 \\ 2 & 1 & 3 & 4 & 5 \\ 2 & 3 & 4 & 1 & 5 \\ 1 & 2 & 4 & 5 & 3 \end{bmatrix}$$

Borda Counts for styles #1, through style #5 = 11, 13, 22, 21, 23

5.2 Classsifer Adaptation

The “self-correcting” adaptation scheme that we use here is independent of the classifier details – it is exactly the same for a linear, Mahalanobis, or quadratic classifier. The classifier parameters are initially estimated on the training set, and the test set is classified with the initial classifier. Next, the test set, with labels assigned by the classifier, is added to the training set, and the classifier is retrained with all the samples (even though some of these samples are mislabeled). Then the test set is reclassified by the retrained classifier. This process can be iterated several times, but in our experience most of the (slight) improvement is generated on the first iteration. The improvement is greatest when the initial accuracy is already high.

Table 8. Classification accuracy vs. Training Set selection and size (averaged over 20 random permutations)

Fraction of data used for training (%)	Method 1 (%)		Method 2 (%)		Method 3(%)	
	Average	STD	Average	STD	Average	STD
0	0.35	--	0.35	--	0.35	--
20	68.4	0.42	70.6	0.6	68.5	2.2
40	69.3	0.53	71.3	0.7	69.6	1.3
60	70.1	0.81	71.6	0.7	69.8	1.8
80	71.3	0.77	72.0	1.1	69.8	2.5
T on T	72.2	--	72.2	--	72.2	--

It is natural to suggest setting a reject threshold so that only confidently classified characters participate in classifier re-training. Natural but wrong. The accuracy invariably drops when only high-confidence samples are added to the training set. Perhaps eliminating characters close to the classification boundary reduces the benefits of additional samples, because it is exactly the representative borderline patterns that are eliminated by the reject procedure.

5.3 Multi-character Classification Accuracy

We validate multi-character recognition using the same procedure as in Section 4. Table 9, shows the improvement in accuracy through multi-character classification averaged over 10 random iterations. Nvote is the number of characters participating in each classification decision, including the unknown itself. The remaining Nvote – 1 samples are selected randomly from the same work. Different "voters" are selected in each random iteration by taking the first Nvote-1 characters from a random permutation of the population of the same work.

As seen in the previous section, single character style classification hovers around 70%, which corresponds to Nvote=1. Even as few as three additional characters result in more than 10% improvement. The key contributor is the information from the additional characters. The Borda Count and Classifier Adaptation help to make slightly better use of this information. Each contributes a cumulative improvement of another percent or so.

When a work is too small to provide the specified number of voters, the classifier takes whatever is available. Here up to 36 test characters (25 on average), depending on which works appeared in the test set, did not have enough voters. Most of the misclassifications are due to the confusions between Running and Cursive styles (#4 and #5) as shown by the highlighted cells of Table 10. The style of a single character in these styles may also puzzle an expert who could readily identify the style of an entire work. We note that the empirical prior probabilities used for Bayesian classification take into account the unequal number of training samples from the different styles.

Table 9. Classification accuracy vs. the number of voters

	Majority Vote		Borda Count		Adaptation	
Nvote	Correc t %	Std	Corre ct %	Std	Corre ct %	Std
1	69.5	1.9	69.5	1.9	67.8	1.5
4	82.8	2.1	84,0	2.1	84.8	1.8
7	87.4	2.1	88.9	2.2	89.9	1.7
10	90.2	2.1	91.1	2.2	92.5	1.6
15	92.3	1.7	92.8	2.1	93.8.	1.7

Table 10. Style-by-style classification results: True StyleID in row header, Assigned StyleID in column header.

	#1	#2	#3	#4	#5	Total
#1	566	0	13	3	17	599
#2	1	807	42	0	0	850
#3	0	51	504	4	2	561
#4	7	2	8	129	54	210
#5	30	10	50	96	361	547
Total	604	870	617	242	434	2767

5.4 Illustrations of Successes and Failures in Style Classification

A visual impression of easy and difficult cases may add insight to the above recognition accuracy statistics. Figure 3 show visual example of characters that all agree on voting the same style.

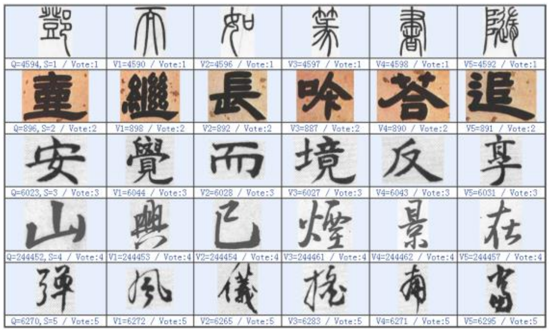
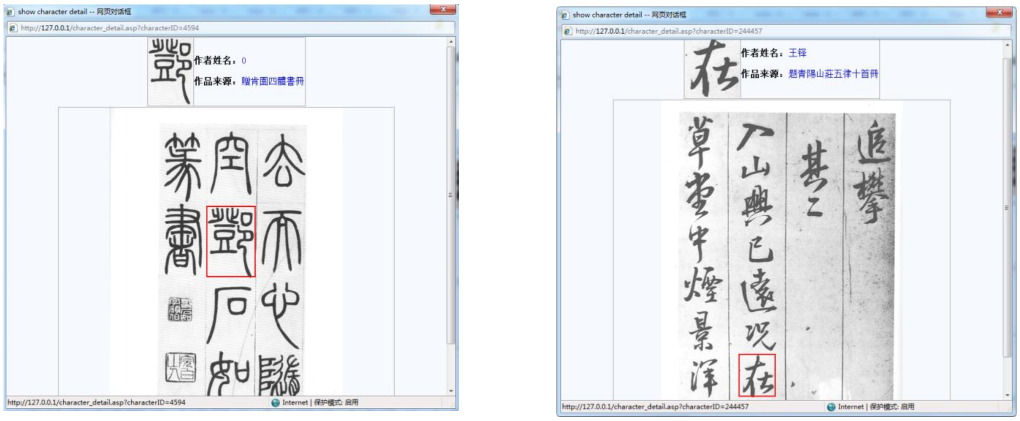


Figure 3. Examples from each of the five styles where every voter agrees on the style: 1st row in style #1, 2nd in style 2, 3rd row in style #3, 4th row in style #4 and 5th row in style #5.

We can inspect any of the characters in their original settings in the page image. The style labels for our experiments where assigned by an expert viewing the characters in their page context. If the first character in the first row of Fig. 3 is clicked, then we can see its original detail page image in the left page of Fig. 4. If the last character in the 4th row is clicked, then we see its page on the right. The red minimum bounding box show the original location of the character.

There were only three cases where all six voters agreed and were all wrong. These are shown in Fig. 5. The voters in all three cases voted for Style #5. The three unknown characters came, however, from a single page on the right that the expert had labeled Style #4. Experts would certainly all agree that the page is either Style #4 or #5, but they might not all agree on one style. This is an example of a borderline style.



show character detail -- 网页对话框

http://127.0.0.1/character_detail.asp?characterID=4594

作者姓名：0

作品来源：贈肯園四體書冊

http://127.0.0.1/character_detail.asp?characterID=4594

Internet | 保护模式: 启用

show character detail -- 网页对话框

http://127.0.0.1/character_detail.asp?characterID=244457

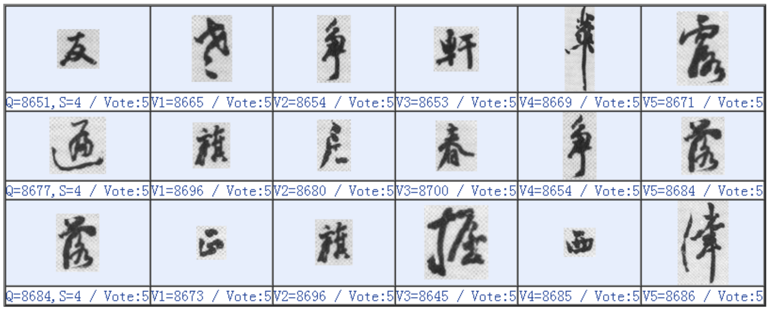
作者姓名：王铎

作品来源：題青陽山莊五律十首冊

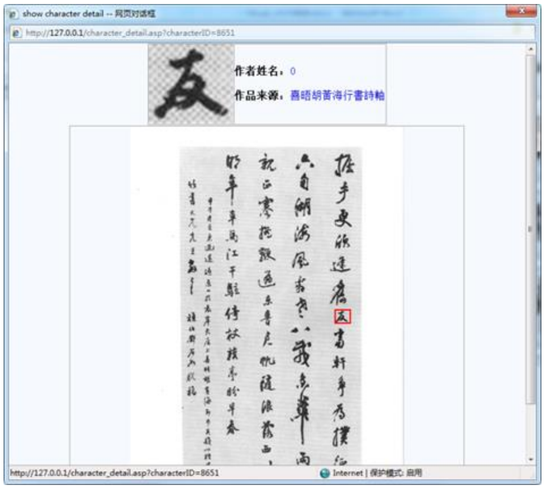
http://127.0.0.1/character_detail.asp?characterID=244457

Internet | 保护模式: 启用

Figure 4. Consensually classified characters in their original page context.



Q=8651,S=4 / Vote:5	V1=8665 / Vote:5	V2=8654 / Vote:5	V3=8653 / Vote:5	V4=8669 / Vote:5	V5=8671 / Vote:5
Q=8677,S=4 / Vote:5	V1=8696 / Vote:5	V2=8680 / Vote:5	V3=8700 / Vote:5	V4=8654 / Vote:5	V5=8684 / Vote:5
Q=8684,S=4 / Vote:5	V1=8673 / Vote:5	V2=8696 / Vote:5	V3=8645 / Vote:5	V4=8685 / Vote:5	V5=8686 / Vote:5



show character detail -- 网页对话框

http://127.0.0.1/character_detail.asp?characterID=8651

作者姓名：0

作品来源：喜晤胡菁海行書詩軸

http://127.0.0.1/character_detail.asp?characterID=8651

Internet | 保护模式: 启用

Figure 5. Examples from a borderline style where all the votes were wrong.
Clicking on the first character above brings up the page on the right,
with the red minimum bounding box showing the character's location.
Even experts may disagree whether this page is written in Running or Cursive style.

6. Conclusion

A study of the distribution of distances between character pairs from various groupings by GB-label, WorkID, and StyleID shows that the local and global features that we have developed are sensitive to style differences. This enables formulation of a vectorial style space partitioned into style regions by the distance of samples from one or more style vectors. The style sensitivity of the features also allow recognition the style of small groups of characters from the same work without training the classifier on any samples from that work. The classification is much more accurate for Seal, Clerical and Regular styles than for running and cursive styles. It is, of course, even more accurate when representative samples are available from each work, but this condition can seldom be realized in the cultural and educational applications of interest (e.g., a smart phone app that identifies the style of a photographed fragment of a calligraphic inscription or scroll).

Combining classifier results with the Borda Count instead of majority vote, and a simple adaptation scheme, raise the overall accuracy to 90% when seven same-source characters are available from a work of unknown style. The correlation between features is dominated by the selected feature set rather than by the style class, therefore there is no advantage in using individual covariance matrices instead of the more robust pooled covariance matrix. We believe that the current classification rate over multiple characters exceeds that of expert classification based on a single character and approaches the limit set by the perceptual overlap between styles.

We intend to classify the works in each main style group into subgroups. The most promising candidate for further partitioning is the #3 Regular style. Unsupervised style classification (clustering) of works offers the opportunity to discover hitherto unknown relationships between different groups of calligraphy artists. This may require further refinement of our current Euclidian distance based measure of calligraphic similarity.

Original calligraphic works are highly valued in China. Therefore forgery detection is remains a potential application of style recognition. Style recognition, like font recognition for Western scripts, is also likely to improve automated transcription of ancient works. Although the most important works have already been transcribed by hand, scholars would like to be able to search and compare with modern document analysis tools the many less significant works that have been scanned but not transcribed. This trove includes millions of family histories (Jiapu), where providing individuals with the ability to automatically transcribe their own records might allay privacy concerns. Of less cultural but greater popular and commercial interest is the generation and rendering of arbitrary text in favorite styles.

Acknowledgments

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

References

1. Chinese calliraphy service of CADAL, [EB/OL]. http://www.cadal.zju.edu.cn/NewCalligraphy /2015-05-04.
2. Universal Digital Library web site, [EB/OL]. http://www.ulib.org /2015-05-04.

3 H. Clifford, Graphology: how to read character from handwriting, with full explanation of the science, and many examples fully analyzed, Penn Pub. Co., Philadelphia, 1905.

4 D. Doermann and K. Tombre, Handbook of Document Image Processing and Recognition, Springer, 2014.

5 H. Bunke and P.S.P. Wang, Handbook of Character Recognition and Document Image Analysis, World Scientific 1997.

6 B.B. Chaudhuri and S.K. Parui (editors), Advances in Digital Document Processing and Retrieval, World Scientific 2014.

7 S. Marinai and H. Fujisawa (editors), Machine Learning in Document Analysis and Recognition, Springer, 2007.

8 M. Cheriet, N. Kharma, C-L Liu, C.Y. Suen, Character Recognition Systems, Wiley-Interscience 2007.

9 S. Ferilli, Automatic Digital Document Processing and Management, Springer, 2011.

10 M. Basu and T.K. Ho, Data Complexity in Pattern Recognition, Springer, 2006.

11 Qiu, Xigui, Chinese writing, translated by G.L. Mattos and J. Norman. Berkeley: Society for the Study of Early China and The Institute of East Asian Studies, University of California. 2000. (English translation of Wénzìxué Gàiyào 文字學概要, Shangwu, 1988.)

12 Y. Zhuang, W. Lu, J. Wu, Latent Style Model: Discovering writing styles for calligraphy works. J. Visual Communication and Image Representation 20(2): 84-96, 2009.

13 S, Xu, H. Jiang, F.C.M. Lau, Y. Pan, An Intelligent System for Chinese Calligraphy, Proceedings of the National Conference on Artificial ntelligence 22, (2), 1578-1583, 2007.

14 S, Xu, H. Jiang, F.C.M. Lau, Y. Pan, Computationally Evaluating and Reproducing the Beauty of Chinese Calligraphy. IEEE Intelligent Systems 27(3): 63-72, 2012.

15 I.T. Yosef, K. Kadem, I. Dinstein, M. Beit-Arie, E. Engel, Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results, Proceedings of the First Workshop on Document Image Analysis for Libraries, Palo Alto, CA 2004.

16 A. Bozkurt, P. Duygulu, and A. E. Cetin, Classifying Fonts and Calligraphy Styles Using Complex Wavelet Transform, arXiv preprint arXiv: 1407.2649, 2014.

17 A.W. Zramdini, R. Ingold, R Optical Font Recognition Using Typographical Features. IEEE Trans. Pattern Anal. Mach. Intell. 20(8): 877-882, 1998.

18 H. Shi and T. Pavlidis, Font recognition and contextual processing for more accurate text recognition, Proceedings of the Fourth International Conference on Document Analysis and Recognition, Vol. 1, 39-44, Ulm, 1997.

19 .S. N. Srihari and K. Singer, "Role of Automation in the Examination of Handwritten Items," Pattern Recognition Journal , 2014

20 S. N. Srihari, Determining Writership of Historical Manuscripts using Computational Methods in On-line Proceedings Automatic Pattern Recognition and Historical Handwriting Analysis, Erlangen, Germany, 2013

21 R. W. Lucky, Techniques for adaptive equalization of digital communication systems, Bell Sys. Tech. J., vol. 45, pp. 255-286, February 1966.

22 Y. Z. Tsypkin, Adaptation, training, and self-organization in automatic systems, Automation and Remote Control, vol. 27, pp. 16-52, January 1966.

23 A.P. Dempster, N.M. Laird, and S. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Proc. Roy. Stat. Soc., B-39:1-38, 1977.

24 H.A. Baird and G. Nagy, A Self-correcting 100-font Classifier, Proceedings of the SPIE Conference on Document Recognition, vol. SPIE-2181, 106-115, San Jose, CA 1994.

25 T.K. Ho, J.J. Hull, S.N. Srihari, Decision Combination in Multiple Classifier Systems IEEE Trans. Pattern Analysis and Machine Intelligence, 16, 1, January 1994.

26 H. Veeramachaneni and G. Nagy, Analytical results on style-constrained Bayesian classification of pattern fields, IEEE Trans. Pattern Analysis and Machine Intelligence 29, 7, 1280-1285, 2007.

27 X. Zhang, Y. Zhuang, Visual Verification of Historical Chinese Calligraphy Works, Lecture Notes in Computer Science, MMM'2007, LNCS 4351, pp: 354–363, 2007.

28 G. Nagy and X. Zhang, CalliGUI: Interactive Labeling of Calligraphic Character Images, Procs. ICDAR'11, Beijing, 2011.

29.X. Zhang, G. Nagy, The CADAL Calligraphic Database, Procs. HIP '11, Beijing, September 2011.

30 X. Zhang, G. Nagy, Style comparisons in Calligraphy, Procs. SPIE/IST/DRR,San Francisco, Jan. 2012.

31 X. Zhang and G. Nagy, Computational method for calligraphic style representation and classification, Journal of Electronic Imaging 24(5), 053003, Sep⁄Oct 2015.

32 G. Nagy and X. Zhang, Simple statistics for complex features spaces, in Data Complexity in Pattern Recognition, pp. 173-195, M. Basu and T. K. Ho, Eds., Springer, 2006

Author Biographies

George Nagy received a B.Eng. in Engineering Physics (1959) and an M.Eng. in Electrical Engineering (1960) from McGill University, and a Ph.D. from Cornell University in 1962. He worked longer at IBM, UNL, and RPI than anywhere else, and conducted some research on pattern recognition, computational geometry and document image analysis. He retired in 2011 and is now Professor Emeritus at RPI, which entitles him at last to do his own programming.

Xiafen Zhang obtained a B.S. from Fushun Petroleum Institute (2000), an M.S. from Liaoning Shihua University (2003), and the PhD. from Zhejiang University (20006), all in Computer Science. She is a Lecturer at the College of Information Engineering of Shanghai Maritime University. She conducts research on image processing for improved digital library access to ancient calligraphy and other historical documents.