CalliGUI: Interactive Labeling of Calligraphic Character Images

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

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

***Abstract*—Calligraphic data entry is accelerated by generating, with a feature-based character classifier, an ordered list of reference candidate labels for each character image. The improvement of labeling throughput depends on the top-N accuracy of the classifier, which in turn is a function of the available already-labeled patterns. Experiments on a database of 13,351 ancient calligraphic characters indicate that clicking on reference labels is more than twice as fast as Pinyin keyboard entry.**

Keywords-calligraphy, computer-aided labeling, shape-based recognition, top-N classification

I. INTRODUCTION

The combination of writing and fine arts has attracted artists from diverse cultures for centuries. Calligraphy was inscribed on stone, bamboo sheet, wood, silk scrolls and paper before the advent of bound books. Exquisite exemplars of Persian, Indic, Arabic, and Chinese calligraphy are preserved in museums throughout the world. Nevertheless, ancient calligraphy is known mostly through graphic reproductions in printed books. In addition to photographs of calligraphic text, such books typically contain interpretive notes about author/artist, historical context, calligraphic style, and technique.

Digital libraries broaden access to ancient calligraphy. Below we describe CalliGUI, a tool developed to enhance digital library holdings of Chinese calligraphy. Han logograms represent words or parts of words as opposed to the vocal patterns of alphabetic scripts. Scholars have identified over 50,000 distinct characters, but 4,000 of them cover over 99% of modern Chinese usage. Standardization has been accelerated by the advent of computer codes (analogous to ASCII) because only characters represented by code words can be readily stored or transmitted digitally.

Our objective is to help increase the granularity of access to Chinese calligraphy from the page-image level to the character-image level. Our source data consists of digitized pages of books of reproductions of original calligraphic works (usually stone rubbings or inked scrolls). First, the minimal bounding box coordinates of each character image on a page of the source book are determined and recorded. Second, the appropriate character label is entered, and stored as a 16-bit GB2312 code with its Pinyin equivalent. This completes the linkage between Dublin-core page-level bibliographic metadata and the graphical and symbolic contents of the book. Fig. 1 shows our data structure.

The novelty of this contribution is that CalliGUI makes use of computer image processing and character recognition to accelerate the above tasks. The remainder of the paper describes character image preparation, the functionality of the labeling interface, and observed performance characteristics of our prototype system. Some interesting options opened up by integrating a character recognition subsystem into CalliGUI are discussed in the last section.

II. CHARACTER IMAGE PREPARATION

A. *Source of Materials*

Most of the source books were published in recent decades and range in length from a few dozen to several hundred pages. The digitized page images are part of the first twenty books on calligraphy digitized by the China Academic Digital Associative Library (CADAL) [1,2] at Zhejiang University as part of the China-US Million Book Digital Library Project [3]. The bilingual multimedia (text, calligraphy, image, audio, video) services provided by CADAL are accessed several hundred thousand times per day by both Chinese and international parties.

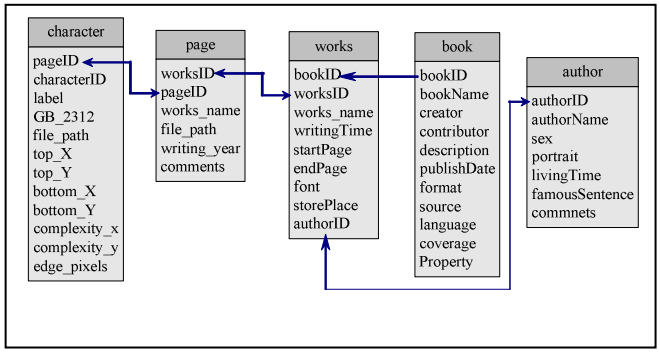
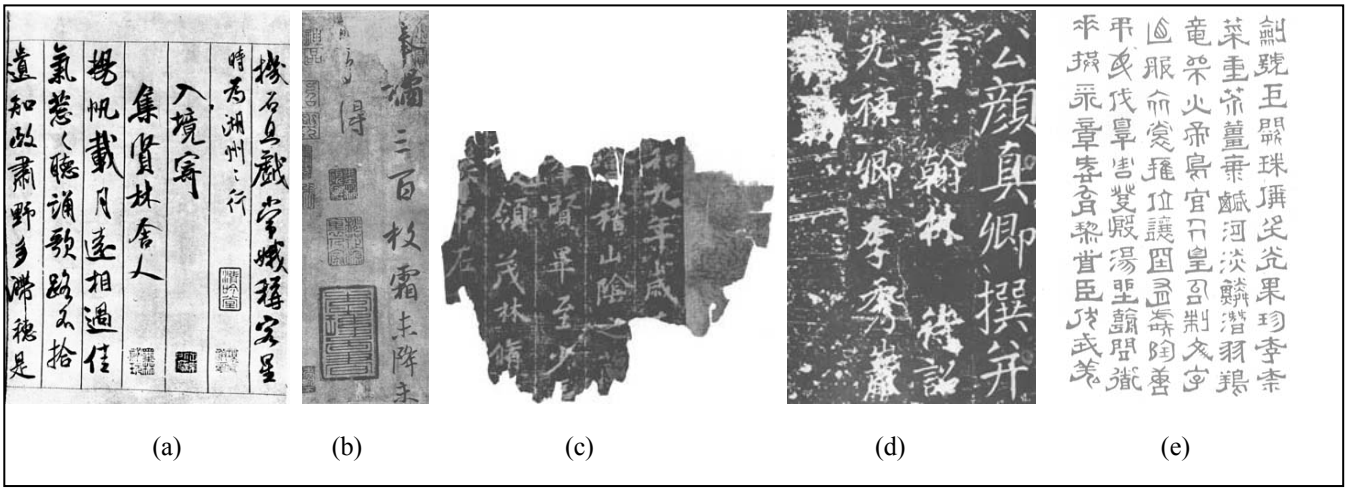


Figure 1. The calligraphy data is organized into five tables: *book*, *works, page, character* and *author*. A book contains many works, which may have consecutive pages. These five tables are related by the primary keys: *bookID*, *worksID, pageID, characterID* and *authorID*..

B. *D*igitization

Each page is scanned at 600 dpi into 256-level RGB TIF (for analysis) and JPG (for display) formats. Since the original scrolls may have to be separated into several pages,



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

Figure 2. 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) Smeared character should be omitted (e) Clean page that can be segmented automatically

a worksID is created in addition to the bookID and pageID. Individual character sizes vary from 40x30 (HxW) pixels to 400x300 pixels because illustrations in different books are presented at different magnifications.

C. Binarization

The RGB images are retained for display purposes, but the pixel amplitudes are globally thresholded to binary values. Wholly automatic segmentation is precluded by the darkness and variability of the background of the reproductions that reflect the aging of the original substrates (Fig. 2). Therefore the thresholds are automatically estimated based on the intensity distribution, but after inspection they are manually adjusted if necessary.

D. Segmentation

The reading order of most calligraphy is top-down, right-to-left. The pages are therefore first segmented into columns by finding the gaps between projections of the pixels onto the horizontal axis. Then each column is segmented horizontally by projections on the vertical axis. Finally the minimal bounding box of each character is located and given a characterID (CID). The column alignment of these hand-written characters is only approximate, and many works exhibit no row alignment. Finding the bounding boxes is also hampered by imperfect binarization, and even more by the presence of "stamps" of successive owners of the valuable original calligraphy (Fig. 2b). Additional information on preprocessing can be found in [4].

E. Feature Extraction and Classification

Two kinds of features are extracted from each segmented character: intersection counts with transects that divide the character images into broad overlapping clusters, and complex shape features based on the relative orientation of pairs of points [5]. The intersection counts are used to generate a candidate set for a computationally intensive Top-N shape-based classifier. The final classifier reports not only the labels of the patterns in the database that best match the query, but also the CIDs of the matching character images.

F. Experimental database

The database currently contains 13,351 characters with 2010 distinct GB labels from 207 "works." The earliest originate from 333 BC. The distribution of labels is skewed by usage: 721 of the 2010 GB labels have only a single image sample. The number of characters per work is distributed roughly according to Zipf's law, with a few large works (the largest has 1245 characters) and many small ones. Characters deemed illegible were excluded from the database. Books, works and pages have eight-digit IDs.

III. CHARACTER IMAGE LABELING WITH CALLIGUI

Fig. 3 shows our web interface for labeling new character images. The operator can either type in a label for the query image, or select a candidate label by clicking (possibly after scrolling down) to transfer it to the horizontal entry box in center right.

For typing in a label, the operator uses Pinyin, where successive keystrokes produce shorter and shorter lists of candidates. Eventually the operator presses a numerical key to select the correct candidate (which, as with the classifier-generated list, is not always in the top position).

Whether or not the final label was selected from the top recognition candidates or typed in, the operator must click on the SAVE button to associate the label with the query image. Clicking on SKIP indicates that the operator was not able to identify this character. Skipped characters will be sent to an expert for labeling. After clicking the SAVE or SKIP button, the next query character image in the normal reading order will come up.

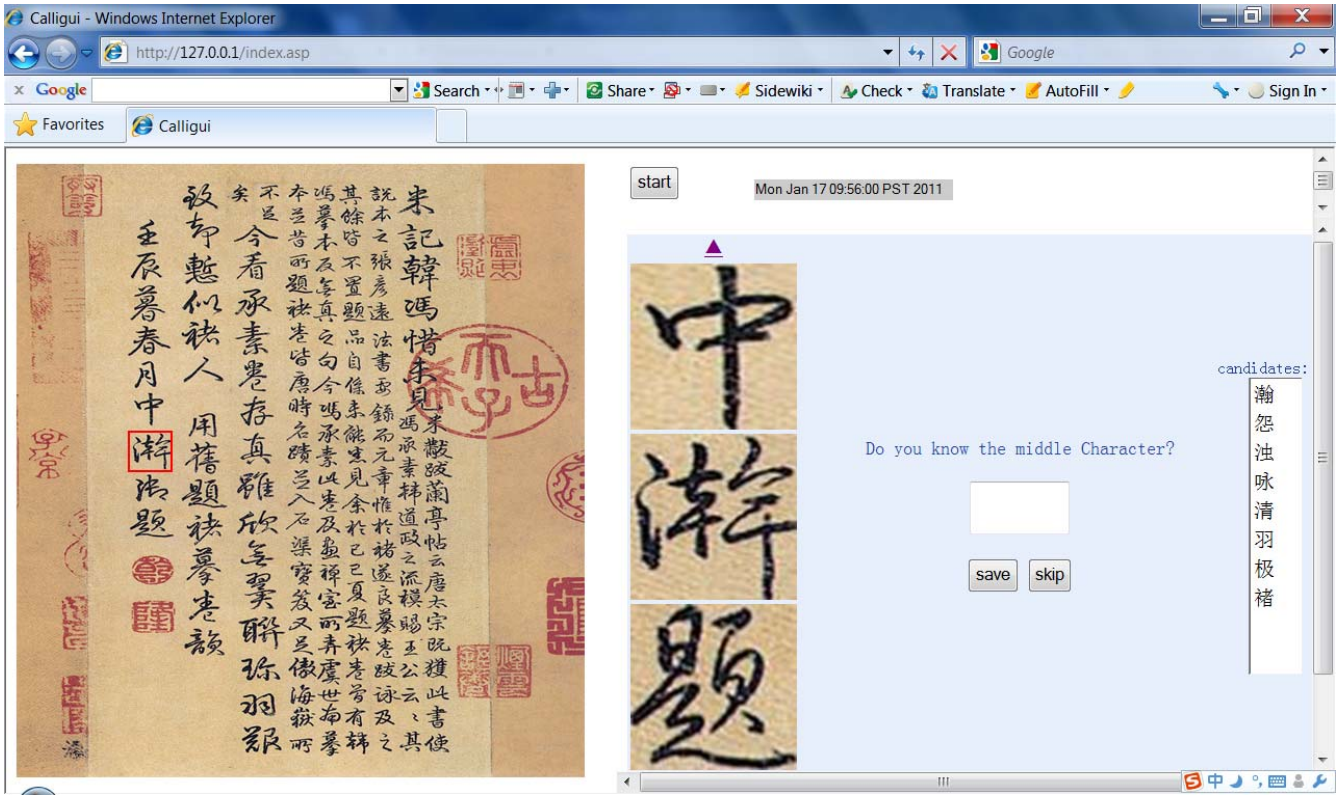


Figure 3. Web interface for labeling character images. On the left is a JPG image of a page in running style from which individual characters were segmented. A red box indicates the location of the current character to be labeled. This display gives the operator the relevant context for identifying ancient, rare, unfamiliar, or deformed characters. The vertical box on the far right 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.

Logging the session begins when the operator presses the START/STOP button after entering his or her name and password, and ends when STOP is pressed. The system keeps track of the current character to be labeled so that the operator can resume at the stopping point later. The primary purpose of the logging system is to determine how much computer assistance raises labeling throughput, and more specifically the relationship between the performance of the recognition subsystem and the speed and accuracy of the labeling task. A secondary purpose of timing operator actions is to facilitate future experimentation on subject-specific factors like education, familiarity with calligraphy, Pinyin keyboarding skill, and perhaps motivation.

time	characterID	label	GB_code	user_label	user_GB	action_type	user_name
20:11:46	1058	日	51413	日	51413	click_ch	Xiafen
20:11:48	1057	也	53938	也	53938	click_ch	Xiafen
20:11:50	1056	天	52460	天	52460	click_ch	Xiafen
20:11:55	1055	朗	49354	朗	49354	save_ch	Xiafen
20:11:57	1054	气	50936	气	50936	click_ch	Xiafen
20:11:58	1053	清	51173	清	51173	click_ch	Xiafen
20:12:0	1052	惠	48093	惠	48093	click_ch	Xiafen
20:12:1	1051	风	47079	风	47079	click_ch	Xiafen
20:12:05	1142	和	47821	和	47821	save_ch	Xiafen
20:12:10	1144	畅	45993	畅	45993	save_ch	Xiafen
20:12:11	1070	观	47579	skip	29547	skip_ch	Xiafen
20:12:13	1069	宇	54254	宇	54254	click_ch	Xiafen
20:12:15	1068	宙	55014	宙	55014	click_ch	Xiafen
20:12:16	1067	之	54958	之	54958	click_ch	Xiafen
20:12:18	1066	大	46323	大	46323	click_ch	Xiafen

Figure 4. Part of a session log for 15 consecutive characters

The timing of the essential actions in labeling every character is logged. Since we wanted to avoid a variable number of time stamps for each character, only the time of the last click (if any) on a recognition candidate is kept. Actions for which the time is recorded are therefore START, STOP, CLICK, SAVE, and SKIP. The CID, the saved label, and true label of the query (available for our experiments from the database) are also recorded. Comparison of the operator-assigned label with the true label of the query yields the error rate of the labeling process. Fig. 4 is a snapshot of 15 consecutive characters from a session log.

IV. EXPERIMENT DESIGN

Three subjects (one of the authors and two graduate students who completed their undergraduate degrees in Xi'an and Hong Kong) used CalliGUI to perform the following experiments, which altogether took about ten hours. Repeated processing of the same works by the same subjects does not affect our results, because visual label recognition, in contrast to *label entry,* is quasi-instantaneous.

Experiment E1. Type in the character labels of three famous works without computer classified reference candidates. This provides the baseline time for label entry using the conventional method of transcribing handwritten Chinese characters.

Experiment E2. Enter the labels of the same works with recognition candidates obtained by a classifier trained on all the remaining works in the database. This experiment will determine whether an existing database of labeled calligraphy can speed up labeling new character images over the conventional manual method.

Experiment E3. Same as *E2*, but with the candidate labels generated for each image using only the other characters in the same work. Training on similar characters should raise classifier accuracy.

Table I shows the title, number of query characters, and the average rank of the correct candidate generated with each training set. Since CalliGUI displays at most 25 reference characters, the maximum rank, even for characters with labels absent from the training data, is 26. In all three works, over 95% of the query characters have character images with a label that also appears in the remaining works.

The advantage of training on the same work (*E3*) is mitigated by the fact that every character that appears only once in the work is misclassified because there is no training character with the same label. Larger works clearly fare better in *E3*.

TABLE I. RELEVANT CHARACTERISTICS OF THREE WORKS

Title	#chars	AveRank *E2*	AveRank *E3*
Lanting Xu	114	3.5	19.3
Shu Su Tie	273	8.7	16.3
Duobao Ta Bei	470	8.1	12.5

We believe that the quantum unit for entering new characters will usually be an entire work by the same calligrapher. If training on the same author rather than on the entire database does lead to faster or more accurate labeling, then future versions of the system should incrementally retrain the classifier after each character or small group of characters is entered. The underlying assumption here is that most of the characters will be labeled correctly by the operator even if the initial Top-1 accuracy of the classifier is far from 100%.

Larger experiments with more subjects are currently underway. We expect to be in a position report results with greater statistical significance by September 2011. We are also adding label context from the two characters preceding the query. We expect that label context, already widely used in OCR, will significantly improve Top-N accuracy.

V. EXPERIMENTAL RESULTS

The average entry times for clicked and for typed characters, and the percentage of characters entered each way, are shown in Table II. *Clicked* here means entered from the recognition candidates, and *typed* means that the operator typed a label using Pinyin. As expected, entry times are lowest with "suggestions" from the best trained classifier (*E2*). Classifier assistance also decreases by 35% the fraction of characters that remain unlabeled. The number of mislabeled images (0-2 per subject) is negligible compared to the number of skipped images.

TABLE II. EXPERIMENAL RESULTS

Experiment & Subject		Average typing time (s)	Average clicking time (s)	Typed %	Clicked %	Skipped %	Average entry time (s)
Expt E1	**S 1**	3.74	N/A	98.6	0	1.4	3.85
	S 2	5.22	N/A	94.3	0	5.7	5.43
	S 3	4.74	N/A	97.2	0	2.8	4.87
Expt E2	**S 1**	5.23	1.82	33.0	65.4	1.6	3.01
	S 2	4.87	3.31	14.3	82.3	3.4	3.75
	S 3	6.58	2.91	15.6	82.5	1.9	3.72
Expt E3	**S 1**	3.93	1.38	54.7	44.6	0.7	2.83
	S 2	6.77	1.73	50.9	45.4	3.6	4.54
	S 3	5.45	1.95	54.6	44.6	0.8	3.91
All Subjects							
	Expt E1	4.57	N/A	96.7	0	3.3	**4.72**
	Expt E2	5.56	2.68	21.0	76.7	2.3	**3.49**
	Expt E3	5.38	1.69	53.4	44.9	1.7	**3.76**

Figure 5. Examples of query characters with operator consensus. (a) Skipped because unrecognized (b) Clicked because one of the reference candidates was correct (c) Typed either because label missing in training set or because the cursive character was misclassified.

Although fewer labels can be clicked when the classifier is trained on the same work, clicking is significantly faster because the correct label usually appears near the top of the list where it is easy to see.

Subject #1 was significantly faster and skipped less than the other two. One subject, from Hong Kong, was less familiar with the Pinyin typing system and benefited most from clicking. Fig. 5 shows samples of characters skipped, clicked, or typed by all three operators.

VI. OTHER POSSIBLE APPLICATIONS OF CALLIGUI

The combination of automated segmentation and classification with human interaction opens up several applications in addition to efficient calligraphic data entry.

(1) *Duplicate detection.* When presented with a query character that is already in its training set, the nearest-neighbor classifier reports 100% similarity with the Top-1 candidate. It can therefore be used for finding duplicates in a database or for avoiding entering duplicates in the first place. Calligraphic images are now widely reproduced on the web, which makes this an actual rather than a hypothetical problem [6].

(2) *Forgery detection.* Some works of calligraphy, like paintings and sculpture, attract a high price from collectors and are therefore frequently counterfeited. Given a suspected page of calligraphy, CalliGUI can be used to compare it to calligraphy from the putative author that is already in the database. If the database contains some character images with the same label as the suspect work, then the task is easy. If it does not, then the classifier may still return a set of characters from the database that are similar to those of the suspect work. These can be visually compared with regard to stroke geometry and configuration. Purely visual comparison can be enhanced with objective numerical comparison of shape features extracted from true and suspect character images [4].

(3) *Style Classification.* Unlike a statistical classifier that estimates only the posterior probability of each label, the nearest-neighbor classifier reports the character images in the database that are most similar to the query. Therefore the style of the work that is the source of the majority of the Top-N candidates of the nearest-neighbor classification of an unknown sample page is likely to have a calligraphic style similar to that of the query. In order to provide explicit style names (e.g., *Great Seal, Little Seal, Clerical, Regular, Cursive*), the style of each of the works in the database must be identified by an expert. (Although a calligrapher may use different styles for different works, style is usually a characteristic of an entire work [7].)

(4) *Calligraphy Retrieval.* Web users often wish to identify unlabelled calligraphic images, to inspect calligraphic images of the same style as their sample, or to compose messages in a given calligraphic style [8]. All of these applications are based on interactive image-based character classification and can be facilitated by a widely accessible web interface with this functionality. With some modifications, the interface could be extended to non-Chinese calligraphy.

(5) *Transcription of modern Chinese text.* Current scanners, fax machines and multifunction printers can all produce page images of writing at a spatial sampling rate (dpi) adequate for classification [9]. It is not unusual to have to enter long passages of text from a printed or hand-written page on which current OCR produces too many errors. With a database of representative characters, a system based on semi-automatic segmentation and recognition, like CalliGUI, may provide an alternative solution to plain Pinyin.

REFERENCES

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

[2] Calligraphy in CADAL (accessed 1/1/2011): http://www.cadal.zju.edu.cn/CalligraphyWeb/listBooks.action

[3] Universal Digital library web site: http://www.ulib.org (1/1/2011)

[]4 G. Nagy, X. Zhang, The CADAL Calligraphic Database, submitted to HIP Workshop, ICDAR 2011.

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

[6] 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.

[7] 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.

[8] X. Zhang, G.zhou 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.

[9] 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.