Brush Writing Style Classification from Individual Chinese Characters

Conference Paper · January 2006

DOI: 10.1109/ICPR.2006.343 · Source: IEEE Xplore

3 authors, including:

Tak-sum Wong
The Hong Kong Polytechnic University
12 PUBLICATIONS **81** CITATIONS

SEE PROFILE

Brush Writing Style Classification from Individual Chinese Characters

Sam T. S. Wong, Howard Leung, Horace H. S. Ip
Department of Computer Science, City University of Hong Kong
{tswong, howard, cship}@cityu.edu.hk

Abstract

Chinese calligraphic artwork usually contains a few or even a single character. In order to perform automatic brush writing style classification on these kinds of Chinese calligraphic images, existing approaches need to be modified because they assume that many characters of the same style exist in the input image. A novel approach is proposed to address the brush writing style classification problem for single-character Chinese calligraphic images by combining the texture analysis and the structural information through a set of parameterised ellipses.

1. Introduction

Recently there are lots of interests in building digital libraries to store information in digital format. Several projects aim at digitising ancient Chinese texts and calligraphic masterpieces [1][2][3][4][5][6]. However, they either only store the photograph or the scanned image of the written calligraphic artwork; or encode the scripts in the form of typed text. A lot of manual effort is required in annotating each artwork to provide a description such as the writing style. In order to reduce manual intervention in the annotating process, image processing techniques can be applied to analyse Chinese calligraphic masterpieces to perform some automatic writing style classification.

Brush writing style classification of Chinese calligraphy is a difficult and time consuming task. Only a few researchers have addressed this issue despite its importance. There are two existing approaches: 1) identification and matching with Most Frequently Used (MFU) characters and 2) Gabor texture analysis.

In the first approach used by *Lin et al.* [7], features from every block of character image in the document are extracted. These include the density of black pixels, projection-profile code, the skeleton temples, *etc*. They are then matched with the templates in the database that contains different font templates of the top-40 Most Frequently Used (MFU) Chinese characters for identification. With a single database of the specified 40 characters, it can only work well for the identification of certain text in a particular period. Since language evolves through time, the MFU characters are different for text written in different historical periods. As a result, the performance of this method is expected to be not good for text from various ancient periods since the MFU characters may not appear in the text. On the other hand, the computation will become very complex if one increases the number of MFU characters in the database.

In the second approach, *Zhu et al.* [8] treat the input document as a texture image. Features are extracted after applying Gabor filters from the input text-image and then matched with the features of the template text in the dictionary. Other research work is done in similar directions on this topic [9][10]. The characters in different passages can vary in texture even with the same font. The performance with this approach is thus not very satisfactory for all fonts.

Furthermore, the above two approaches assume that a passage of text with many characters is present in the input image. However, in ancient Chinese calligraphic artwork, sometimes the number of characters contained in the artwork is very few. As a result, due to the large variations among individual characters even in the same style, one needs to extend existing approaches to the case when a few or even only a single character exists in the input Chinese calligraphic image and perform the writing style classification.

In this paper, a novel methodology in content independent and automatic writing style classification is proposed to identify the font styles of individual characters written by traditional Chinese bristle brush. The proposed approach is illustrated in Figure 1. The image characteristics of each writing style are analysed after parameterising each character by a set of ellipses. This is similar to our prior work on the parameterisation for calligraphic rendering in [11] but with different purpose. Afterwards, features are extracted from the set and are combined with the output channels from Gabor filter for identification.

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

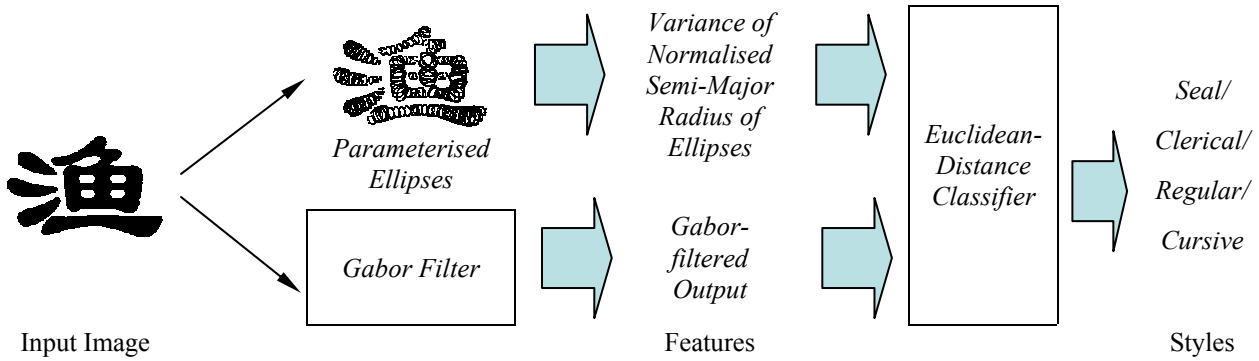
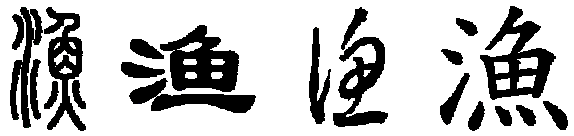


Figure 1 Flow of our proposed algorithm

The rest of this paper is organised as follows. The classification by texture analysis approach alone is reported in section 2. Our proposed parameter representation of characters and additional feature for classification are described in section 3. The result with the proposed classification scheme is discussed in section 4. Conclusions and future work are given in section 5.

2. Classification by Texture Analysis

In this paper, four types of brush writing styles are used to illustrate our algorithm. They are the seal character, clerical script, cursive script and regular script that are shown in Figure 2.



(a) Seal character (b) Clerical script (c) Cursive script (d) Regular script

Figure 2 Four traditional types of brush writing styles in calligraphy for the same character

A common way for classifying brush writing styles is by global texture analysis with Gabor filter as in [9]. In [9], the set of parameters recommended for the classification of the single-font text-block include orientation $\theta = \{25°, 55°, 75°\}$ with spatial frequency $\omega \le \frac{N}{4}$ for textures of image resolution $N \times N$ pixels. We have conducted some experiments to find out the best set of orientations and spatial frequencies to be used for writing style classification from individual characters. The 9 best Gabor-channel features from our experiments are listed in Table 1 in descending order of classification rate.

In our experiment, out of the 452 character images ($N \ge 120$) per class mentioned above, one tenth is used for training while the rest are used for testing. Various sets of θ's and ω's have been tested. The best average testing-set classification rate is 84.51% by making use of the Gabor-channels with $\theta = \{25°, 55°, 75°\}$ and $\omega = \{24, 18, 12, 6\}$. The mean and standard deviation are computed from each filtered image and form a 24-element feature vector for classification. Euclidean-distance classifier is used to identify the styles. The confusion matrix using the best Gabor parameter set is shown in Table 2.

Table 1 Results with various Gabor parameter sets

Gabor-Channel Features		Training Rate	Testing Rate
ω	θ(°)		
24,18,12,6	**25,55,75**	**86.71%**	**84.51%**
32,24,16,8	25,55,75	76.88%	76.05%
40,30,20,10	25,55,75	72.90%	75.06%
8,6,4,2	25,55,75	65.82%	68.69%
16,8,4,2	25,55,75	65.98%	68.53%
32,16,8,4	0,60,120	75.56%	68.20%
40,20,10,5	25,55,75	65.54%	67.87%
32,16,8,4	25,55,75	68.89%	67.70%
32,16,8,4	0,45,90,135	73.89%	66.54%

Table 2 Confusion matrix by global texture analysis for the Gabor parameter set $\theta = \{25°, 55°, 75°\}$ and $\omega = \{24, 18, 12, 6\}$

	Seal	Clerical	Cursive	Regular
Seal	71.24%	0.00%	13.50%	15.27%
Clerical	0.00%	97.12%	0.88%	1.99%
Cursive	2.88%	0.88%	91.59%	4.65%
Regular	10.18%	0.00%	11.73%	78.10%

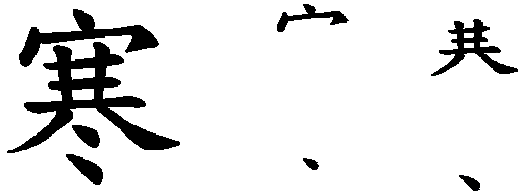
It can be seen in Table 2 that the classification rates of the clerical and cursive scripts are both over 90%. However, the result for seal character and regular script are not very satisfactory. It can be seen that in many cases the seal character is mixed up with the regular script. A parametric representation of characters is thus proposed to assist in improving the classification rate of writing style from individual characters.

3. Parametric Representation and Additional Features for Classification

We propose to first parameterise a character image with a set of ellipses that fit onto the character region. To have a parametric representation, each character image is first divided into connected regions. For each connected region, the boundary is extracted and an initial set containing all possible ellipses covering the region is found. Afterwards, a subset of ellipses that covers the whole region and represents the structure of the character is selected from the initial set. Finally geometric features are extracted from the subset of ellipses and then are augmented with the features from the texture analysis to perform the classification.

3.1. Parameterisation with Ellipses

To find a set of ellipses covering the supporting region of a character, the character is first divided into connected components. The corresponding connected components of Figure 3(a) are shown in Figure 3(b).



(a) Input character (b) Connected components

Figure 3 Character with its connected components

The boundary of each connected component (region) is extracted for finding an initial set of ellipses. Figure 4(a) shows a region in Figure 3(b) and its result of boundary's extraction. The line segment joining every two points on the boundary are selected as the major-axis to construct an ellipse with a fixed aspect ratio. If none of the pixels in that ellipse are outside the supporting region, this ellipse is taken as one of the members in the initial set. In our algorithm, the aspect ratio, which is the ratio between the major radius and minor radius of the ellipse, is set to be 12 to 7.

(a) Region and its boundary (b) Subset of ellipses

Figure 4 Extraction of ellipses

To cover the supporting region, however, it is not necessary to use all the ellipses in the set but only a subset is sufficient. Therefore, for every pixel on the boundary, we pick the ellipse with the median angle of rotation among all the ellipses touching that pixel to be included in the subset of ellipses. By using this criterion, as only one ellipse is selected for one boundary point, the number of ellipses is greatly reduced from the set while maintaining a large coverage.

Figure 4(b) shows the subset of ellipses. For most of the ellipses, the major axes are always perpendicular to the tangent of the boundary. This suggests that most of the fitted ellipses are good. The ellipses in the subset are sufficient for representing the structure of the character.

3.2. Geometric Features

The variance of the semi-major radius normalized by their maximum is computed from the subset of ellipses. This geometric feature is then used to augment with the features from the texture analysis to assist in distinguishing between seal character and the other three types of scripts. The scatter plot of this geometric feature from 452 character images per style is shown in Figure 5.

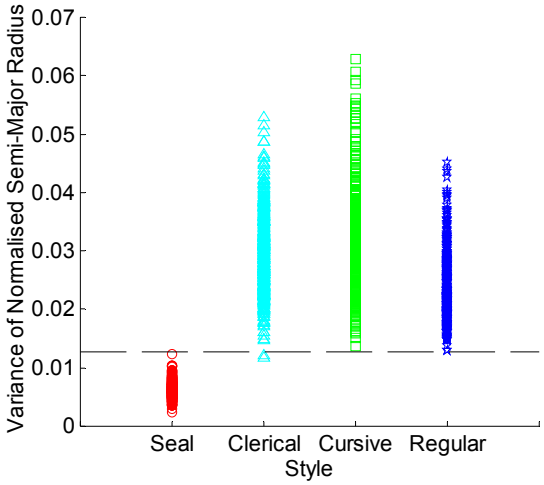


Figure 5 Scatter plot of the variance of the normalised semi-major radius among the four styles

In Figure 5, it can be seen that the variance of the normalised semi-major radius, which characterises the uniform thickness of the seal characters, can be used to classify between the seal character and other scripts.

4. Experiments and Classification Results

The variance of the normalised semi-major radius is augmented with the Gabor-channel features to form a

combined feature vector. A scaling factor is required to make the two kinds of features comparable. As the magnitude of our proposed geometric feature is in the order of 10^{-2} and the magnitudes of Gabor channels are in the order of 10 to 10^2, a coefficient $c = 8800$ is multiplied to the geometric feature as the scaling factor.

The confusion matrix of the Euclidean-distance classifier with the combined feature vector is given as in Table 3. The average classification rate is 94.0%. It can be seen that the classification rate of the seal script is almost perfect with an improvement of about 30% compared with Table 2. There are also some improvements for the classification rate of the cursive and regular scripts while that of the clerical script has only dropped by less than 1%. As a result, the average result has been improved prominently by about 10%. However, there are still rooms for improvement of the classification of the regular script.

Table 3 Confusion matrix by combination of texture analysis and proposed geometric feature

	Seal	Clerical	Cursive	Regular
Seal	99.78%	0.00%	0.22%	0.00%
Clerical	0.00%	96.24%	1.11%	2.65%
Cursive	0.00%	1.55%	96.02%	2.43%
Regular	0.22%	2.88%	12.83%	84.07%

From a historical point of view, the cursive script is derived from the regular script so it also possesses some of the properties of the regular script. The relatively lower classification rate is probably due to the similarity between them.

5. Conclusions and Future Work

In this paper, a novel methodology in content independent and automatic brush writing style classification is proposed for the identification of the font style of individual characters. Our approach combines global texture analysis with the geometric feature from the parameterised ellipses. The proposed geometric feature captures the underlying structure of the characters.

The classification result of the seal characters, clerical and cursive scripts are quite good. However, the classification rate for the regular script is not satisfactory enough due to the similarity between the fonts. Therefore, as future work, more features will be sought and considered by further analysing the set of parameterised ellipses to improve the classification rate. In addition, more scripts will also be considered for testing the flexibility of our approach under more variations.

6. Acknowledgements

The work described in this paper was supported by a grant from City University of Hong Kong (Project No. 7001614).

7. References

[1] J. Derming, J.-H. Wang, C.-Y. Lai, C.-C. Hsieh, L.-F. Chien, and J.-M. Ho, "Resolving the Un-encoded Character Problem for Chinese Digital Libraries", *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, pp. 311–319, June 2005.

[2] Scripta Sinica, *Hanji dianzi wenxian*, Academia Sinica, http://www.sinica.edu.tw/~tdbproj/handy1/.

[3] NDAP, National Digital Archives Program, Academia Sinica, http://www.ndap.org.tw/.

[4] G. Yang, and T. Zhang, "The Development of the China Digital Library", *Electronic Journal of Academic and Special Librarianship*, Vol. 3 (3), 2002.

[5] China-America Digital Academic Library, http://www.cadal.zju.edu.cn/index.jsp.

[6] C.W. Ho, "CHANT (CHinese ANcient Texts): a Comprehensive Database of All Ancient Chinese Texts up to 600 AD", *Journal of Digital Information*, Volume 3, Issue 2, Article No. 119, Aug. 2002.

[7] C.H.F. Lin, Y.F. Fang, and Y.T. Juang, "Chinese Text Distinction and Font Identification by Recognising Most Frequently Used Characters", *Image and Vision Computing*, Vol. 19, pp. 329–338, 2001.

[8] Y. Zhu, T.N. Tan, and Y.H. Wang, "Font Recognition Based on Global Texture Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23 (10), pp. 1192–1200, 2001.

[9] F. Yang, X.-D. Tian, and B.-L. Guo, "An Improved Font Recognition Method Based on Texture Analysis", In *Proceedings of the First International Conference on Machine Learning and Cybernetics*, Peking, pp. 1726–1729, November 2002.

[10] T.N. Tan, "Rotation Invariant Texture Features and their Use in Automatic Script Identification", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20(7), pp.751–756, July 1998.

[11] T.S.S. Wong, H. Leung and H.H.S. Ip, "Model-based Analysis of Chinese Calligraphy Images", *Proceedings of 9th International Conference on Information Visualisation*, London, July 2005.