Flexible Computer Assisted Transcription of Historical Documents Through Subword Spotting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

***Abstract*—In the absence of accurate handwriting recognition for historical documents, computer assisted transcription (CAT) methods move into the spotlight. We explore some of the weaknesses of current CAT systems and propose a CAT system which relies on subword spotting that overcomes most of these. The system is ideal crowdsourcing transcription to mobile users.**

I. Introduction

Manual transcription of handwritten historical documents is a costly process, requiring many man-hours. Current handwriting recognition technology is an inaccurate replacement; during a recent competition for handwriting recognition on historical documents, the top method had a word error rate above 25% [1].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 an effective means of transcribing large corpi of handwritten documents (e.g. FamilySearch Indexing). We propose a CAT system which is directed at crowdsourced work, and is particularly adaptable to mobile users. The ubiquity of smartphones means mobile users represent a large resource for crowdsourced work, but these potential users generally limited with small screens, no keyboards (limited data entry), and shorter attention spans. Our system addresses these limitations.

II. Prior Work

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 (See Fig. 1). 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].

[1]You can find a demo of their system at http://cat.prhlt.upv.es/iht/

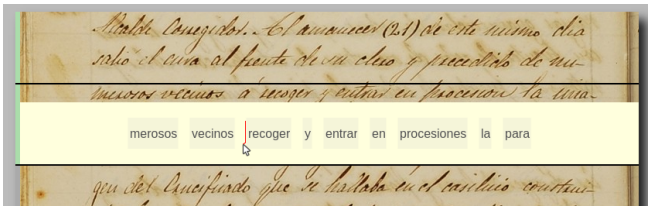

Fig. 1: A screenshot of a demo of Toselli et al's multimodal CAT system. The red line is drawn by the user to indicate the need to insert a word into the automatically obtained transcription.

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, as seen in Fig. 2. This leveraged the human user's natural ability to discriminate. Zagoris et al [5][3] also designed a CAT system which uses word spotting, as seen in Fig. 3. 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, if at all, in documents (e.g. names).

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. 4. Then in a word session, a word image is shown to

[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/

(a) A small window shows matching words from the column. The user can get rid of bad matches (e.g. "Wife") by clicking on them.

(b) The matched words are given the same label, indicated by the highlighting, and the red box is advanced to the next word to be transcribed.

Fig. 2: An example of Clawson's Intelligent Indexing, a CAT system for tabular documents.

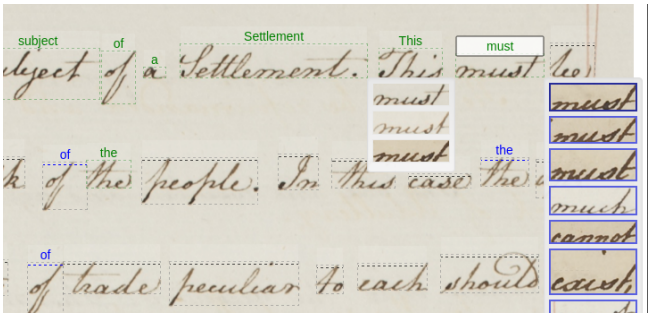

Fig. 3: A screenshot of a demo of Zagoris et al's word spotting based CAT system. The matched words of previously transcribed words have also been given the same label). The current word's matching results are shown as the long list on the right and the user confirmed spottings the short list to its left.

the user with possible transcriptions for the word, from which the user selects the correct one.

There are three key strengths of the system presented in [6]. One is that there is no dependence on a language model (unlike [2] and [3]), as long as a documents' characters can be segmented, it can transcribe the 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 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. The CAT system for handwritten documents we propose follows this system's flexibility for document types, simple user tasks and parallelizable framework.

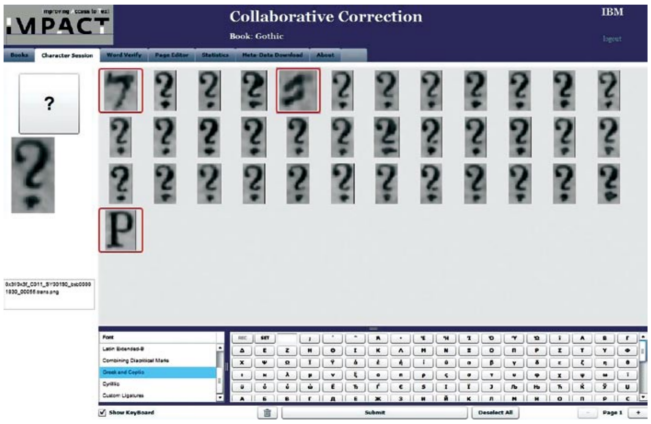

Fig. 4: Screen shot of character session for "?" from Neudecker and Tzadok's CAT system, taken directly from their report [6]. Both this method and Intelligent Indexing use an interface that makes it easy for users to simply click on erroneous classifications.

III. Proposed System

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 medium, to word spotting and OCR, is character n-gram spotting, which is spotting short subwords (bi- and trigrams) within the words of the document. Character n-grams have more frequent repetition than words do, but are large enough to spot (i.e. don't require character segmentation), meaning there should be a relatively large information gain for the work of spotting a single n-gram. This provides the backbone of the CAT system we propose.

Our system follows a similar pattern as [6]; Fig. 5 shows an overview of the process. N-grams are spotted in the document images. Low confidence spottings are then presented to users to indicate incorrect spottings (Fig. 6 might be how this would be presented to a user). From the spotted n-grams partial transcriptions of words (we know some, but not all of the letters) are found, from which the list of possible transcriptions can be narrowed considerably. Additional reduction can be done by scoring the possible transcriptions on the word image with an ordinary word spotting or handwriting recognition algorithm, and thresholding the scores. Once this list has been narrowed down to a few words, this list is presented to users to select the correct transcription (Fig. 7 might be how this would be presented to a user). Additionally, spotted n-grams and transcribed word images provide information the system can learn from to improve later spotting iterations.

IV. Justification

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 50% recall (i.e. we actually spot only

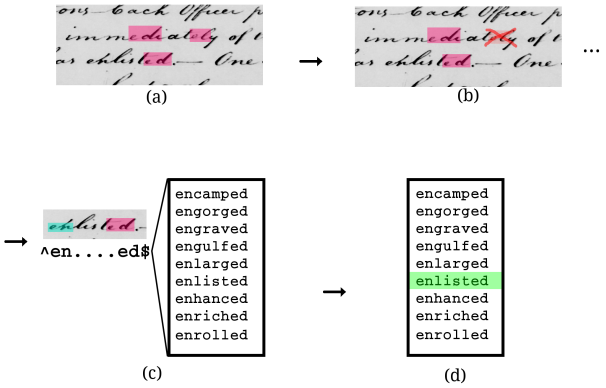

Fig. 5: Work-flow of proposed CAT system. (a) N-grams are spotted by the system ("ed" as an example here). (b) User removes false-positives ("el"). (c) After some iterations of n-gram spotting, a regular expression is generated (from "en" and "ed") and used to query the lexicon. (d) If 10 or less words are returned, present the list to a user to select the correct transcription ("enlisted").

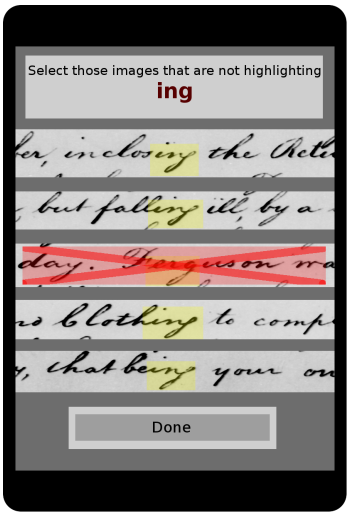

Fig. 6: A mock-up of what the user might see when verifying character n-gram spotting in the proposed system. The highlighted images are from the server spotting a particular n-gram. The red-boxed image has been selected by the user as it is a false-positive.

50% of the occurrences of each bigram), 41% of the words in the corpus can be narrowed down to 10 or fewer possible transcriptions. This is using a lexicon of 108,028 words and 6,939 names. 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 50% recall, 73% of the corpus can be transcribed with 250 spottings (i.e. going through the 100 bigrams 2.5 times). See Fig. 8 for more

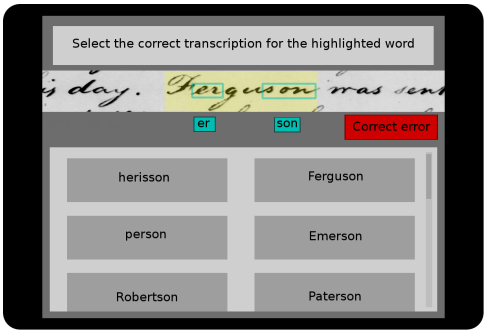

Fig. 7: A mock-up of what the user might see when selecting a correct transcription.

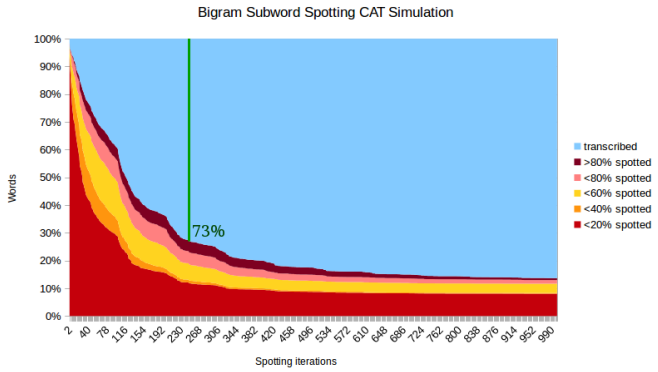

Fig. 8: Results of a simulation showing how much of the GW dataset can be transcribed after a given number of spotting iterations. The chart is drawn so one can observe the progress of spotting as well as transcription, each category (color) indicating the portion of words which have the given percent of their characters recognized by spottings (or indicating the portion of words transcribed).

thorough results of simulating this process. Preliminary results in spotting character n-grams in the GW dataset have yielded a mean-average-precision of 64% for bigrams and 72% for trigrams, using a naive sliding window adoption of Almazán et al's [8] spotting method.

V. Discussion

The verifying of spotted n-grams can be broken up into small tasks of a handful of confirmations. The selection of transcriptions will be from a short list making them small tasks. With the added element that these tasks do not require typing, they are ideal for mobile users. Each task takes a few seconds of judgement and one or more taps. Tasks can be rapidly completed for any amount of time a user is willing to spend; this dynamic will appeal to casual users in a way many other transcription methods (CAT or manual) do not. The rapid nature of the tasks also would lend to a variety of gamification methods that the system could be injected into.

There are a handful of limitations to this system. First, it is dependant on word segmentation. Even if the n-gram

spotting method is segmentation free, the words still need to be segmented to create the lexicon queries. Another limitation is that better results for sentence structured documents would be achieved if including NLP processing. Though it would be possible to include this, by processing possible transcriptions with the words around them, we have left it out to allow a greater variety of documents. The other limitations are based on performance of some of the pieces. The n-gram spotting algorithm may require some contextual exemplars (e.g. for training), requiring a few pages of a new corpus to be manually transcribed before the system can operate. If the n-gram spotting performs poorly, either users will be required to reject many spotting results or a low recall will have to be accepted, both options hurting the effectiveness of the system. If the word spotting/handwriting recognition algorithm, is unable to prune the list of possible transcriptions effectively, the lexicon size may have to be restricted in size. But we believe these limitations are negligable given the gains of the system and our preliminary subword spotting results.

A positive feature of the reliance on subword spotting is that the verification of spotted n-grams is somewhat language agnostic, meaning users might feel comfortable completing this task for a language other than their native tongue. An additional strength of our system is the large lexicon size is supports. The simulation described above was done with a lexicon size almost doubling what many large-vocabulary handwriting recognition systems use. Our system will be able to transcribe far more words, particularly uncommon names, than other CAT system's reliant on handwriting recognition methods.

We have presented a highly parallelizable CAT system which leverages character n-gram spotting to form partial transcriptions and, user interaction to maintain high accuracy and complete transcriptions. User interaction takes place as small selection tasks, ideal for mobile users. While it has yet to be implemented, we believe this system could make a large impact on the work that can be done with crowdsourced transcription by both providing an effective tool as well as one appealing to a largely untapped user-base.

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.

[8] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.