Unsupervised Extraction of Training Data for Pre-Modern Chinese OCR

Donald Sturgeon
Fairbank Center for Chinese Studies
Harvard University
djs@dsturgeon.net

Abstract

Many mainstream OCR techniques involve training a character recognition model using labeled exemplary images of each individual character to be recognized. For modern printed writing, such data can be easily created by automated methods such as rasterizing appropriate font data to produce clean example images. For historical OCR in printing and writing styles distinct from those embodied in modern fonts, appropriate character images must instead be extracted from actual historical documents to achieve good recognition accuracy. For languages with small character sets it may feasible to perform this process manually, but for languages with many thousands of characters, such as Chinese, manually collecting this data is often not practical.

This paper presents an unsupervised method to extract this data from two unaligned, unstructured, and noisy inputs: firstly, a corpus of transcribed documents; secondly, a corpus of scanned documents of the desired printing or writing style, some fraction of which are editions of texts included in the transcription corpus. The unsupervised procedure described is demonstrated capable of using this data, together with an OCR engine trained only on modern printed Chinese to retrain the same engine to recognize pre-modern Chinese texts with a 43% reduction in overall error rate.

Introduction

Research into Chinese OCR techniques has tended to focus on OCR of contemporary documents, particularly modern printed and handwritten documents (e.g. Huo et al. 2001, Su et al. 2009, Zhang et al. 2009, Liu et al. 2013, Zheng et al. 2013, Chen et al. 2014) with a few notable exceptions (e.g. Yang and Peng 2013). While much of this work is also applicable to OCR of pre-modern documents, there are nevertheless both challenges and opportunities specific to the premodern OCR problem which may benefit from greater attention. As the cost of digitization and digital image storage continues to decrease and libraries continue to digitize ever larger portions of their historical collections, the need for OCR techniques more specifically suited to pre-modern documents increases also. At the same time, as digital libraries of pre-modern materials grow in size, it is inevitable that their contents gradually come to represent a larger fraction of the total extant historical corpus. As a result, opportunities for exploiting similarity within a historical corpus of material for purposes including OCR bootstrapping are also continually increasing. In the case of historical Chinese texts in particular, one aspect of this is the existence of and desire to digitize and transcribe multiple editions of texts instantiating the same abstract work for the purpose of scholarly comparison of editions.

One particular challenge for pre-modern Chinese OCR is the relative difficulty of assembling appropriate training data. Modern printed documents are typically created digitally, and font data for typefaces similar or even identical to the style of writing to be recognized is often easily available. This allows for almost trivial generation of arbitrary amounts of synthetic training and test data for modern printed OCR. This type of font data is not typically available for pre-modern writing styles; in order to obtain high levels of accuracy on historical documents, training data must normally be assembled manually.

This problem is further exacerbated for languages such as Chinese by the large character set and need to identify adequate exemplars of each distinct character. Pre-modern Chinese OCR requires the recognition of – at a minimum – 3000 or more distinct characters. These characters, however, are not all equally common, with some occurring very frequently and some far more rarely. As a result, whereas manual markup of tens of characters requiring consultation of only a few tens of pages of English writing might be sufficient to create training data for a pre-modern English writing style, for Chinese the equivalent process would require thousands of characters to be manually identified, and identifying these character exemplars would likely require the examination of tens of thousands of pages of text.

At the same time, many pre-modern forms of Chinese writing have a high degree of similarity with modern printed writing forms, to the extent that character recognition models trained on modern printed character data can often produce meaningful results on pre-modern data, though with significantly reduced accuracy. This paper presents a methodology for exploiting this fact together with existing corpus data to generate large amounts of training data for pre-modern Chinese OCR. The unsupervised approach presented takes as input two unaligned, unstructured, and noisy datasets: firstly, a corpus of transcribed historical documents; secondly, a corpus of scanned documents of the desired printing or writing style, some fraction of which are believed to be editions of texts included in the transcription corpus. After initial image pre-processing and character segmentation stages, the training data extraction procedure consists of: 1) applying a character recognition model trained purely on modern printed documents to pre-modern documents to generate a low-accuracy, intermediate OCR result; 2) making use of this intermediate result to align the image data with its probable transcription; 3) extracting labeled character images on the basis of this alignment; 4) identifying suitable training exemplars from these labeled character images. The data produced can then be used directly without review to train a new character recognition model achieving significantly higher accuracy on similar material.

Source Data

The results presented in this paper use texts from the *Siku Quanshu* collection for both training and evaluation because of their convenience, consistency, and availability; it is also one of the most well-known collections of pre-modern Chinese writing (Wilkinson 1998:262), and is available online.[1] The texts in this collection were transcribed by hand by many independent copyists in the 18th century, though the work of these copyists remained remarkably consistent in terms of writing style. In this evaluation, only a small fraction of the complete corpus (four texts for training data extraction and one for evaluation, out of over 3000 total) is used, in order to demonstrate the viability of the procedure even with relatively small input corpora. Scanned images of the *Siku Quanshu* texts as well as transcriptions of these texts were obtained from the Chinese Text Project digital library.[2]

The methodology presented in this paper has also been successfully applied to other historical writing styles such as woodblock printed texts; the only pre-requisites to applying it are sufficiently large corpora written in a consistent style.

Image pre-Processing

Many pre-modern Chinese texts follow a relatively fixed layout: text is organized into columns, each read from top to bottom, the columns themselves being read from right to left. Often (though not always) the divisions between these columns are marked by vertical lines drawn on the page. Very frequently, one or more solid borders are also drawn around the region of the page used for writing. Additionally, as many "pages" in the sense considered here were originally halves of much wider sheets printed together, at the far left or right edge of each page is frequently found a column of information (part text, part non-text, possibly containing information about the volume, section, page number etc.) which should be ignored when reading the contents of the text itself – i.e. this information is not part of the primary text flow, and is thus typically not part of the desired transcription when attempting to transcribe an entire text.

All of these aspects of typical page layout can be leveraged at the image pre-processing level to significantly simplify the OCR problem for pre-modern Chinese texts. To perform this step, each image was first deskewed so as to ensure that any lines and borders present became as close to perfectly horizontal or vertical as possible. Next, projection-based algorithms were applied to the transformed image to identify contiguous and fragmentary horizontal and vertical line segments of lengths sufficiently long that they could not plausibly be character components, and the lines identified erased from the images. Additional heuristics were used to remove all content bounded by a detected vertical line close to an edge, as these marks typically do not correspond to any part of the desired transcription and frequently include both incomplete character fragments and non-character marks. This procedure resulted in input images of the type shown on the left of Figure 1 being transformed into images of the type shown on the right of the same figure.

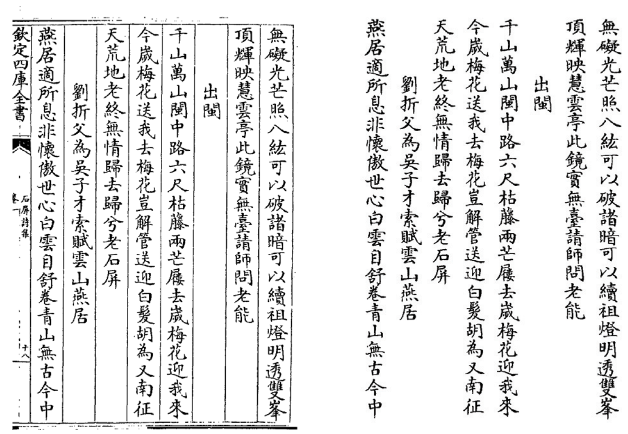

Figure 1: Page image before (left) and after (right) line removal.

[1] http://ctext.org/library.pl?if=en&collection=4

[2] http://ctext.org

Character Segmentation

While many pre-modern Chinese texts at first glance appear to follow a simple grid-like layout, in which characters are aligned both horizontally and vertically and/or have largely constant width-height ratios – properties which would greatly simplify character segmentation – in general the reality is much more complex (Heijdra 2006). Character size ratios can vary significantly, and grid-based layouts, though common, are far from universal. Columnar layouts are, however, near universal in these works, and thus column segmentation was performed as the first step towards character segmentation. Starting with the pre-processed images, column identification was again accomplished using a projection-based algorithm (Yang and Peng 2013, Liu et al. 2014) together with simple assumptions about plausible column widths, taking advantage of the whitespace separating columns after line removal. A similar procedure was then applied within each column to identify characters, this time also taking into account constraints of plausible character dimensions.

Training Data Extraction

In contrast to modern printed fonts, for which OCR training data can be easily created directly from existing font data, in order to train many types of character recognition model to accurately recognize pre-modern forms of Chinese writing such as woodblock print or a particular style of handwritten script, labeled character images of each individual character to be recognized must first be assembled. This is challenging for Chinese due to the large numbers of characters, as well as additional constraints on images used for training data important when these are obtained from real-world historical texts: each labeled character must be a "typical" instance (i.e. not an unusually proportioned or drawn instance of the character it represents due to, for instance, the slip of a scribe's writing implement), properly segmented, and free of noise and damage (e.g. due to poorly inked woodblocks, subsequent physical damage to the page, or errors introduced during digitization). These constraints, together with the fact that not all Chinese characters are equally frequent – some will be highly infrequent, and most works will not contain all characters – mean that manual assembly of such training data for any given writing style would be an extremely time-consuming task requiring consultation of many thousands of pages of text in order to identify and label several thousand character instances.

To address this difficulty, a fully unsupervised procedure for obtaining large amounts of suitable training data was developed. This takes as input two collections of material: 1) a set of transcribed texts, each labeled with a title; 2) a set of scanned texts, each labeled with a title. No strong assumptions are made about the relationships between the transcriptions and the scanned texts: a transcription and a scanned text may have the same title, yet represent entirely different works, or may be based upon editions of the same text containing similarities as well as significant differences.[3] The weak assumption required for the procedure is that *some* of the transcribed texts are broadly similar to some scanned texts labeled with the same title.

The procedure consists of four main stages: 1) first-pass OCR on the scanned texts, using an OCR model trained on modern printed Chinese; 2) unsupervised alignment of existing transcriptions to first-pass OCR output; 3) extraction of probable character labelings from the alignment data; 4) selection of appropriate character exemplars for use in OCR training.

For the purposes of this paper, first-pass OCR was performed using the image processing and character segmentation methods described above, followed by the open source Tesseract engine, which provides a pre-trained model for modern printed Chinese used to perform recognition of individual characters.[4] In the work presented here, Tesseract is used only as a character recognition engine, recognizing individual character images only; thus other features of Tesseract such as page layout and segmentation, language modeling, etc. are not used.

Alignment

Input data consisted of sequences of page images and full-text transcriptions, which were neither aligned nor expected to be alignable in all cases. Given a candidate pair of a sequence of images and a corresponding transcription selected according to similarity of title, the following procedure was performed to either align the pair if possible, or else reject the candidate pair. Firstly, all pages of image data were processed and segmented as described above, and character recognition attempted using the existing OCR model for modern printed Chinese provided with the Tesseract OCR engine. As expected, this data contained many misrecognized characters, and achieved an overall character accuracy of around 78%. This initial data was firstly used to estimate the approximate number of characters per page c in the image data for the given candidate text.

[3] A natural though more computationally expensive extension to this procedure involves ignoring (and thus not requiring) title labels of both sets, and simply comparing all transcribed texts with all scanned texts. This approach may be appropriate for lower-resourced languages for which less source material is available.

[4] https://github.com/tesseract-ocr

Alignment between the existing transcription of the entire work and the sequence of scanned page images was then attempted firstly at page-level granularity. The first stage of this process involved locating the first page of image data corresponding to the beginning of the transcribed text. For the texts considered, this was typically not the first page of the image data, nor in general the first page of image data containing a significant amount of text, because the scanned materials often contained additional information (such as cover pages, contents pages, prefaces, etc.) not present in the available transcriptions. To locate the starting page, pages n and n+1 of raw page image OCR were successively compared with the first 2c characters of the chosen transcription; the page s for which the concatenation of OCR data for pages s and s+1 had the lowest edit distance to the expected transcription (i.e. minimum total number of single-character insertions, deletions, and substitutions required to transform the raw OCR output into the expected transcription) was chosen as the start page.

The remainder of the matching process consisted of successively identifying the locations in the transcribed data most probably corresponding to page boundaries in the image data by comparison with the raw OCR results for each page. To do this, 2c characters of the transcription were aligned with the raw OCR data for pages n and n+1 by means of Longest Common Subsequences (LCS), and the corresponding boundary according to this alignment taken as the probable boundary within the transcribed data between pages n and n+1 of the image data. At the same time, a quality score was calculated for each page on the basis of this alignment and the edit distance between raw OCR and expected page transcription given the estimated alignment. This process was then repeated for page n+1 with the next 2c characters of transcription starting from the estimated page boundary between pages n and n+1.

If more than 5 consecutive pages m, m+1, … m+4 were found to have low quality scores suggesting that the alignment was not proceeding successfully, the procedure for identifying the start page was repeated, this time searching for the 2c characters of data expected to be found on pages m and m+1 according to the transcription and previously identified page alignments. If such a page could be found, the matching process resumed from that location; if not, the alignment was considered unsuccessful, and the candidate pair of image data and textual transcription rejected entirely.

For texts which aligned successfully at page granularity, a similar procedure was then performed within each page to successively align each column of OCR output with part of the expected transcription for that page. Thus the procedure resulted in large numbers of columns of text paired with probable transcriptions of these columns.

Character Labeling Extraction

The first stage of character label extraction was to reject data for any individual page with a low quality score (i.e. high edit distance between raw OCR and expected transcription according to the identified alignment) so as to avoid introducing unnecessary noise. Next, the raw OCR and expected transcription for each column were aligned using LCS. Within this alignment, aligned segments $A_1B_1C_1$ of the existing transcription and $A_2B_2C_2$ of the raw OCR data were compared; cases where A_1=A_2, C_1=C_2, and B_1 and B_2 each contained a single character were used to infer that the correct labeling for character B_2 in the OCR output (and corresponding to a known region of image data) should be B_1. A second pass of the same process was performed after these inferences had been made to extract additional labelings from the alignment data (Figure 2 rightmost two columns).

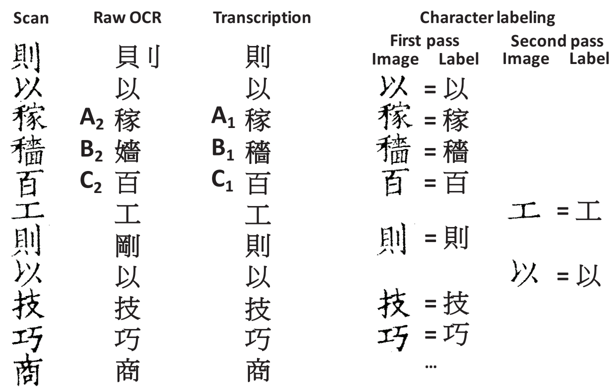

Figure 2: Character image label extraction from aligned OCR and transcription.

Finally, in order to minimize introduction of noise due to vertically adjacent character images within a column not being cleanly separable from one another due to overlapping elements or parts of the two character images being connected, cases in which the optimal horizontal line dividing two vertically adjacent characters from one another as determined by the segmentation algorithm crossed any black pixel of image data were excluded from the set of extracted character labelings used to create training data (Figure 3).

Figure 3: Cleanly separable character images (left), images with overlapping elements (middle), and connected images (right).

This procedure resulted in the extraction of a large volume of labeled character images, including labelings for characters completely unrecognizable by the OCR engine used to generate the initial intermediary OCR results. From the four successfully matched texts used in this experiment together totaling 6148 pages and 803,896 characters, 340,017 labeled instances in 5582 character classes were extracted.

Character Exemplar Selection

Not all correctly segmented and labeled characters are suitable training exemplars, and a small proportion of characters identified may be mislabeled (in practice less than 1% of character images were observed to be mislabeled). In order to avoid selecting as training exemplars character images which had been poorly segmented, contained noise, or were incomplete, damaged, or mislabeled, the set of all character images for each character class was first sorted in order to facilitate selection of the most "typical" exemplars of the class. To do this, each character image was first normalized to a fixed size, and feature vectors constructed based on 8-directonal projections (Bai et al. 2005, Zhang et al. 2012). Images were then sorted using dynamic time warping (Zhang et al. 2012), resulting in an ordering from most to least typical (Figure 4). After applying this ordering, the top three most typical characters according to the ordering were chosen as training exemplars. Manual inspection confirmed that in all cases the chosen character images were of the correct character class (i.e. no incorrectly labeled characters were present in the top three after sorting).

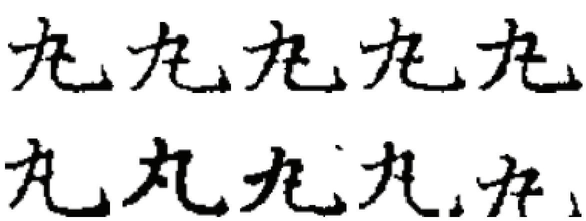

Figure 4: Most typical (top) and least typical (bottom) character images extracted for the character "丸".

Evaluation

In order to evaluate the practical utility of the procedure, the same OCR engine used to generate the initial alignment data was retrained using the extracted exemplars, and performance evaluated against test data not used in exemplar extraction. Two types of evaluation were performed: raw character accuracy on extracted labeled characters not used in training for each character class, and full-process accuracy in a complete OCR system using a text not used in any part of the procedure.

Raw character accuracy was evaluated by randomly selecting for each character class extracted labeled characters other than those used in training, manually reviewing these images to exclude clearly damaged or mis-segmented characters, and then evaluating OCR performance on these images using the trained model. In this evaluation, accuracy increased from 82% to 92%, a 55% reduction in raw character error rate.

To evaluate the effect on full-process accuracy, a practical OCR system including the image processing, character segmentation and recognition steps described above was implemented, together with character trigram language modeling (Zhuang et al. 2004) based on a 1 billion character corpus of pre-modern Chinese writing,[5] and incorporating information about frequently confused characters (Zhuang et al. 2004) easily extractable from data created during the present study, enabling a more meaningful gauge of the typical OCR accuracy improvement offered by the procedure. The number of errors present in the OCR output is calculated in terms of edit distance: the total minimum number of character insertions, deletions, and substitutions which would be required to transform the actual OCR output into the gold standard reference text. The rate of character accuracy of the procedure is then calculated as follows:

$$accuracy = \frac{characters - errors}{characters}$$

This value, expressed as a percentage, is reported as the full-process OCR accuracy, which for these historical works includes errors due to noise, damaged characters, and segmentation errors, as well as character recognition errors. Written Chinese does not use explicit markers such as spaces to delimit words; nor, typically, do pre-modern Chinese documents use punctuation or other marks to delimit phrases or sentences. As a result, this paper reports only *character* accuracy, which is easily calculated for this type of material. Full-process OCR accuracy increased from 78.9% to 88.0%, a 43% reduction in error rate.

Conclusions and Future Work

The procedure described offers a fully unsupervised method for extracting practically usable training data from noisy historical input data. As a result, it provides a mechanism to significantly improve both character-level and real-world

[5] Created using data from the Chinese Text Project digital library http://ctext.org.

OCR accuracy on historical documents by automated extraction of appropriate training data using existing character recognition techniques. The methodology makes few language-specific assumptions, and is particularly well-suited to languages with large character sets such as Chinese, Japanese, and Korean due to the practical difficulties of manually assembling large amounts of appropriate training data for these languages.

For historical texts, alignments between scanned materials and existing transcriptions of the type generated as part of this procedure are often also valuable by-products in themselves, since many pre-modern texts exist in multiple editions, all of which may potentially be of interest to scholars of such texts. Where a scanned edition corresponds closely to an existing transcription, the alignment information generated by this procedure can be used to implement full-text search of image data using the contents of the transcription in place of raw OCR results which would normally be used for this purpose. The ability to reliably extract large volumes of labeled character instances using an unsupervised procedure may similarly have other seemingly independent applications – for example, it raises the possibility of future statistical studies of changes in pre-modern Chinese writing practices over time based on the analysis of large amounts of labeled character image data available on such a scale as would be entirely impractical to assemble by hand.

A practical OCR procedure for pre-modern Chinese based upon the steps described in this paper has been implemented as part of the open access Chinese Text Project digital library, and has so far been used to transcribe and make searchable over 25 million pages of pre-modern Chinese material, the results of which are freely available online.[6]

References

Bai, Z., and Huo, Q. 2005. A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition. In *Proc. ICDAR 2005*.

Chen, J., Wu, Y., and Cao, H. 2014. Confusion Network Based Recurrent Neural Network Language Modeling for Chinese OCR Error Detection. In *Proc. International Conference on Pattern Recognition 2014*.

Heijdra, M. 2006. A Tale of Two Aesthetics: Typography versus Calligraphy in the pre-Modern Chinese Book. In Wilson, M., and Pierson, S. ed., *The Art of the Book in China*. London: SOAS 2006, p. 15-27.

Huo, Q., Ge, Y., and Feng, Z.D. 2001. High Performance Chinese OCR Based on Gabor Features, Discriminative Feature Extraction and Model Training. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

Liu, C., Yin, F., Wang, D., and Wang, Q. 2013. Online and Offline Handwritten Chinese Character Recognition: Benchmarking on New Databases. In *Pattern Recognition* 46 (2013) 155-162.

Liu, M., Suo, Y., and Yinan, D. 2014. Research on Optimization Segmentation Algorithm for Chinese/English Mixed Character Image in OCR. *Proc 4th International Conference on Instrumentation and Measurement, Computer, Communication and Control*.

Su, T., Zhang, T., Guan, D., and Huang, H. 2009. Off-line Recognition of Realistic Chinese Handwriting Using Segmentation-free Strategy. In *Pattern Recognition* 42 (2009) 167-182.

Wilkinson, E. 1998. *Chinese History: A Manual*. Cambridge: Harvard University Press.

Yang, L., and Peng, L. 2013. Local Projection based Character Segmentation Method for Historical Chinese Documents. *Proc. SPIE-IS&T Electronic Imaging*.

Zhang, H., Guo, J., Chen, G., and Li, C. 2009. HCL2000 – A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition. In *Proc. 10th International Conference on Document Analysis and Recognition*.

Zhang, X., and Zhuang, Y. 2012. Dynamic Time Warping for Chinese Calligraphic Character Matching and Recognizing. *Pattern Recognition Letters* 33 (2012) 2262–2269.

Zheng, G., Li, K., and Yuan, L. 2013. Chinese Characters Recognition Based on HALCON. In Proc. ICCNCE 2013.

Zhuang, L., Bao, T., and Zhu, X. 2004. A Chinese OCR Spelling Check Approach Based on Statistical Language Models. In *Proc IEEE International Conference on Systems, Man and Cybernetics*.

[6] http://ctext.org/instructions/ocr