The CADAL Calligraphic Database

Xiafen Zhang
College of Information Engineering
Shanghai Maritime University
Shanghai, P.R. China 201306
13564911632
xfzhang@shmtu.edu.cn

George Nagy
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY, USA 12180
0015182766078
nagy@ecse,rpi.edu

ABSTRACT

A set of 13,351 digitized calligraphic characters were segmented and labeled, with 12,918 characters extracted from 21 books scanned by the CADAL scanning center located in Zhejiang University's library, and 1443 characters from calligraphy works from web sources. The database contains calligraphy from 208 works, some from over 1000 years ago. Statistics are provided on provenance, character size and shape, and label frequency distribution. Specific problems encountered in creating a calligraphic database are illustrated and discussed. Progress is reported on a classifier-based interactive labeling system that halves the human labor necessary to expand the database.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture–*Graphics recognition and interpretation.*

General Terms

Experimentation.

Keywords

Chinese calligraphy, historical documents; ground truth; styles.

1. INTRODUCTION

Literally, *calligraphy* means beautiful writing. The art of fair or elegant penmanship is still taught in elementary school in many countries with a tradition of aesthetic appreciation of handwriting. Outstanding examples, some hundreds or even thousands of year old, are treasured in museums all over the world.

Because of the ideographic foundations of Chinese characters and the popularity of writing for nearly 3,000 years, there are many collections of calligraphy in China and calligraphy occupies an important place in Chinese culture. It is therefore not surprising that the digitization of calligraphic books was a priority in the China-US Million Book Project.

As in the case of other historical documents, considerable work is required to make such a collection useful for the diverse objectives of art historians, students, and the public. At a minimum it appears desirable to be able to search the collection by author/artist and work, and to be able to visualize how a work was written originally. *Work* here denotes an instance of a calligraphic composition by an artist. A work is also a message from the past to the present, because people wrote messages by hand before the invention of printing. It is not unusual, however, to find copies of works, by the same artist, that are admired as much as earlier versions.

Our objective here is to make the above collection accessible at a finer grain of individual characters. This requires character segmentation and linking each character image with a label, a pointer to its physical location in the original source material, and a reference to the relevant bibliographic data (author, title, date, etc.). The increased level of access will facilitate studies of the evolution of calligraphic styles and shapes, developing style-based taxonomies, and detecting duplicates, copies and forgeries [1,2,3,4,5]. Access to individual character images will also increase the usefulness of the database for further experiments on calligraphic document image processing. The characteristics of a useable calligraphic database, described below, differ from those of test data for research on printed or handwritten Chinese character recognition [6,7,8,9].

Section 2 describes the data collection process. Section 3 provides diverse statistics on the database. Section 4 is a report of our experience with a web-based interactive labeling system designed to accelerate the time-consuming process of labeling new characters. The final section summarizes our observations and outlines possible future work.

2. DATA COLLECTION

In this section we describe the sources of the character images and the processing steps necessary to convert them into a useful digital format.

2.1 CADAL

The China Academic Digital Associative Library (CADAL) [10] manages the China-US Million Book Digital Library Project, and is an important part of the Universal Digital Library (UDL) [11]. By November 2010, 1.323 million Chinese books have been scanned at the CADAL scanning center. There are 50 UDL scanning centers, scanning books in different languages day by day. The books are linked to Dublin Core (bibliographic) metadata. CADAL offers bilingual multimedia services including content-based image and video search. According to a policy promulgated in 2006, "the service of Chinese calligraphy character retrieval is provided in the CADAL digital library, treating them just as they are images without recognizing them like OCR does".

HIP '11, September 16 - September 17 2011, Beijing, China Copyright 2011 ACM 978-1-4503-0916-5/11/09…$10.00.

2.2 Sources of Calligraphy

The original substrates for ancient calligraphy—stone, bamboo sheet, silk scroll and rice paper—are inaccessible or too delicate for digitization. The material in our collection was extracted from 21 of 53 scanned books of calligraphy selected by a noted scholar. Each book has some theme: for example, works from the same dynasty. The two examples below illustrate the range of styles

Lanting Xu ("Preface to the Poems composed at the Orchid Pavilion") was written in 353 AD by Wang Xizhi and is admired for its early Running Style that displays economy of hand movement (and ink), as well as for its literary qualities (Fig. 1). It is the introduction to the poems written at a gathering of forty-two poets meeting as friends. Three hundred years later, Emperor Tai Zong of the Tang Dynasty (619-907) found the original after a legendary search, ordered his calligraphers to copy it, distributed the copies to the court, and had the original buried with him.

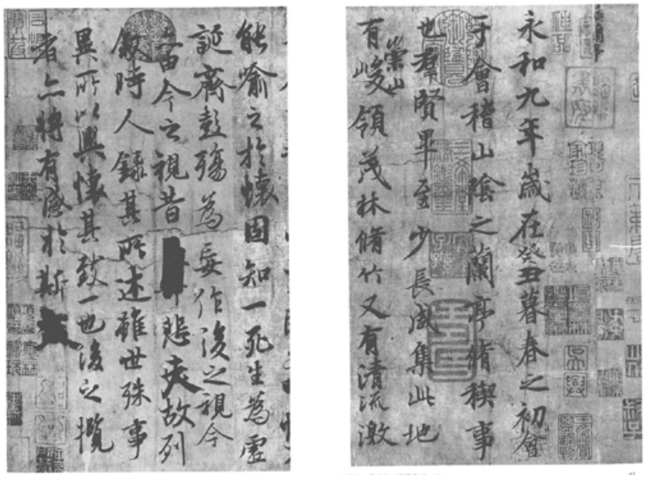


Figure 1.Beginning (on the right) and end (on the left) of the work *Lanting Xu*. The seals of successive owners are clearly visible.

The second work is a stone rubbing of the first two pages of a work of Yan Zhenqing, who was a Tang emperor's household calligraphy teacher (Fig. 2). It recounts the story of the monk who brought back Buddhism from India to China. The Regular Style is in sharp contrast with the fluent Running Style of *Lanting Xu*, where one of the characters that occur twenty times is never written the same way. The Regular Style is often the one taught first to newcomers to writing with a brush.

Fiigure 2. Extract from a work of Yan Zhenqing. The calligraphy is notable for its extreme regularity. It is almost like print.

All the characters, except those deemed illegible, were extracted from several dozen works. However, because of the skewed distribution of character labels in Chinese, this resulted in relatively few distinct labels. (Note that in English the letter frequency distribution is also highly skewed: "e" occurs about 170 times more often than "z".) Therefore the remaining works were sampled for character images with rare labels or unusual structures or shapes. To increase diversity, an additional 1443 characters were obtained from on-line sources (often with incomplete book and page references).

2.3 Digitization

The calligraphy books were scanned with an Avision FB6080E (a high-speed flatbed CCD graphics scanner for up to A3 sheets) instead of the larger Konica Minolta PS7000 overhead book scanner used at CADAL for most books. The scanned pages were digitized at 600 dpi (23.6 lpm) optical resolution into 24-bit per pixel RGB TIFF (for archival and image processing) and JPG (for presentation). Both formats are preserved in the database – JPG is more convenient for viewing page images, but the compression artifacts hamper character classification.

Either black or red ink was generally used for brush work, and also for the seals that were used to show ownership. Some books are not printed in color. Stone rubbings are light gray on dark gray background.

Most of the digitized books on calligraphy were published in the last two decades. In addition to the photographic illustrations of calligraphy, they contain scholarly notes about author/artist, historical context, calligraphic style, and technique. The aspect ratios of scrolls, stone tablets and rubbings don't necessarily fit the page size of modern books, so some works are paginated arbitrarily. Page numbers are recorded as metadata during digitization.

The fidelity of reproduction varies from work to work. While all calligraphy books attempt to preserve the "essential qualities" of the calligraphy, the size of the printed reproductions varies by a large integer factor. This accounts for more of the variability of height and width (in pixels) of the characters from different works than does size variation in the original.

We don't know how the quality was monitored and whether any test charts were stored as part of the digitization process, but visual comparison of the digitized pages with the hardcopy books did not reveal any major problems.

2.4 Binarization and Segmentation

The darkness and variability of the background of the reproductions that reflect the aging of the original substrates render automatic foreground/background segmentation problematic. Furthermore, the page numbers and figure titles are darker than the calligraphy. Therefore global thresholds for both the calligraphy and the added printed material were automatically estimated from the reflectance distribution, and adjusted manually when necessary.

Semi-automatic segmentation (Fig. 3) comprised the following steps: (1) Eliminate seals that can be differentiated by color by clicking on the seal and deleting neighboring pixels of the same color. (2) Binarize the page images. (3) Cut the page into column blocks by projecting the foreground pixels onto the horizontal axis and cut the column blocks into character blocks projecting the column pixels onto the vertical axis. (4) Segment connected characters block into individual characters. (5) Eliminate blocks outside the normal size range (6) Detect algorithmically and merge partial character blocks into complete character blocks. (7) Extract the RGB character images using the minimum bounding boxes of the binarized characters. (8) Re-binarize the character images using individual thresholds. (9) Filter out noise specks and small “hairs”. (10) Save the coordinates of the minimal bounding box with a unique eight-digit character identifier (CID).

The acceptable widths and heights of complete characters and of character components to be merged were determined algorithmically.

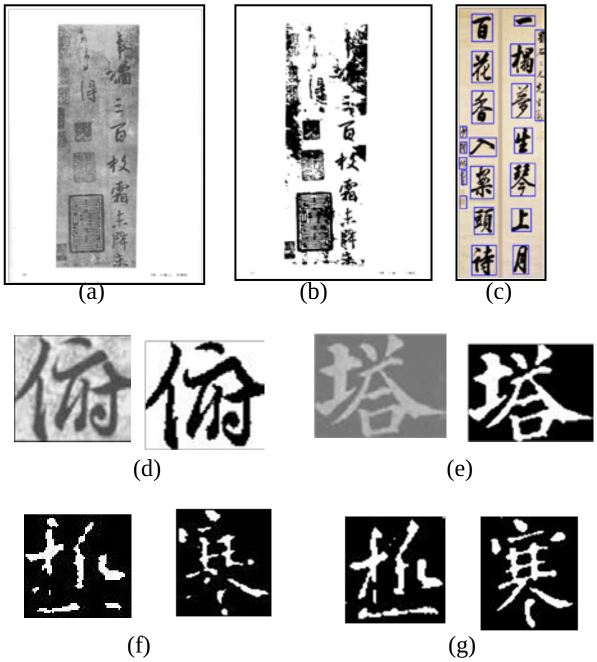


Figure 3. (a) page image with page number and title of work; (b) binarized with a single threshold; (c) blocks outside the normal size range on left and right are eliminated; (d) regular script; (e) stone rubbing; (f) character image after page binarization; (e) after re-binarization.

A few bounding box coordinates are off by 4-5 pixels because of noise specks. Examples of binarized calligraphic characters are shown in Fig. 4, where several styles of calligraphy and several kinds of noise can be observed. The lack of context here highlights the difficulty of recognizing individual characters.

2.5 Labels and Metadata

The diversity in the style, shape and aspect ratio of the characters is clearly a challenge for automatic label recognition. Although the original plan was to use this collection only for content based image retrieval by character shape, character labels are necessary for many applications. To test our proposed classifier-based interactive labeling system (Section IV), we needed a labeled database. Therefore the segmented character images were manually labeled with 16-bit GB 2312 codes. GB 2312 encodes 6763 characters. About 150 characters that we could not recognize were labeled with null codes.

The 32-bit Chinese National Standard GB 18030-2000, which covers most of the traditional characters among its 70,244 assigned codes, is a superset of both ASCII and GB 2312. The GB 2312 codes form the 16 low-order bits of the corresponding GB 18030 codes. Furthermore, GB 18030 can be mapped 1:1 to 32-bit Unicode. The standard was expanded further in 2005 with the addition of Tibetan and Mongolian characters. As more ancient and rarely used characters are added to the database, it may be desirable to convert the current encoding to 32 bits.

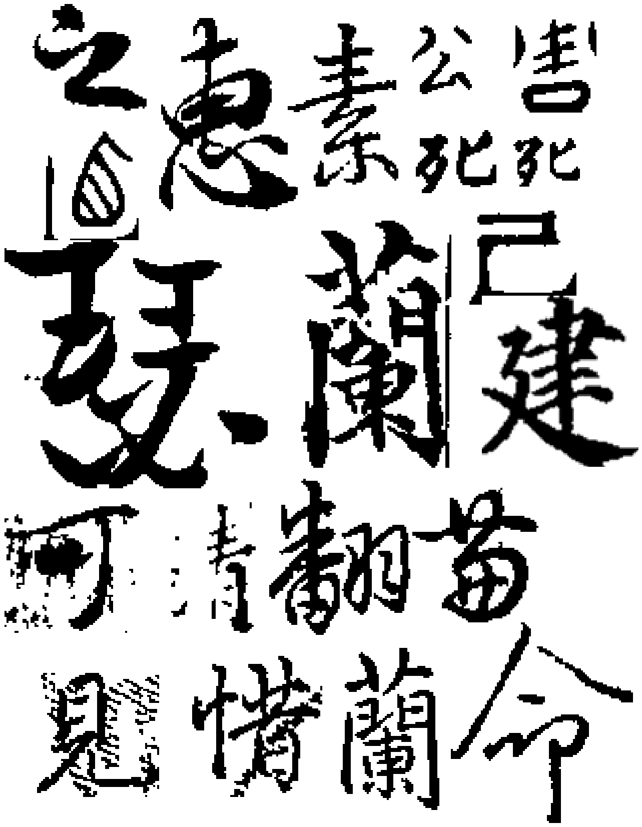


Figure 4. Characters printed at the same scale to show variations in size.

The interfaces to external metadata, including catalog entries for the source page, title of the book, publication date, and the author, are shown in Fig. 5.

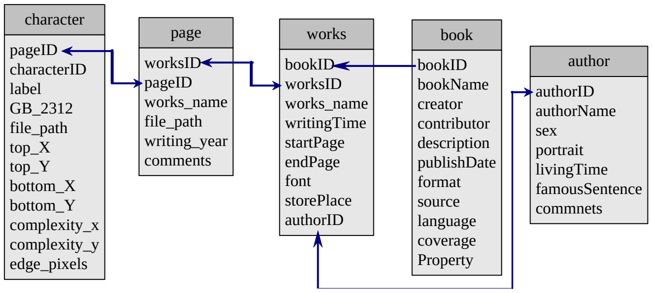


Figure 5. Architecture of calligraphy data. Embedded URLs point to both additional bibliographic information and to the digitized books themselves.

The data is stored in an Access database. Some of the internal metadata for 20 characters from two books and two works can be seen in the spreadsheet of Table I. The three columns on the left show the source of each character image. The four columns on the right are the top-left and bottom-right coordinates of its minimum bounding box. Only part of the file path, which includes the character identification (CID), is shown here.

Table I. An Extract from the Calligraphic database

bookID	worksID	pageID	CID	label	GB2312	file_path	X1	y1	x2	y2
6100015	126	261	9289	奉	47086	00009289.	393	390	457	479
6100015	126	261	9292	个	45755	00009292.	386	562	440	613
6100015	126	261	9293	垩	51922	00009293.	381	631	436	703
6100015	126	261	9294	泠	49380	00009294.	507	266	576	321
6100015	126	261	9295	为	52906	00009295.	508	484	569	567
6100015	126	261	9296	宗	55258	00009296.	505	594	560	662
6100015	126	261	9297	伏	47100	00009297.	503	336	572	388
6100020	127	42	9298	啄	55236	00009298.	179	55	244	96
6100020	127	42	9299	峰	47077	00009299.	173	367	233	469
6100020	127	42	9300	讵	55978	00009300.	175	495	222	540
6100020	127	42	9301	得	46531	00009301.	179	566	219	621
6100020	127	42	9302	金	48624	00009302.	171	647	229	689
6100020	127	42	9303	磬	50637	00009303.	176	705	228	764
6100020	127	42	9304	雨	54250	00009304.	273	129	316	176
6100020	127	42	9305	羞	53471	00009305.	272	195	311	253
6100020	127	42	9306	掀	53190	00009306.	266	292	318	335
6100020	127	42	9307	轻	51169	00009307.	264	372	315	423
6100020	127	42	9308	溥	45473	00009308.	267	444	317	524
6100020	127	42	9309	风	47079	00009309.	271	556	314	606
6100020	127	42	9310	幸	53458	00009310.	272	645	305	696

3. STATISTICAL SUMMARY

3.1 Books and Works

Only five books and two works consist of more than 1000 characters. The distributions for the larger books and works are shown in Figs. 6 and 7 respectively.

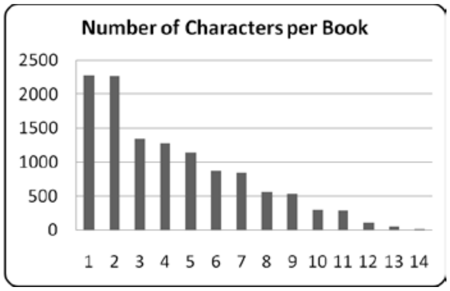


Figure 6. Nine books account for about two-thirds of the data.

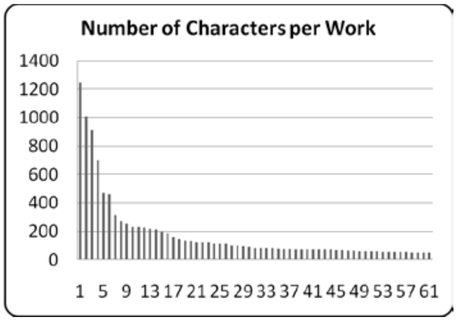


Figure 7. Number of characters in works with more than 50 characters.

Only 62 of the 208 works have more than 50 characters. The source of 47 characters from the web is not identified. The densest page has 235 characters, and there are 128 pages (out of a total of 483) with fewer than 10 characters.

3.2 Labels

The most common label (GB 54958 - 之) occurs 392 times among the 13,351 characters (~3%). There are 9064 instances of labels that never appear more than once in any work, which clearly makes it essentially impossible to divide each work into a meaningful training and test set for classification.

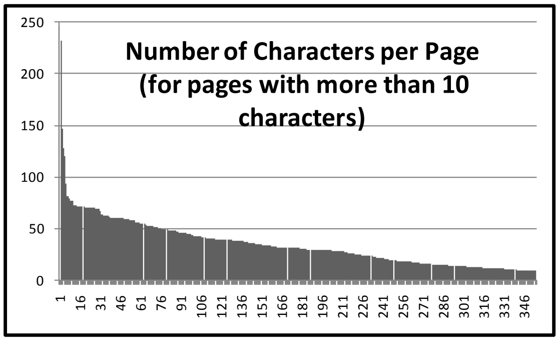


Figure 8. Character density per page.

Many demographic and linguistic phenomena (such as the populations of towns and word frequencies) obey Zipf's Law. To test whether the label frequencies do, Fig. 9 shows them on a log-log plot. It is seen that initially the frequencies do not fall off as rapidly as 1/n. Perhaps the poetic nature of many works forces more repetitions than plain text.

3.3 Size and Shape

The average height and width of the character images are both 59 pixels. The maximum width is 460 pixels, and the maximum height is 413. The average aspect ratio is almost exactly unity, with a standard deviation of 0.3. The fraction of foreground pixels is 32% of the bounding box area.

The average number of four-connected foreground components per image is only 3.5 (including noise specks.

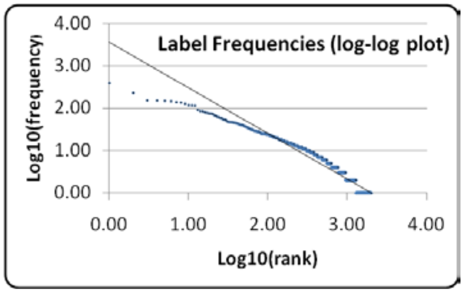


Figure 9.Frequency distribution of character labels.

The measure of character complexity is the number of intersections of the character strokes with uniformly spaced transects. The average number of intersections with horizontal transects is 2.10, and with vertical transects 2.27. Spacing and location of the transects are shown in Fig. 10.

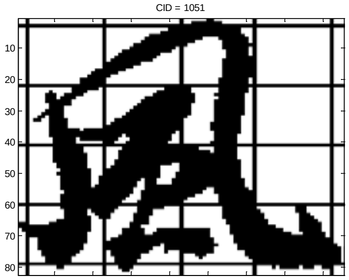
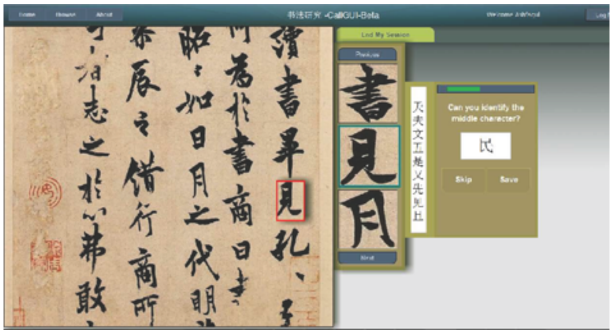


Figure 10. Complexity. This character has 14 intersections with the five horizontal transects and 10 intersections with the five vertical transects.

4. Interactive Labeling

CalliGUI is a prototype classifier-assisted interactive labeling interface. Experimental results on a PC platform, three subjects, and three works, reported in [12], show that label entry with CalliGUI is more than twice as fast as Pinyin keyboard entry. The assistance provided by the classifier consisted of the top 25 label candidates generated after training on the database but excluding the current work. If the character to be labeled appeared on this list, the operator could just click on it (this happened about 77% of the time). Otherwise the character was typed in pinyin or, if unrecognized, skipped (about 2% of the characters.)

Our new web interface built with ASP.NET is shown in Fig. 11. It is superficially similar to that reported in [12] but contains changes in the links to the database, improved logging facilities, more consistent tracking of the characters on the page, better layout, and improved ability to accommodate the user's browser display. On the left is an image of a page in running style from which individual characters were segmented. This display gives the operator the relevant context for identifying ancient, rare, unfamiliar, or deformed characters. A red box indicates the location of the current character to be labeled. The vertical box in the center is a list of candidate labels obtained by a nearest neighbor classifier by comparing the query image (in the red box) with already labeled images in the database. The candidate labels are shown in Song-style font.



Can you identify the middle character?

Figure 11. CalliGUI web interface for labeling character images. The user can (1) click on the appropriate candidate from classification, (2) type in pinyin the correct label, or (3) skip the character image.

5. Discussion

We reported the characteristics of a new calligraphic database. This database contains 13,351 characters from 21 books and 208 ancient and famous works. Among the 2010 distinct GB labels, 721 have only a single image sample, and 9064 characters have labels that appear at most once in any work. The styles of calligraphy represented in the database are *Great Seal, Small Seal, Clerical, Regular, Running and Cursive,* with many transitions among them.

The differences in shape and style among the works and the skewed label frequency distributions impose a barrier to the application of any trainable classifier. Nevertheless, a two-stage nearest neighbor classifier offered high enough Top-25 accuracy to reduce the time for interactive labeling by a factor of more than two. Experiments are underway to test an improved version of our prototype labeling interface with users from a wider variety of background in calligraphy.

In addition to experimentation on recognizing calligraphic text, the database is designed for research on shape-oriented content based image retrieval. In this scenario, instead of a label, the user input is either a scanned character image or a mouse or stylus drawn character.

Other potential lines of investigation are duplicate and forgery detection. The variety calligraphy in the database can also facilitate research on the evolution of calligraphic styles. All such projects would, of course, benefit from further expansion of the database (Fig. 12).

Figure 12. Calligraphy awaiting digitization at CADAL scanning center.

The CADAL database is a work in progress. The segmentation is being improved and more characters are being added. When the current round of improvements is completed, the database will be submitted to calligraphic scholars who will correct any residual labeling errors and identify most of the remaining unknown characters. After these steps, the entire database will be posted for public access.

The current set of 13,351 labeled character images is available now from the first author in either full color or binarized format.

6. Acknowledgment

We thank RPI student Ashfaqul Islam for improving CalliGUI and gratefully acknowledge the support of the Rensselaer Center for Open Software and of the Science & Technology Program of Shanghai Maritime University

7. REFERENCES

[1] D. Doermann, H.P. Li, O. Kia, The detection of duplicates in document image databases, IVC(16), No. 12-13, 24 August 1998, pp. 907-920.

[2] R. Dai, C-L Liu, B. Xiao, Chinese character recognition: history, status and prospects, Frontiers of Computer Science in China, Volume 1, Number 2, 126-136, 2007.

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

[4] X. Zhang, G. Liu, J. Wu, C. Luan, A Quick Search Engine for Historical Chinese Calligraphy Character Image, In Procs of 1st Int'l Congress on Image and Signal Processing, pp. 355-359, 2008.

[5] 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.

[6] R.G. Casey and G. Nagy, Recognition of Printed Chinese Characters, IEEE Transactions on Electronic Computers, vol. 15, #1, pp. 91-101, February 1966.

[7] T. Su, T. Zhang, D. Guan, Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text, IJDAR 10:27-38, 2007.

[8] Y.J. Liu, J.W. Tai, J. Liu, An introduction to the 4 million handwriting Chinese character samples library. In: Proceedings of the International Conference on Chinese Computing and Orient Language Processing, Changsha, pp. 94–97 (1989)

[9] H. Zhang, J. Guo, Introduction to HCL2000 database. In: Proceedings of Sino-Japan Symposium on Intelligent Information Networks, Beijing (2000)

[10] CADAL web site: http://www.cadal.zju.edu.cn (accessed 1/1/2011)

[11] Universal Digital Library web site: http://www.ulib.org (25/3/2005)

[12] G. Nagy, X. Zhang, CalliGUI: Interactive Labeling of Calligraphic Character Images, to appear in Procs. ICDAR11, Beijing, 2011.