Large-scale Optical Character Recognition of Pre-modern Chinese Texts

Donald STURGEON

Donald STURGEON is College Fellow in the Department of East Asian Languages and Civilizations at Harvard University. His research interests include language and knowledge in early Chinese thought, and the application of digital methods to the study of pre-modern Chinese language and literature. His current projects include the adaptation of optical character recognition techniques to historical Chinese documents, the application of machine learning to dating and authorship attribution of pre-modern Chinese texts, and the study of text reuse relationships in the pre-modern Chinese corpus. Since 2005, he has developed and managed the *Chinese Text Project* (https://ctext.org), a widely used online digital library of pre-modern Chinese writing.
Email: djs@dsturgeon.net

International Journal of Buddhist Thought & Culture Vol. 28. No.2 (December 2018): 11–44.

https://doi.org/10.16893/IJBTC.2018.12.28.2.11

The day of submission: 2018.10.12
Completion of review: 2018.11.30
Final decision for acceptance: 2018.12.15

Abstract

Optical character recognition (OCR)—the fully automated transcription of text appearing in a digitized image—offers transformative opportunities for the scholarly study of written materials produced prior to the digital age. Digitization, in the sense of photographic reproduction, is now a well-established and efficiently performable mechanical process, and one with significant value in its own right for purposes of preservation as well as access to rare materials. As a result, hundreds of millions of pages of pre-modern Chinese works have already been digitized by libraries and academic institutions around the world—a significant portion of this increasingly being made freely available online.

To make use of this material efficiently, transcriptions of the textual content of these images are needed. Given the enormous volume of image data in existence—and its ongoing production as digitization continues—this task is only feasible if it can be fully automated: performed by software without manual human intervention. Individually, reliable transcriptions produced by OCR offer enormous time savings to researchers, enabling efficient navigation of materials in ways not possible without digital transcription. In aggregate, however, these transcriptions facilitate entirely new ways of exploring historical materials—for example, rapidly identifying material that one suspects might exist somewhere, *without knowing in advance where that might actually be. It is also a prerequisite also to virtually any type of statistical analysis of these materials—the potential utility of which continues to increase as a larger and larger proportion of the extant corpus is transcribed.*

This paper introduces a procedure for OCR of pre-modern Chinese written materials, both printed and handwritten, describing the complete process from digitized image through to automated transcription and manual correction of remaining errors, with particular attention to issues arising in this domain. The process described has been applied to over 25 million pages of pre-modern Chinese works, and the paper also introduces the Chinese Text Project (https://ctext.org) platform used to both make these results available to scholars as well as provide a distributed, crowdsourced mechanism for facilitating manual corrections at scale as well as further analysis of the materials.

Key words: Optical Character Recognition, Chinese, Digitization, Digital Libraries, Crowdsourcing

Introduction

OCR of pre-modern Chinese texts presents challenges distinct from those faced by OCR of modern documents as well as of pre-modern documents in many other languages. Mainstream OCR techniques typically rely in part upon supervised machine learning, in which large amounts of correctly labeled training data—such as images of characters or text lines, each paired with a reference transcription—are used to train a model capable of producing analogous transcriptions from unseen images with similar characteristics to the training data. To obtain optimal results, the training data should be as close as possible in form to that of the documents to be recognized. For modern typeset materials, arbitrary amounts of synthetic (i.e. computer generated) training data can often be created digitally, as the character forms to be recognized correspond closely to those of digital fonts available to the OCR system. For historical materials, such data is typically not available, and a natural approach to improving accuracy is therefore to train using data extracted from real images of text in the same historical writing style. For alphabetic languages with small numbers of character types, the amount of training data required is often small enough that this data can be prepared manually, by identifying clean representative exemplar images of each character type and providing the correct labelings for them. However, the large number of character types involved in Chinese writing—a minimum of 3000 or so for a useful pre-modern OCR system, and ideally many more—usually makes this task impractical to perform by hand.

This paper introduces a practical procedure for obtaining high quality OCR results for pre-modern Chinese texts based on an unsupervised method for the extraction of training data from historical images, together with adaptations to many steps of the OCR processing pipeline. First, as historical documents do not consist solely of clean pages of text, and all non-character marks on a page have the potential to complicate the recognition process and reduce overall accuracy, specialized image preprocessing is used to remove as much non-text content as possible prior to recognition. In the Chinese case, pages frequently contain many additional non-text features having some degree of regularity, including borders and lines between columns of text, and marks introduced during both the transmission of the text itself and the subsequent digitization process, and a variety of approaches are used to identify and

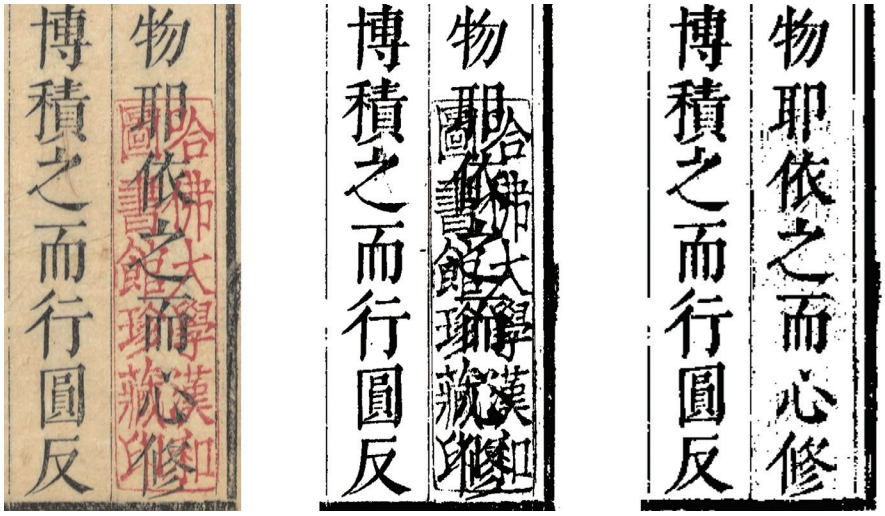
remove these. Next, page images are segmented into individual characters, and OCR models trained on modern Chinese materials are applied to a large corpus of pre-modern page images. The outputs obtained from this procedure are used to align page images with transcriptions of similar texts, such as transcriptions based on other editions of the same work. Similarity metrics are used to automatically identify cases where such alignment has been successfully performed, and successfully aligned pages are then further aligned at the character level, yielding correspondences between image regions and their predicted transcriptions, which can be further processed to automatically produce training data for new models. Finally, the OCR models trained on modern materials are discarded, and replaced with new models trained on the extracted training data for historical character forms. When combined with the image pre-processing steps and additional post-processing, this results in a system capable of significantly more accurate recognition of historical documents than other available methods.

Image Pre-processing

An OCR system is required to take marks made upon a page, represented in the form of a two-dimensional image, and from this data infer a sequence of abstract characters, each represented by a codepoint—a unique number corresponding to one abstract character, irrespective of how it might have been input or formatted. This is a complex process, typically involving many steps, the initial stages of which consist of simplifying the problem to be solved to the greatest extent possible. One of the first questions an OCR system must typically answer is: which things are "marks"—that is, which of the vast number of possible regions of the image are those which could *potentially* belong to a character of some kind? For modern typeset documents, printed on clean, bleached white paper and scanned in optimal conditions, this question may have an easy answer: all those parts of the image that are entirely black constitute potential character components, and all those parts that are white do not.[1] While some OCR systems do attempt to directly make use of more color information than simply these two levels of pure black and pure white, many opt for the simplicity of the 2-level or "binarized" approach used in this paper. In this approach, where an input image is not already binarized (i.e.

contains more distinct shades of grey and/or color than pure black and pure white), the first stage of image pre-processing consists of converting these images into binarized images containing only those two colors, while doing this in such a way as to maximize the chances that all black marks on the resulting images belong to characters, and all white marks do not. This seemingly simple process already introduces many possibilities for erroneous results—and, being the first step in the long and complex OCR procedure while simultaneously affecting every part of the image, has the potential to dramatically change the accuracy of the final result. Most obviously, images of pre-modern works will in general contain a range of shades and hues not present in the simplest case of modern typeset text on bleached white paper. Physical page colors vary significantly, and may be different from background colors corresponding to off-page material; text color, even when expected to be the result of black ink in most cases, can appear in an image as a widely varying range of actual shades, to the extent that a color representing text content in one image (and obviously so, to a human reader) may also occur in another document in which it equally clearly (again, to a human reader) represents part of a page upon which text is written in a darker shade of grey. In some cases, usually due to physical damage to the pages prior to digitization, such as water damage, these two cases can coexist upon the very same page: in some regions one color represents text as contrasted with background, while in other regions the same identical color constitutes background and not text. To deal with this problem, adaptive binarization methods including the Otsu method (Otsu 1979), which take into account local information about relative contrast, are used to avoid losing information due to varying contrast across a page.

Additional problems occur where information is "layered" on the page: where text overlaps a set of marks—sometimes other text—and the two are distinguished visually to a human reader by differences in color. In pre-modern Chinese works, this case occurs relatively frequently in the form of seals stamped upon documents to indicate ownership, typically in red ink. This often has the effect of leaving any text in black clearly readable to a human due to the color differences, but entirely illegible after an otherwise accurate binarization is performed because the shades of red and black used will typically both be dark relative to the paper background (Figure 1). Because these marks are sometimes added systematically by collectors or libraries, in some cases they may be expected to occur many times on every book in a collection.[2]



[Figure 1]
Binarization of an image fragment (left) containing a red seal with (right) and without (middle) color-channel pre-processing.

By taking account of the common convention in Chinese history to use *red* ink of particular ranges of shades for these seals, they can be removed relatively straightforwardly by excluding certain portions of color information from the full-color image prior to binarization, making the black text accessible to later parts of the procedure.[3]

Having determined a reasonable way to binarize the image, there will nevertheless remain in general many marks on the page which do not correspond to characters that should be transcribed; as will be seen below (Results section), such marks represent one common source of error in OCR output. Any such mark has the potential firstly to be misrecognized as one or more characters, which would then need to be deleted manually from the output, but also more damagingly to cause the OCR system to make incorrect decisions about classifying other nearby marks which *do* belong to real characters. In the worst case, such marks can cause a system to make mistaken decisions on page content at a high level, such as treating two adjacent columns of text as a single column, causing entire lines of text to be omitted or mis-transcribed.

In the case of pre-modern Chinese materials, some of the most commonly

occurring cases of such marks are borders drawn around the main page content, and vertical lines drawn between columns; depending upon the digitization and binarization processes used, additional borders towards the edges of the image (extending beyond the limits of the physical page and/or book) may also be present. These marks all have the potential to interfere with subsequent stages of OCR, but also have relatively predictable features: they tend to occur towards the edges of a page, should usually be "outside" the main body of text (i.e. main text should not occur between these marks and the edge of the page image), are likely to be larger than any plausible character to be recognized (i.e. the characters we wish to transcribe will be small relative to the page size, but borders will frequently have extents constituting a significant fraction of page dimensions), etc. While there are complicating factors—these lines are generally hand-drawn or hand-carved, and are often neither perfectly straight, perfectly perpendicular, nor forming continuous uninterrupted strokes—their distinguishing features are nevertheless sufficient to reliably remove a large proportion of such marks from page images prior to performing any further analysis of page contents.

In order to facilitate this process, a useful first step is to deskew each page image, attempting to remove any complete rotation of the page from the vertical (typically, though not always, this means skew applying equally to both the body text and any borders present). Such deviations from the vertical are easily introduced during the original creation of the text (whether handwritten or printed), as well as the scanning process (where the rarity or uniqueness of historical documents may limit the options available for capturing images of the pages). Normalizing page rotation can simplify line removal as well as further steps such as character segmentation and recognition, because it reduces the range of potential inputs which these later stages of the process must be able to correctly handle. After identifying an optimal rotation—a process described in Bloomberg et al (1995)—the image is rotated to align vertical columns of content to be as close to the true vertical as possible.[4] Identification of semi-continuous, near-horizontal and near-vertical lines of length larger than any plausible character components is then performed by analyzing pixel densities within regions of the image, and any such detected lines are replaced with white pixels; additional heuristics are used to erase all content appearing between such lines and the page image boundaries in cases where no text is expected to appear between the line and the image boundary. This results in images containing many fewer black marks not corresponding to character

[Figure 2]
Image pre-processing showing a single scanned page image from the 萬曆版大藏經 (left), together with the same image after binarization (middle) and automated line removal and cleanup (right).

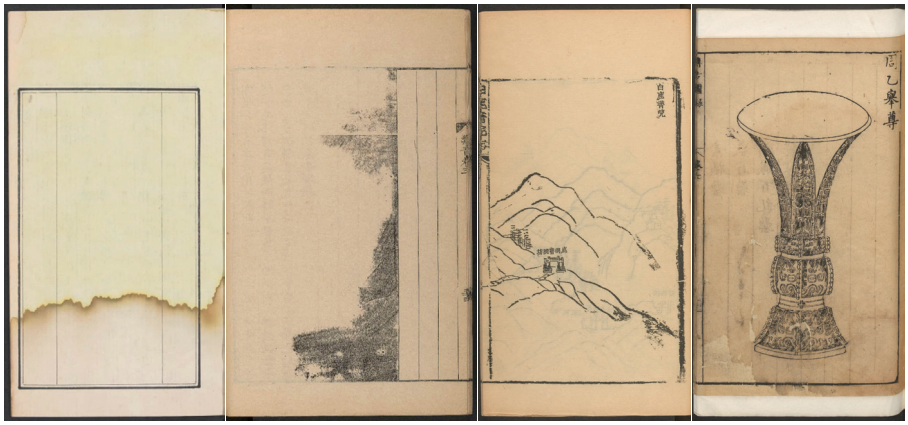
components (Figure 2). An additional benefit of this procedure is that it also excludes from the image marks occurring outside the border corresponding to information located outside the main text flow and containing marginal information such as volume title, page number, etc.—but which in many cases, due to textual production methods, consist only of partial characters which will likely be illegible to OCR (see Figure 2, right-hand edge of page images).[5]

Finally, additional techniques are used to remove more isolated noise from the image. This includes firstly isolated black dots of sizes and distances from other components such that they could not plausibly constitute character components, often resulting from fragments of hand-drawn lines. In block-printed texts, the opposite problem of white noise appearing in black areas also occurs where blocks have been unevenly worn or poorly inked (Figure 3); both can be corrected for small amounts of noise by using information about adjacent pixels. Depending on the algorithms used for character recognition, this step can be particularly useful in character recognition training, as it can help avoid algorithms incorrectly learning features that are due solely to noise in training images rather than general properties of a character form.

A final type of pre-processing is rejection of images (or image regions)

[Figure 3]
A fragment of a block printed text (left) with sub-optimal ink transfer, resulting in small "holes" in the binarized image (middle) which can be corrected automatically (right).

that do not contain any text at all. This task is considerably more technically challenging than those considered so far, because it involves complex global assessments based on image content rather than localized processing. Additionally, while there will be many easy cases such as near-blank pages, there are also harder cases where no *textual* content exists on a page, but other semantically meaningful content (such as an illustration) does appear; yet at the same time the presence of a significant volume of marks on a page does not in itself guarantee the presence of meaningful content (Figure 4). Particularly when dealing with images of real historical objects—many of which are hundreds of years old—marks appearing on what a human reader would consider to be an empty page can often appear in patterns superficially very similar to those of either textual content or illustrations. Failure to exclude processing of such page images in the OCR process will almost inevitably introduce many errors due to some marks being incorrectly identified as text. For small volumes of material this type of error is relatively easy to correct, since all output produced for any given page will be localized in the transcription, and so can be deleted together manually in a single operation. For larger volumes, such as the tens of millions of pages considered in this paper, this becomes more of a concern, as it may not be practical to manually review all pages even for trivial errors such as whether a blank page or illustration has resulted in erroneous output.



[Figure 4]
Two pages containing no semantic content (leftmost two, water damage and printing artefacts respectively), and two pages with illustrations (rightmost two).

One approach to removing this type of noise is to perform post-processing on the OCR output to predict whether or not the output should be considered valid, for example using language modeling (see below) to estimate whether or not the output contains sequences of characters that might plausibly appear in pre-modern Chinese text, or by using character-level confidence values produced by the recognition system to reject pages where many character-level images did not appear to closely resemble characters. In practice, both of these techniques provide poor results due to the difficulty of finding an acceptable cutoff point: when operating at scale, some individual pages will contain only globally improbable sequences of characters, and some pages (due to poor print quality or other technical issues inherent in the images) will contain only characters recognized with low-confidences—both types of pages are all too easily removed incorrectly by post-processing algorithms.

A second approach, used in this study, is to attempt this type of noise rejection *prior* to page recognition, purely on the basis of the page image itself. In this approach, a neural network is trained on the basis of tens of thousands of page images to distinguish between several classes of image: pages containing text, pages containing illustrations, and pages which have no semantic content (i.e. are effectively blank). This task—recognizing the presence or absence of text on a page—is significantly less difficult than the OCR task of recognizing what

text actually appears. It also can plausibly be completed by examining images at a much less fine-grained level of detail: for example, human readers—even those unable to read Chinese—have no difficulty in distinguishing the text-containing images in Figure 2 from the non-text examples in Figure 4, and could perform this task equally well even on reduced-resolution images in which the individual characters of the text were not legible at all. In other words, text occurs on page images not randomly but in consistent patterns which distinguish it from non-text content even at a glance. On the basis of these intuitions, a convolutional neural network was trained using 3,000 manually categorized images of each class. Training data was augmented to increase generalizability by including random transformations (rotation, zoom, shear, etc.) of each training image, after which images were reduced in resolution to a fixed size. The trained model was then used to filter out pages lacking any content, as well as pages containing illustrations, which could then be treated as such later in the transcription process.

Character Segmentation

Given a clean image, ideally now containing nothing except marks belonging to characters, the next task in a traditional OCR approach is to identify which precise regions of the image correspond to distinct characters. The difficulty of this task varies considerably with types of writing style involved, and particularly with the *range* of writing styles that a single segmentation procedure is required to correctly handle (often in the absence of prior information as to which type of writing style will be present in which images). Depending on the style involved, character proportions may be expected to be entirely consistent, consistent within classes (e.g. one proportion for main text, and a different proportion for interlinear commentary), or vary widely depending on the character being written (Figure 5 and Figure 6). Additionally, even where character proportions are highly consistent within a page, the proportion and spacing are generally not known in advance, and can vary significantly between works (Figure 6 and Figure 7). One potentially useful fact about some—but not all—pre-modern Chinese texts is their use of a grid-like layout of characters on the page, in which characters are aligned regularly in both horizontal and vertical axes. Such grid-like layouts evolved over time and eventually became

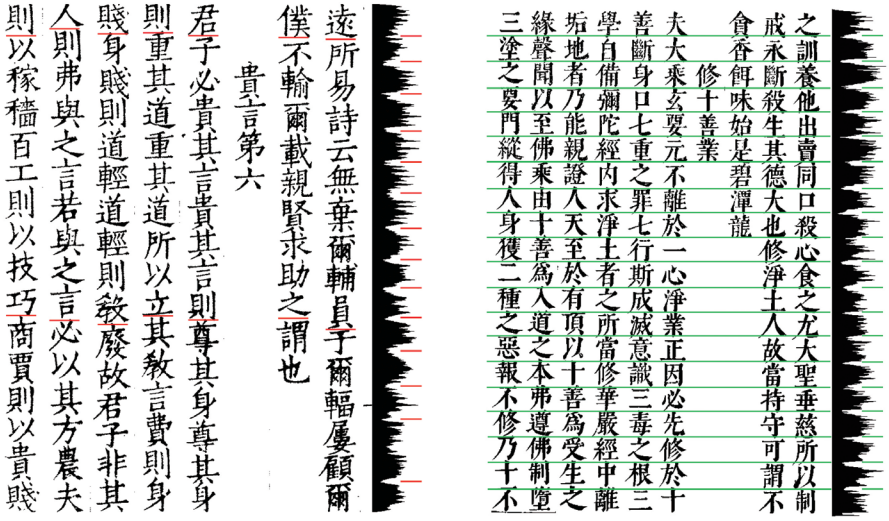
[Figure 5]
Varying character proportions in the same writing style and size from a single page of the 四庫全書. https://ctext.org/library.pl?if=en&res=5575

[Figure 6]
Comparatively regular character proportions and largely fixed character height. https://ctext.org/library.pl?if=en&res=79589

[Figure 7]
Regular character proportions with significantly different aspect ratio to those from Figure 6. The third character from the previous examples is omitted as it does not occur in this volume. https://ctext.org/library.pl?if=en&res=90709

standard for block-printed works (Heijdra 2006, 20–21); where applicable to a given page, this fact can be used to greatly simplify the task of segmenting an image into characters by reducing the range of segmentation possibilities that must be considered—but only when it is known in advance that a particular page will have such a layout.

While one possibility would be to use metadata to determine the presence of a grid-like layout, when working with large sets of materials this data is not always available. Additionally, even where such data exists, there is the further problem that conventions used may differ among sections of a single work—a common example being where a preface is written in an entirely different style to the main text. For this reason, a useful first step is to establish whether or not any given page appears to follow a regular grid-like pattern, and if so, determine what the average character height is. Both goals can be accomplished using horizontal projections of the text region of the image



[Figure 8]
Variable vertical alignment of characters (left) versus regular, grid-like vertical alignment (right). Troughs in a horizontal projection of pixel densities repeat at regular intervals in the regularly aligned case.

(Figure 8): where a grid-like layout is present, local minima in the horizontal projection will correspond to vertical breaks between characters, and these will occur at regular intervals corresponding to a combination of vertical character height and vertical spacing, whereas in general for non-grid-like layouts these minima will not occur with the same degree of regularity.

By contrast, division into columns is less problematic due to the consistent presence of at least some spacing between columns and more rigid conventions observed in horizontal placing of characters within a column. Together these mean that once any vertical lines have been removed from the image, vertical projections of pixel densities can be used to divide the image into separate columns. However, an additional complexity for historical Chinese materials is the frequently used writing convention for including comments within a work, in which smaller characters are used to distinguish other text such as commentary from main body text, and additionally this smaller text is frequently written as a group of sub-columns within the overall columnar layout of the page, complicating text flow. As this convention is not used in

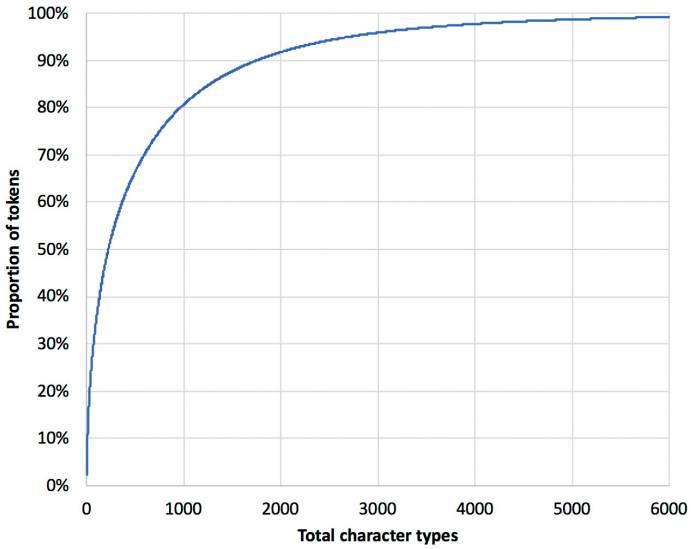
modern Chinese works, mainstream OCR software is typically not able to handle the resulting logical text flow, and so—even where it is capable of performing character-level recognition on such pages—would not be able to produce an accurate, readable page transcription due to incorrect ordering of the characters.[6] This information can be captured using similar projection-based techniques to those used to identify columns of text and characters within columns.

Training Data Extraction

Character recognition is generally achieved using supervised machine learning: a model is trained using large numbers of example images of text, each associated with a "label"—a transcription into characters of the text contained in the image. A training algorithm attempts to extrapolate patterns of how a large range of distinct possible concrete representations of the same text in different images relate to the text itself, encoding this knowledge in a model, which can then be used to transcribe text from future images—these being expected to be similar (though in general not identical) to images used in training. For OCR of modern printed documents, a commonly used and technically straightforward method is automatic generation of synthetic training data by means of preexisting font data. Fonts created for the purpose of displaying and printing text in modern computer systems contain all the information necessary to create character images of any character having a Unicode representation. Multiple fonts can be used to account for expected variation in character styles present in actual documents, and the images created for training can also be augmented with images incorporating subtle transformations—such as random rotation by up to several degrees, enlargement or reduction in size, or the addition of noise—to closely mimic the properties of real-world digitized images to be recognized. The successfulness of this procedure is limited by the degree to which character forms used in documents to be recognized are similar to those used in training; for modern printed documents, which will generally have been produced by computer software using similar types of font data—possibly even one of the same fonts used in training—this technique can train highly accurate models that perform very well on real-world data.

For historical materials, the accuracy achievable using this approach is

limited by differences between handwritten and hand-carved character forms used in pre-modern works and the forms used in modern standard fonts. While creating adequate new fonts modeled upon each historical writing style to be recognized would in principle provide a solution to this problem, in practice this is rarely feasible, and would be prohibitively expensive for languages with large character sets such as Chinese, in which the total number of character types is measured in tens or even hundreds of thousands, and even a baseline OCR system must be able to accurately distinguish several thousand of the most common types in a variety of styles. An alternative is to identify and label images of characters as they occur in historical documents containing the same styles of writing as those to be recognized. This consists of identifying multiple examples of every possible character, creating bounding boxes (i.e. precise regions of the page image containing only the character), and providing the correct label—the correct transcription of that character. For alphabetic languages with small character sets, the total number of symbols to be identified may be relatively small—modern English, for example, would require instances of the letters "a" through "z" in both upper- and lower-case, numbers 0–9, and a variety of common symbols—perhaps less than 100 distinct character types in total, something which could be performed manually in a reasonable amount of time. For historical Chinese, this task becomes greatly more challenging for two reasons: first, the number of characters to be identified is at least an order of magnitude greater than for alphabetic languages; second, because Chinese characters contain a high degree of semantic content (in classical Chinese texts, many words consist of a single character), they function in some respects more similarly to words than characters of alphabetic languages. This second aspect is important because it further increases the difficulty of the task: some characters are only used to write a very limited number of words, and in some cases almost only ever used to express only one word or concept. For example, both the characters in "葡萄 *putao* (grape)" are used almost exclusively in writing that single word (and later, in writing "葡萄牙 *Putaoya* (Portugal)"). Thus, unlike the lesser difficulty in English writing of finding example images of less commonly used letters such as "x" or "z," the appearance of a particular Chinese character can often only be caused by the mention of a particular word or concept—hence making examples of some Chinese characters problematic to find. Yet these characters themselves are not so uncommon as to make it seem reasonable to forgo attempting to recognize them altogether—both



[Figure 9]
Proportion of all character tokens accounted for by top n most common character types, based on 1.2 billion characters of pre-modern Chinese writing.

characters do appear in the top 5000 most frequently used characters.[7] While there is typically a tradeoff in overall expected recognition accuracy versus total number of character types considered—Unicode contains over 80,000 Chinese characters, many of which are obscure to the point that they might be expected to occur once in a million pages of material, and the more characters a system attempts to recognize, the greater its chances of misrecognizing a commonly occurring character as an obscure and improbable one—the most frequently occurring 5000 characters together account for 98.7% of characters in a large and varied corpus of pre-modern Chinese writing (Figure 9)—leaving 1.3% still unaccounted for. Since failure to include a character type in the training step guarantees that it will be misrecognized on every future occasion upon which it occurs, any OCR system aiming to reach low single-digit character error rates on most pre-modern Chinese materials will need to include many relatively rare characters in its training data.

Thus, the difficulty of the problem is considerably greater than that of finding and annotating all the letters of the alphabet in a body of English writing, and arguably closer to that of finding and annotating examples of

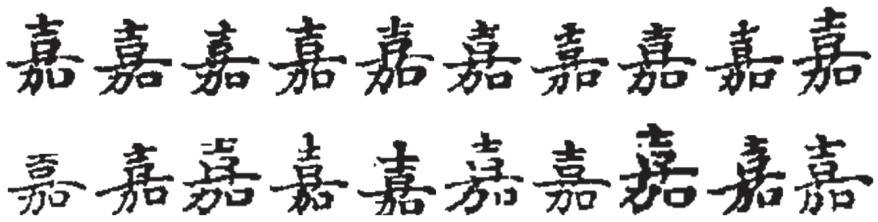
all *words* accounting for 99% of observed English language use. Supposing that characters are randomly distributed on pages in proportion to the rate at which they occur in a large existing corpus, though the probability of finding the most common characters "之" and "不" on any given page is over 90%, the chance of finding the 5000th most common character on a given page is approximately 0.1%.[8] Since *all* characters need to be found—ideally in several locations, and some examples will be incomplete or damaged and have to be thrown out—this suggests that typically tens of thousands of pages of text would need to be examined just to find a single example of each of these most common characters.[9]

In order to address this difficulty, a fully automated method is needed to extract character images and labels without manual intervention. To do this in the first instance, an existing corpus of manually transcribed material was used. This corpus in consisted only of texts and their titles, with no information about which parts of each text occurred on which page, or which edition was used for the transcription. An automated procedure was developed to use this corpus, together with a large set of scanned texts, to reliably infer relationships between image regions and character labels without manual intervention (Sturgeon 2017). An off-the-shelf character recognition engine trained on modern font data was used together with the procedures described above for image processing and character segmentation to produce approximate transcriptions for all available pages of scanned material. Based on a fuzzy comparison of titles of transcribed works and scanned editions, a list of candidate transcriptions and image sequences was created—pairs of transcriptions and scanned texts which *might* represent the same work based solely on similarity of title. The next stage of the process involved making an automated comparison of results obtained with the off-the-shelf recognition method to the expected result—i.e. the corresponding manual transcription of a text with a similar title to the scanned work. Though the off-the-shelf OCR results were not of satisfactory quality to use directly, they were easily of sufficient quality to compute two things: 1) whether or not the transcription and the scan are the same text, and 2) if so, which pages of the scan correspond to which parts of the existing *manual* (i.e. not OCR-derived) transcription. This can be viewed as a sequence alignment task, i.e. the goal is to find the best possible alignment of every character in the OCR data against every character in the manual transcription. Candidate pairs which were low confidence matches

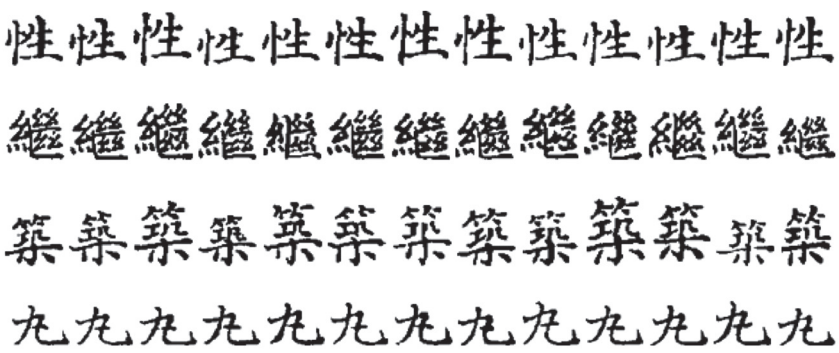
were discarded, leaving only those pairs which showed strong agreement between the expected (manual input) and observed (OCR) results. This resulted in data on hypothesized correct transcriptions of large numbers of pages.

Next, the data within each page was aligned in a similar way to form hypotheses about which characters of text (according to the manual transcription) appeared in which column on a given page. Again, low-confidence pairs in which correspondences were uncertain due to clusters of OCR errors or deviation between different editions were discarded. Lastly, the alignment was repeated a final time to infer which individual *characters* of the manual transcription corresponded to particular characters of the OCR output. Because every OCR-derived character corresponds precisely to some known image region, this correspondence yields a label for an image region believed to correspond to some character. Note that this correspondence must be made on the basis of overall alignment and not simply by selecting cases where the manual transcription agrees with the OCR output for a given character, for this would limit the output to only those characters which the existing OCR model was already capable of recognizing in at least some cases. Instead, using the alignment it is possible to infer from context *which* character a misrecognized character image actually is, thus creating correct character labels even for character images which the existing model was never able to recognize.[10]

By applying this procedure to a set of scanned materials having some consistent writing style, large numbers of labeled character images were obtained. As suggested above, tens of thousands of pages needed to be processed to ensure that images for all character types could be identified. While on the whole these character labels were highly accurate with an error rate of less than 1%, to train an accurate model it is imperative that *all* labels used are accurate—if not, the trained model can be expected to systematically misrecognize the mislabeled characters. Additionally, not all correctly labeled characters are good examples for use in training: images which are of the correct character, but contain significant amounts of noise due to issues in textual production, digitization, or character segmentation are often undesirable as training examples. To address both of these issues, all images for each character type were compared and ranked in order of decreasing similarity to the remainder of character images for that same type (Figure 10). Since most images were correctly labeled, and noise tended to be randomly distributed and



[Figure 10]
Most typical (top) and least typical (bottom) character images extracted for the character "嘉" from the 四庫全書 collection.



[Figure 11]
Examples of character images selected for model training extracted from handwritten text in the 四庫全書.

also absent from the majority of images, selecting a small number of the most *typical* example images within each type avoided the possibility of choosing both incorrectly labeled and damaged characters (Figure 11). Finally, these correctly labeled images were used to train a model to recognize historically attested writing styles. As the entire process of data extraction and character image selection and labeling does not rely upon human intervention, it can be easily repeated for different writing styles to produce directly usable training data for each. The models created at the end of this process can then be used to much more accurately recognize character images written in those styles used to train them. Depending on the type of model used, typically the resultant output consists of a list of the most probable transcriptions of the image,

together with a confidence representing how likely (according to the model) each transcription is to be correct.

Language Modeling

Human readers implicitly make use of context when reading written text, partly as a means to distinguish between visually similar letters. For example, in English, depending on typeface, a lowercase letter "l" (12th letter of the alphabet) and a capital "I" (9th letter) may be visually similar or even indistinguishable. Yet human readers are not usually confused by this fact, because the letters normally appear in sequence in ways that—together with prior knowledge of the written form of the language (as opposed to merely knowledge of the letters with which it is written)—indicate which character should be read in a given case, as in the capitalized word "Illustration." Language modeling provides a means to make similar use of contextual information in performing OCR, by considering not just characters in isolation, but in sequence as they appear on a page. In alphabetic scripts—particularly scripts in which words are delimited in writing by the appearance of spaces between them—dictionary data can be used to post-correct OCR output, by comparing "words" (i.e. space-delimited sequences of characters) observed in the OCR output with expected lists of dictionary words, allowing correction of individual characters within words where it seems probable that the word as a whole is a mistake for an expected dictionary word. In Chinese, this is generally less feasibly due to the lack of explicit word delimiters (and for classical Chinese, would be less useful in any case due to the high proportion of single-character words to which such a technique cannot be applied).

With Chinese materials, instead of using words, a common approach is to use character-based language models, which attempt to model the probability of any given sequence of characters being observed in the type of writing considered. One of the simplest such models is an n-gram model, in which statistics are computed from large bodies of existing transcribed material, expressing the likelihood (on the basis of the existing corpus) that any given sequence of *n* characters will occur in unseen data. For example, in Chinese writing, the 3-gram "天下曰" (all under heaven say) is a relatively common sequence of characters; the visually very similar 3-gram "夭下曰" (under [the character] "yao" it says) is very improbable (though nevertheless can, rarely, occur).[11] Statistics on all sequences of

characters occurring anywhere in a large existing corpus can be easily computed, and then used to make decisions about the likelihood of different character sequences. In this example, where an OCR model predicts a character that looks like it could be either a "天" or a "夭", examining the candidate transcription for adjacent characters can enable the correct decision to be made in the majority of cases. In a different context, a visually similar (or even identical) image might occur; supposing the immediately preceding characters were believed likely to be "桃之", this would greatly increase the chances of the character in question being "夭" and not "天", due to its appearance in an often quoted line of classical poetry, "桃之夭夭" (the flourishing peach [tree]) causing the 3-gram "桃之夭" to be quite common, as compared with the unattested and improbable occurrence of "桃之天" (the sky/heaven of the peach).

Many enhancements of this general approach are possible; in this paper, 3-gram and 2-gram language models created from a 1.2 billion character corpus of classical Chinese were used in recognition. Additionally, for every volume being processed, a 2-gram language model was created on the basis of high-confidence OCR results occurring repeatedly throughout the volume, and then used to correct individual page-level results. This accounts for the intuition that, within any particular work, certain terms and phrases (e.g. proper names) will frequently be repeated many times, and this information can be used to correct borderline instances where there were multiple candidate transcriptions. For example, although the 2-gram "公孫" (Gongsun) is globally much more frequent than "公輸" (Gongshu) within the complete corpus, locally where a text frequently references "公輸" (e.g. as part of repeated references to the person "公輸般"), this increases the probability that other instances that appear to the character model visually similar to both "公孫" and "公輸" are more likely to in fact be the latter. All of these factors are combined by running OCR experiments on a test corpus to determine the weightings which produce the most accurate results.

Results

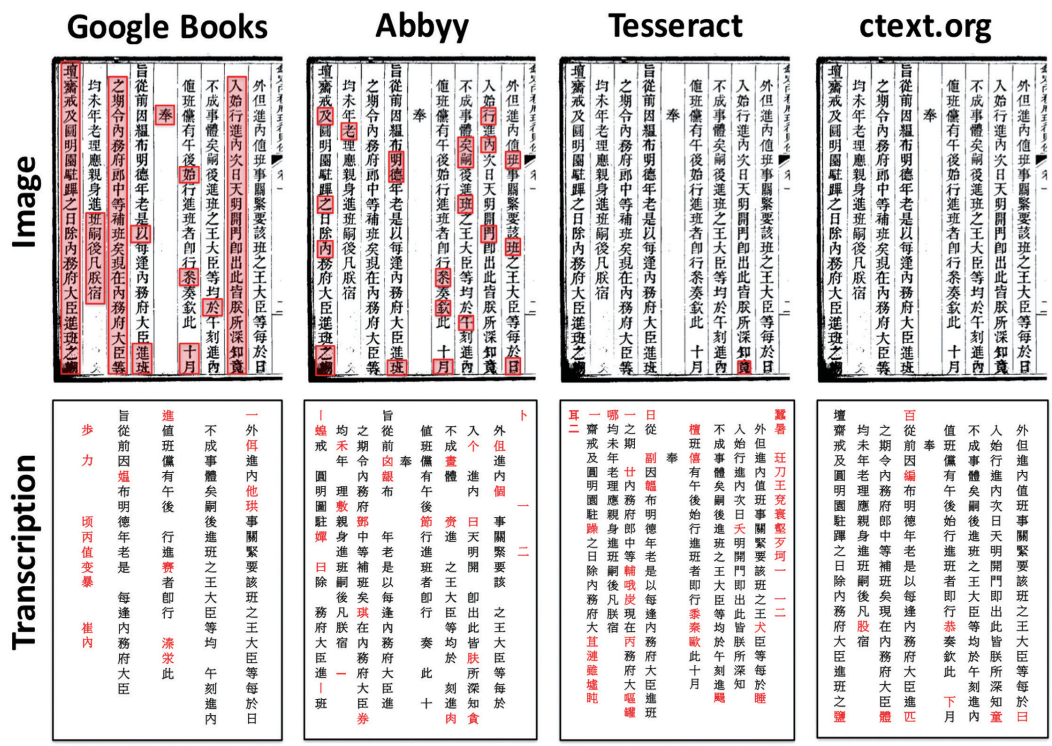
Standard methods of measuring OCR accuracy include character error rates and word error rates: intuitively, these correspond to the proportion of characters or words actually appearing on a page which were transcribed

incorrectly by the OCR procedure. In the case of Chinese text, the lack of explicit word delimiters such as spaces in the written language make calculating word error rate problematic, while at the same time due to the relatively short average length of words in pre-modern Chinese and the high per-character semantic content of Chinese writing, less divergence between word and character error rates can be expected in the Chinese case—hence in this paper, only character error rates will be considered. The character error rate can be defined more precisely as the minimum total number of insertions, deletions, and substitutions which would have to be made to the actual output of an OCR procedure in order to yield a completely correct transcription, divided by the total number of expected characters.

OCR of real-world, historical Chinese documents can be a difficult problem in practice. In order to provide a meaningful comparison, precise character error rates are compared among existing available OCR systems using an identical source image as input in each case. In the following example, an image scanned as part of the Google Books project is used, in order firstly to use a real-world example of a historical image scanned using sub-optimal conditions, and secondly to include the OCR method used by Google Books itself in the comparison.[12] In order to provide a fairer comparison, the output of those methods also capable of outputting non-Chinese marks (English letters, punctuation, Arabic numerals, etc.) was first filtered to remove all such marks—this has the effect of decreasing the error rate for these methods, and is a reasonable post-processing step as it is both easily accomplished and a reasonable assumption that such marks will not occur in most (though not all) pre-modern Chinese works. Performing this procedure, the results for four methods: Google Books, Abbyy FineReader Mac, Tesseract (Smith 2007; Smith et al 2009), and the ctext.org procedure introduced in this paper, are summarized in Table 1. The same results are presented visually in Figure 12, in which the images on the top row highlight on the input image those parts which were incorrectly not transcribed by the method (i.e. the insertions which would be required to correct the output), and the images in the lower row contain the OCR output from each method, with transcription errors highlighted (a combination of substitutions and deletions). From these results it can be seen that in this case, the character error rate varies significantly across methods, with the ctext.org method representing a ten-fold reduction in error rate over Google Books OCR, from 61% to 6%, and a 4- to 5-fold reduction in

[**Table 1**] Character error rates for the page image shown in Figure 12.

Method	Substitutions	Insertions	Deletions	Total errors	Error rate
Google Books	7	76	11	94	61%
Abbyy FineReader	20	22	4	46	30%
Tesseract	24	1	17	42	27%
ctext.org	10	0	0	10	6%



[**Figure 12**]
Visualization of character error rates for a single page from Google Books using four different OCR systems.[13]

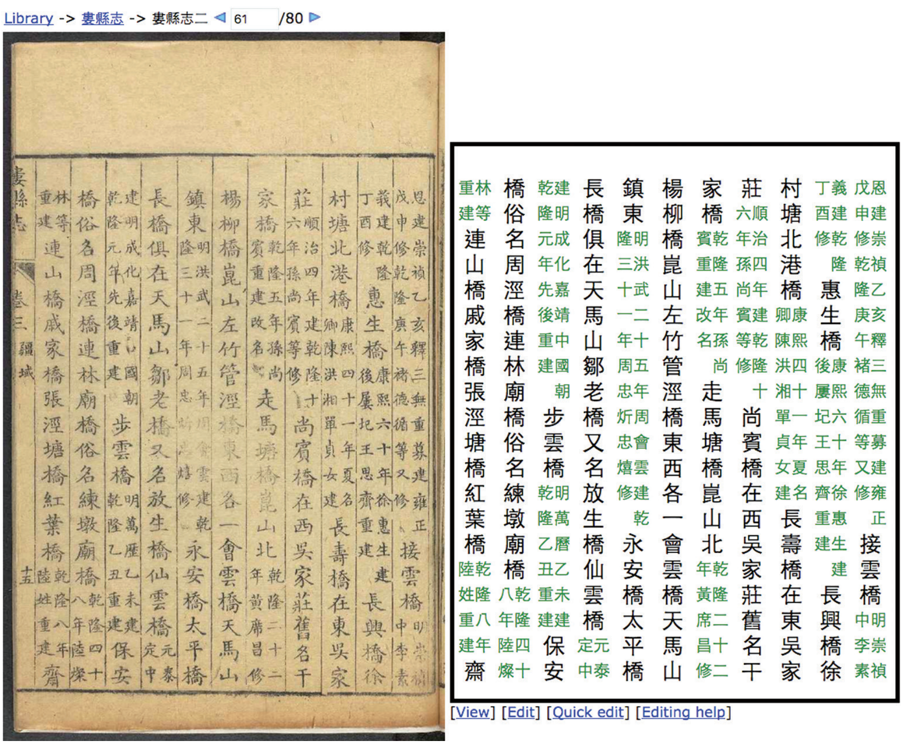
error rate over off-the-shelf commercial (Abbyy FineReader) and open source (Tesseract) general purpose Chinese OCR software. Interestingly, the patterns of error observed in each of the cases is quite different, with Google Books omitting large amounts of material,[14] Abbyy omitting a much smaller amount, and Tesseract, despite having an overall comparable error rate to Abbyy, omitting only a single character, instead incurring penalties due to noise most likely caused by non-text content occurring towards the edges of the body text. In the ctext.org result, all errors are substitution errors, and half of these occur in the bottom-most position in each column; closer inspection of the image indicates that these errors are likely due in part to the presence of noise in this region of the digitized image.

Using the Data

The textual transcriptions of image data created using OCR can be directly useful to researchers, since in many cases historical materials have not previously been transcribed digitally at all. One of the most basic—and easily achieved—applications of OCR of historical materials is to enable full-text search of the digitized image data. This is a particularly attractive application for two reasons: first, most pre-modern Chinese materials lack modern navigational aids such as contents pages with page numbering; second, distinct historical editions—even when highly similar in content—are often non-identical in places, and these variations are often of interest to scholars; lastly, and most importantly, in scholarly work direct citation of primary source materials is very often desirable or essential, and being able to locate precisely the source of textual material within a work makes this process vastly more efficient.

In order to facilitate full-text search, all materials processed with the OCR procedure described in this paper are ingested into an open, online digital library system called the Chinese Text Project (https://ctext.org), which provides access to the page images and transcriptions, together with full-text search functionality. Transcriptions of entire works are made by joining page transcriptions in order; a combination of available metadata (such as *juan* 卷 numberings and their start locations in the page sequences, where this information is available) and rules designed to infer structure from the page-level transcription

features (e.g. use of vertical offsetting to indicate titles, and the presence of blank or unusually short pages at the end of chapters or *juan*) are applied to make the transcriptions more readable and easier to navigate. This results in two distinct natural ways to view the transcribed text: a "physical" view, in which the transcription is viewed page by page in direct correspondence to the original edition, and a "logical" view, where text is divided into units such as chapters and paragraphs, but otherwise presented as a single flow irrespective of the divisions between pages as they occur in the transcribed edition. Both of these views are useful to the researcher: the physical view allows efficient comparison between the OCR-derived transcription and the original images, but the logical view provides a more easily readable and searchable version of the content (Figure 13 and Figure 14). Since the same data lies behind both views, the two views can be linked to one another to facilitate switching between



[Figure 13]
Image and transcription (physical) view of a page showing an OCR-derived transcription subsequently corrected by crowdsourcing.

23 四十三保，橋二十有六。

24 谷源永壽橋，跨張家浜，俗名中橋明天啟丙寅釋三無等改建，雍正乙卯僧省機募修，舊有二橋，今圮其一；會源橋，俗名陸家浜橋明嘉靖戊午御史馮恩建，崇禎乙亥釋三無重募建，雍正戊申修，乾隆庚午褚德循等又修；接雲橋明崇禎中，李素義建，乾隆丁酉修；惠生橋康熙六十年徐惠生建，後屢圮，王思齊重建；長興橋，徐村塘北，港橋康熙四十一年夏名卿、陳洪湘、單貞女建；長壽橋，在東吳家莊順治四年建，乾隆十六年孫尚賓等修；尚賓橋，在西吳家莊，舊名干家橋乾隆五年孫尚賓重建改名；走馬塘橋，崑山北乾隆二十二年黃席昌修；楊柳橋，崑山左；竹管涇橋，東西各一；會雲橋，天馬山鎮東明洪武二十五年周會雲建，乾隆三十一年周忠炘、忠燻修；永安橋、太平橋、長橋，俱在天馬山；鄒老橋，又名放生橋；仙雲橋元泰定中建，明成化、嘉靖中、國朝乾隆元年先後重建；步雲橋明萬曆乙未建，乾隆乙丑重建；保安橋，俗名周涇橋；連林廟橋，俗名練墩廟橋乾隆四十八年陸燦林等重建；連山橋；戚家橋；張涇塘橋；紅葉橋乾隆八年陸姓重建；齋口橋，在秀溪上康熙十年周岳鎖建；峴村橋，在峴溪上周忠燻建。

[Figure 14]
Textual (logical) view of the same text as shown in Figure 13 (corresponding content manually highlighted in blue).

logical and physical representations at any point. The full text generated in the logical view can also be exported via Application Programming Interface, allowing for its use in other tasks such as text mining or annotation.[15]

Crowdsourcing of Corrections

Despite the significant improvements in accuracy possible using techniques such as those described in this paper, OCR of historical materials inevitably results in errors of transcription even under close to ideal conditions. While ideally researchers would like access to manually reviewed and corrected materials containing error rates as close to zero as possible, the sheer volume of data available—tens of millions of pages in this project, and many more outside its scope—makes this option impractical due to the formidable investment that would be required to perform this on the full dataset. An alternative to the traditional approach of proofreading and correcting data prior to making it available is a crowdsourcing model, as popularized by websites such as Wikipedia, and used in a range of transcription projects including Distributed Proofreaders (Newby and Franks 2003) and Transcribe Bentham (Moyle et al, 2011; Causer and Terras 2014). In a crowdsourcing approach, instead of a centrally

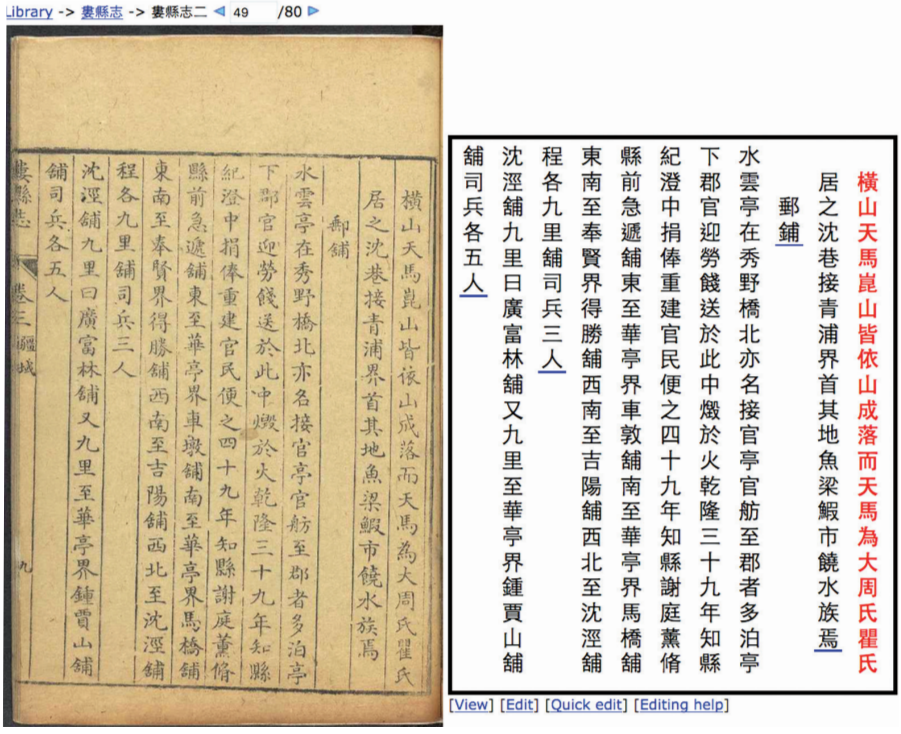
coordinated group of editors correcting the material, corrections are made by a distributed, uncoordinated group of editors, each working independently. Often these editors are volunteers, who may be a self-selecting group not directly affiliated with the project itself other than through their voluntary participation.

In the Chinese Text Project, texts transcribed using OCR are placed in a wiki system analogous to (but technically distinct from) the MediaWiki software used by projects such as Wikipedia and Wikisource. The key goal of this system is to facilitate storage of textual transcriptions in a manner that simultaneously allows for: 1) physical page-wise viewing and searching of a text (Figure 13); 2) logical textual viewing and searching (Figure 14); and 3) version-controlled editing of both the textual content and logical structure enabling those two functions. To achieve this, a single representation of each text is created during the OCR process, containing the textual content, but also additional markup—predefined codes indicating the relationship between every character of the transcription and its location on a particular page of a scanned edition. The crowdsourcing system allows for corrections to be made in two ways: by directly manipulating the complete internal representation, which includes both transcription and markup expressing the relationship between text and its appearance on physical pages; and by editing the transcription for a single page in a simplified interface in which this markup is hidden from the user (but still retained by the system during editing).

The system is entirely open, and is edited by a largely self-selecting group of pseudonymous volunteer users: anyone can create an account through the online interface and immediately contribute to correcting transcriptions. Edits made by any user through the online interface immediately become part of the current version of the text; however, all previous versions are also preserved and can be easily restored in future if necessary, as well as visually compared to identify precisely what changes have been made by which users. Edit logs—lists of recent changes, either globally or for any part of a particular work—are available to all users, so that others can verify the correctness of newly submitted edits as well as the reliability of new users who may not be familiar with editing rules and conventions. Entries in the edit logs can be visualized to rapidly see what precisely has changed during an editing operation: the parts of a text which have been changed are listed in context, in their state prior to the edit in question on the left, and in their state after the edit on the right, with deletions and substitutions highlighted

Original	New
<scanend file="117519" page="48" /><scanbegin file="117519" page="49" y="1" />橫山天馬崑山皆依山戒落而天馬為大同氏瞿氏<scanbreak file="117519" y="1" />居之沈巷接青浦界首其地魚梁鰕市饒水族焉<scanbreak file="117519" y="2" />郵鋪	<scanend file="117519" page="48" /><scanbegin file="117519" page="49" y="1" />橫山、天馬、崑山皆依山成落，而天馬為大，周氏、瞿氏<scanbreak file="117519" y="1" />居之。沈巷接青浦界首，其地魚梁鰕市饒水族焉。<scanbreak file="117519" y="2" />***郵鋪

[Figure 15]
Visualization of a part of a single edit operation from the edit log in the online interface showing three corrected characters as well as added punctuation and logical markup (indicated by asterisks).



[Figure 16]
The corresponding image and transcription view shown when part of the edit log shown in Figure 15 is clicked on enabling subsequent review of the edit.

in the prior version, and additions and substitutions likewise highlighted in the edited version (Figure 15). This visualization itself links to the search interface of the physical view, meaning that any part of the modified version can be immediately visually verified by comparison to the scanned image of

the particular page in question, with the correct region highlighted (Figure 16). Since this system was publicly released in 2015, over 80,000 edits have been made by thousands of users, with the system currently averaging over 100 edits each day. Moreover, as one "edit" in the sense used here corresponds to one submission of a revised version of a text or page—and in practice, many editors choose to correct an entire page in a single operation—the total number of *corrections* made is considerably greater.

Conclusions and Future Work

OCR of historical documents is a challenging task due to a variety of factors, most fundamentally the wide range of sources involved. Often there is considerable variation within even those ideal cases that do not suffer from printing artefacts and physical damage. Inevitably some level of human intervention will be needed for the foreseeable future to produce accurate transcriptions of these materials, but at the same time, improvements in OCR technology are rapidly being made, and these can be expected to further reduce error rates. Many of the processes described in this paper relate to domain adaptation of existing OCR techniques to one particular case—historical Chinese works—and as such, many of the procedures described here will remain valuable when applied together with newer OCR approaches in the future. At the same time, much recent work on OCR involves the application of neural network techniques, which have the potential to bring not only significant improvements in accuracy, but also simplification of a long and complex procedure by integrating steps traditionally carried out using a mixture of different techniques into a single procedure which learns to make effective decisions on its own purely on the basis of training data. For example, neural network approaches can perform the complex process of character segmentation—and, in principle, establishing the presence of and correct reading order for interlinear commentary—as part of a single learning process which simultaneously tackles character recognition (Breuel et al 2013); other parts of traditional OCR pipelines such as binarization can also benefit from being built into a unified training process in a similar way (Yousefi et al 2015). However, some challenges raised in this paper which are largely specific to pre-modern Chinese OCR, such as the unusually large character set, may have

an impact on the practicality of some such approaches. One possibility is to make better use of an important fact about Chinese characters: many of them (i.e. a majority of character types) are composites, composed of subcomponents themselves appearing in similar or identical form within many different characters, and/or being independent characters occurring in their own right. While neural network approaches to some extent implicitly model this effect, more targeted approaches may be able to help with the problem of accurately recognizing rare and obscure characters.

The corpus created through the work described in this paper—in addition to its utility for its intended purpose of facilitating scholarly access to historical materials—also represents a significant body of training data with which to explore the potential of new and emerging approaches to historical Chinese OCR. In order to evaluate and compare the accuracy of future methods, a balanced corpus of test data is needed, incorporating material in a variety of writing styles as attested in the historical corpus—though developing this will require human involvement, the processes described here will greatly reduce the effort needed to create it. As a large volume of corrections continue to be made by the crowdsourcing community, these edits themselves also provide valuable feedback which could be harnessed to further improve OCR results and identify systematic biases in OCR output. Future work will also likely involve applying similar domain adaptation techniques to historical works from regions with related cultural and textual traditions to the Chinese case—while some different issues arise, OCR of historical Japanese, Korean, and Vietnamese works also face many challenges overlapping with those presented here.

Notes

1 In reality, even in this simplest type of case, shades of grey lying between black and white will be present at some part of the process, and how these are dealt with can significantly affect the accuracy of the procedure.

2 This is the case with the image shown in Figure 1, which is from (and states ownership of) the Harvard-Yenching Library—this seal is present at the start of every volume of most books in that library's rare books collection.

3 It is equally possible to use this process to isolate the seals themselves, either to recognize text in them, or mark their presence in the resulting transcription, though this is not presently performed by the procedure described.

4 Vertical alignment is privileged over horizontal alignment, since incorrect *vertical* alignment has the potential to cause greater difficulty in segmenting text into columns.

5 As with the case of color information, this data could be processed separately and the results included in an appropriate way if desired—though the problem of fragmentary characters remains.

6 This is one reason why OCR for the purposes of creating adequate transcriptions, as considered in this paper, is inherently more difficult than OCR intended solely for the purpose of implementing text search in image sequences.

7 This assessment is based upon rankings of the most frequently occurring characters in all distinct manually transcribed or corrected works in the Chinese Text Project.

8 The assumption that characters appear randomly of course does not hold in reality, and character distributions will not be random in practice—though this does not necessarily make the problem any easier. These calculations are based on an assumption of 200 characters appearing on each page.

9 An alternative would be to use a premodern dictionary as the source for manual annotation; this would reduce the difficulty, but also limit the range of writing styles available for training to those represented in such works, and still present the same problems if any of the headwords should be damaged and unsuitable for training.

10 For a detailed explanation of this procedure, see Sturgeon (2017).

11 The former occurs many thousands of times in the Chinese Text Project corpus; the latter occurs just once, in a commentary on the 說文解字 *Shuowen Jiezi* dictionary.

12 The OCR used in Google Books can only be applied to new images by Google itself; however, its transcriptions for pages included in Google Books are accessible through the Google Books interface.

13 Image source: https://books.google.com/books?id=zXwsAAAAYAAJ . The image shown is from page 105 of the downloadable PDF. Note that the images in the PDF are not identical to those in the online Google Books viewer, though they likely derive from a common source.

14 Much of this omitted content is actually transcribed as sequences of non-Chinese

punctuation, which have been removed for comparison—as noted earlier, this process can only decrease the error rate for a method and cannot increase it as no punctuation is expected to appear.

15 https://ctext.org/tools/api

References

Bloomberg, Dan S., Kopec, Gary E. and Lakshmi Dasari 1995 — "Measuring Document Image Skew and Orientation." *SPIE Conference: Document Recognition II,* 302–316.

Breuel, Thomas M., Ul-Hasan, Adnan, Azawi, Mayce Al and Faisal Shafait 2013 — "High-Performance OCR for Printed English and Fraktur Using LSTM Networks." *12th International Conference on Document Analysis and Recognition (ICDAR 2013).*

Causer, Tim and Melissa Terras 2014 — "Many Hands Make Light Work. Many Hands Together Make Merry Work." In *Crowdsourcing our Cultural Heritage*, ed. Mia Ridge. Farnham: Ashgate Publishing.

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

Moyle, Martin, Tonra, Justin and Valerie Wallace 2011 — "Manuscript Transcription by Crowdsourcing: Transcribe Bentham." *Liber Quarterly* 20 (3): 347–356.

Newby, G. B. and C. Franks 2003 — "Distributed Proofreading." *Proc. Joint Conference on Digital Libraries* 2003.

Otsu, Nobuyuki 1979 — "A Threshold Selection Method from Gray–level Histograms." *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1): 62–66.

Smith, Ray 2007 — "An Overview of the Tesseract OCR Engine." *9th International Conference on Document Analysis and Recognition (ICDAR 2007).*

Smith, Ray, Antonova, Daria and Dar-Shyang Lee 2009 — "Adapting the Tesseract Open Source OCR Engine for Multilingual OCR." *MOCR '09: Proceedings of the International Workshop on Multilingual OCR.*

Sturgeon, Donald 2017 "Unsupervised Extraction of Training Data for pre-Modern Chinese OCR." *Proc. FLAIRS-30.*

Yousefi, Mohammad Reza, Soheili, Mohammad Reza, Breuel, Thomas M. Kabir, Ehsanollah and Didier Stricker 2015 "Binarization-free OCR for Historical Documents Using LSTM Networks." *13th International Conference on Document Analysis and Recognition (ICDAR 2015).*