Flexible Computer Assisted Transcription of Historical Documents Through Subword Spotting

Brian Davis, Robert Clawson and William Barrett
Department of Computer Science, Brigham Young University
Provo, Utah
Email: briandavis@byu.net

Accurate transcription of important historical documents is a costly process, requiring many man-hours. Current handwriting recognition technology does not allow fully automated solutions; during a recent competition for handwriting recognition on historical documents, the top method had a word error rate above 25% [1]. However, computer assisted transcription (CAT) methods offer a middle ground between manual and fully automated transcription. CAT methods aim to harness handwriting recognition technology and human efforts together in an effective way. We will explore a few prior CAT systems to examine the state of the art methods. Additionally, crowdsourcing has been a popular means of transcribing large corpi of handwritten documents (e.g. FamilySearch Indexing). We will propose a CAT system which is directed at crowdsourced work, and is particularly adaptable to mobile users.

Toselli, et al [2][1] have explored the realm of CAT using the idea of user-verified prefixes. They use a fairly standard HMM recognition model as the backbone of their approach. The recognition is done for a line of text and the user corrects the first error. Recognition is run again reusing the computation up to that point and the correction. Their approach relies on a language model, which means this approach cannot be used to effectively transcribe documents containing non-sentence writing, such as tables and lists. Serrano, et al also have pursued a similar approach, where the user corrects the n words the recognition model had the least confidence in [3].

Robert Clawson designed Intelligent Indexing [4][2], a CAT system for handwritten documents. Intelligent Indexing relies on finding matching word images in a document and assigning them the same user-specified label. The user oversight of matches was accomplished by showing the user a list of matches (with an adjustable threshold for sensitivity) from which the user removed the false-positive matches. This leveraged the human user's natural ability to discriminate. Zagoris et al [5][3] also propose a CAT system which uses word spotting. Rather than focusing on having a user remove bad spots, as the user confirms correct spottings, a relevance feedback loop helps select better results from the word spotting. Both of these approaches allow a few user actions to transcribe many words. However, both of these approaches are limited as they require frequent word repetition to be effective. There are some commonly repeating words for certain documents, but there are many words which repeat infrequently, or not at all, in documents.

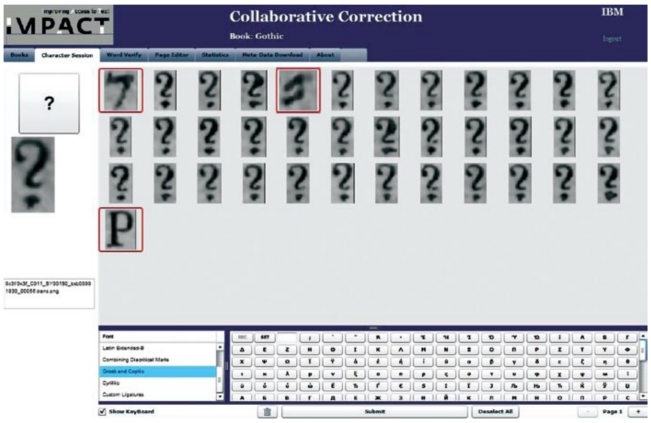

Fig. 1: Screen shot of character session for "?" from Neudecker and Tzadok's CAT system, taken directly from their report [6]. Both this method and Intelligent Indexing use an interface that makes it easy for users to simply click on erroneous classifications.

Neudecker and Tzadok [6] presented a CAT system for historical printed documents which is very similar to the CAT system we are presenting here. Their system first segments the individual characters of the documents and runs an OCR engine on them. Those characters with low confidence are then presented to a user for verification in a character session. A single character session contains all the low-confidence character images classified to a single character; the user merely needs to select the incorrect classifications. An example of their system's character session for the character "?" is given in Fig. 1. Then in a word session, a word image is shown to the user with possible transcriptions for the word, from which they select the correct one.

There are three key strengths of the system presented in [6]. One is that as long as a documents' characters can be segmented, it can work with that document. The second is that it formats all user tasks as selections, rather than typing, thereby minimizing the time to complete each task and reducing human errors from typing. This also creates a much

[1] You can find a demo of their system at http://cat.prhlt.upv.es/iht/
[2] You can view a short demo of his system at http://tiny.cc/intelind
[3] You can find a demo of their system at http://vc.ee.duth.gr/ws/

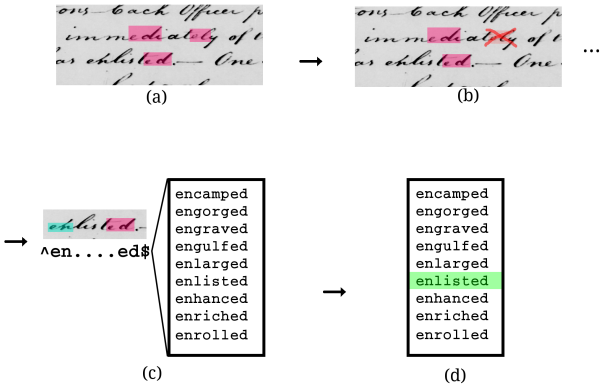

Fig. 2: Work-flow of proposed CAT system. (a) N-grams are spotted by the system ("ed" as an example here). (b) User removes false-positives ("el"). (c) After some iterations of n-gram spotting, a regular expression is generated (from "en" and "ed") and used to query the lexicon. (d) If 10 or less words are returned, present the list to a user to select the correct transcription ("enlisted").

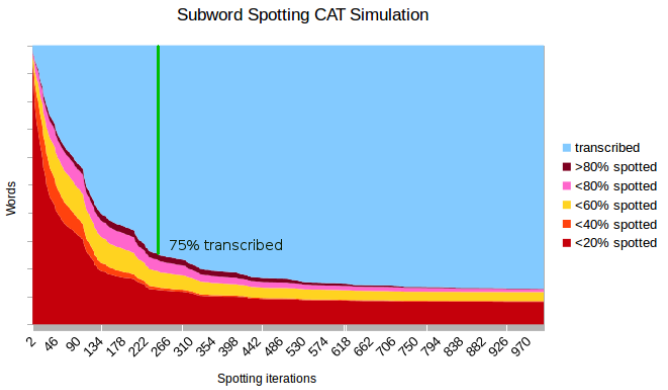

Fig. 3: Results of a simulation showing how much of the GW dataset can be transcribed after a given number of spotting iterations. The chart is drawn so one can observe the progress of spotting as well as transcription, each category (color) indicating the portion of words which have the given percent of their characters recognized by spottings (or indicating the portion of words transcribed).

more enjoyable experience for the user and could be easily adapted to a small touch screen. The third key strength is that it is highly parallelizable for crowd-sourced transcribing. This parallelism is achieved as all character sessions are independent of one another and all word sessions are independent of one another. Our CAT system for handwritten documents will follow this system's flexibility for document types, simple user tasks and parallelizable framework.

Clawson [4] and Zagoris et al [5] rely on word spotting to transcribe, which is dependent on frequent word repetition. Neudecker, Tzadok [6] relies on OCR to transcribe, which is dependent on character segmentation, a difficult problem for handwriting. A happy median, to word spotting and OCR, is character n-gram spotting, which is spotting short subwords (bi- and trigrams) in the words of the document. N-grams have more frequent repetition than words do, but are large enough to spot (i.e. it doesn't require character segmentation). This will provide the backbone of the CAT system we are proposing.

Our system would follow a similar pattern as [6]. N-grams are spotted in the document images. Low confidence spottings are then presented to users to verify (discarding/ignoring incorrect spottings). From the spotted n-grams we are able to generate partial transcriptions of words (we know some, but not all of the letters), from which we can narrow the list of possible transcriptions considerably. Once this list has been narrowed down to a few words (from one or more n-grams being spotted in the word image), this list is presented to a user to select the correct transcription. Fig. 2 shows this process. Additionally, spotted n-grams and transcribed word images provide information we can learn from to improve later spotting iterations.

Let us examine the George Washington (GW) dataset [7] as an example of how effective this might be. If the 100 most frequent bigrams in the English language are spotted in the GW dataset with 49% recall (i.e. we actually spot only 49% of the occurrences of each bigram), 53.6% of the words in the corpus can be narrowed down to 10 or fewer possible transcriptions. This is assuming we can correctly guess the correct number of characters we haven't spotted in the words and using a 32,000 word lexicon. From this list of 10 or fewer words a user can easily select the correct transcription. More words can be transcribed as we use online learning to create new spotting queries; character n-grams that we missed will be spotted with subsequent queries. If subsequent queries also have 49% recall, 75% of the corpus can be transcribed with 250 spottings (i.e. going through the 100 bigrams 2.5 times). See Fig. 3 for the results of simulating this process. Preliminary results in spotting bigrams in the GW dataset have yielded a mean-average-precision greater than 60%.

Like [6], the proposed system is flexible to all handwritten documents, is highly parallelizable, and has very simple user tasks. Of interest, small user tasks are ideal for casual mobile users. This can attract a larger audience of users, and this, combined with the parallelizable nature of the system, will make it ideal for crowdsourced transcription of historical handwritten documents.

References

[1] J. A. Sánchez, A. H. Toselli, V. Romero, and E. Vidal, "ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset," in *Proc. ICDAR*, 2015. [Online]. Available: icdar.org/proceedings

[2] A. Toselli, V. Romero, M. Pastor, , and E. Vidal, "Multimodal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1814–1825, 2010.

[3] N. Serrano, A. Giménez, J. Civera, A. Sanchis, and A. Juan, "Interactive handwriting recognition with limited user effort," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 1, pp. 47–59, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10032-013-0204-5

[4] R. Clawson, "Intelligent indexing: A semi-automated, trainable system for field labeling," Master's thesis, Brigham Young University, 2014. [Online]. Available: scholarsarchive.byu.edu/etd/5307/
[5] K. Zagoris, I. Pratikakis, and B. Gatos, "A framework for efficient transcription of historical documents using keyword spotting," in *Proc. HIP*. ACM, 2015.
[6] C. Neudecker and A. Tzadok, "User collaboration for improving access to historical texts," *Liber Quarterly*, vol. 20, no. 1, p. 119128, 2010.
[7] V. Lavrenko, T. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proc. DIAL*. IEEE, 2004, pp. 278–287.