Applications of Subword Spotting

Brian Davis
Brigham Young University
briandavis@byu.net

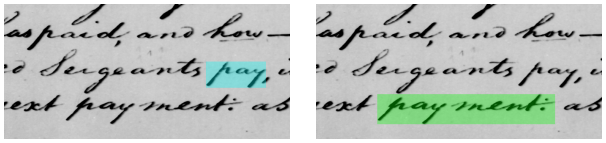

Figure 1: Examples of word spotting for 'pay' and 'payment'.

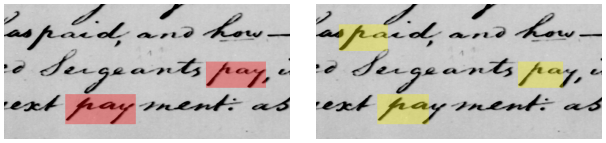

Figure 2: Examples of subword spotting for the character trigram 'pay' (left, red) and bigram 'pa' (right, yellow).

1 Introduction

In the domain of family history there are billions of handwritten documents which are being captured as digital images. The transcription of these lags far behind the capturing, and the gap is growing.

Automated handwriting recognition is a potential solution. While the state-of-the art methods work relatively well, particularly in single author scenarios [4, 6, 7], they do require large training sets.

Because being able to search over documents is sometimes the only desired utility, word spotting [1, 2, 5] has emerged as an alternate method of processing documents. In word spotting, the goal is to make a collection of document images searchable without transcription. The search results are based on visual features the system extracts. The result of word spotting is the location, and potentially bounding box, of all instances of the query (see Figure 1).

Where word spotting finds words matching a query, **subword spotting** relaxes this to finding any instances of the query, even within a word. As seen in Figure 2 (left, red), an additional instance of "pay" is found compared to Figure 1. And in Figure 2 (right, yellow), an additional instance of "pa" is found, which turns out to be a different form of the word"pay".

Applications of subword spotting can extend search capabilities and enhance transcription. Important use cases for subword spotting include searches (1) where a root or part is desired to be searched, such as querying "pay" wanting to find instances of "payment", "payments", "prepay", etc., (2) where there are character(s) which a human transcriber cannot initially recognize, but if instances elsewhere in the document in familiar words could be found in context, would become discernible to the transcriber.

We first briefly describe our method of subword spotting. We next show exploratory results in applying subword spotting to suffix spotting and assisting human transcribers in finding instances of unrecognized characters. We do not demonstrate polished results, but show that these tasks are intuitively accomplished with the use of subword spotting.

2 Method

Our subword spotting is built on the segmentation based word spotting method PHOCNet [5] which we adapted to perform a sliding window over word images. [5] uses a deep convolutional network trained on word images with pyramidal histogram of characters (PHOC) [1] as the target vectors. PHOC encodes characters and their approximate location as a fixed length vector. This method can spot using both query-by-string (QbS) and query-by-example (QbE); that is, a query may be a text word or an image.

3 Suffix Spotting

There are certain situations where one may want to search for a partial word, such as a prefix or suffix. For example, if one wanted to find names of towns in

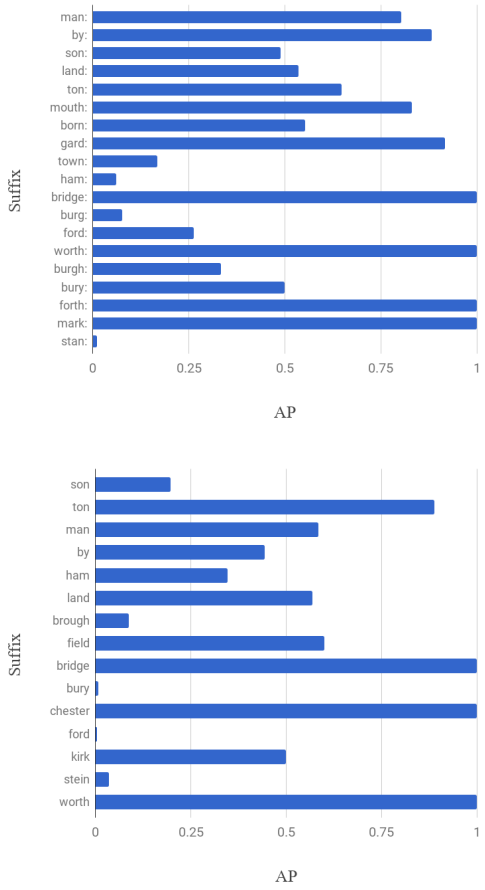

Figure 3: Suffix spotting average precision (AP) of individual suffixes for the IAM (top) and the census names (bottom) datasets. Arranged in descending order of frequency in test sets.

a corpus of German documents, spotting words with the suffix "-berg" should return many town names.

We identified a list of 41 suffixes and evaluate spotting the suffixes present from this list in the IAM dataset [3] and a set of cropped names from the US 1930 Census. These are spotted in a QbS fashion, restricting the sliding window to be on the right-hand side of the word.

We show the average precision for the suffixes individually in Figure 3. We believe this method of search could be extended to include text search similar to regular expressions.

4 Manual Transcription Assistant

Frequently when a person is transcribing a handwritten document they come across a handwritten word they do not recognize. A common solution is to scan the document for similar shapes which are present in the difficult word. If the transcriber can find the same letters in the context of a word they do recognize, the transcriber can identify the letters. However, characters may be time consuming to find due to the density of the document and rarity of the characters.

Using the unknown characters' image as a query allows QbE subword spotting to automate this scanning task. We have created a proof-of-concept assistant program which does this. Figure 4 demonstrates how our assistant works.

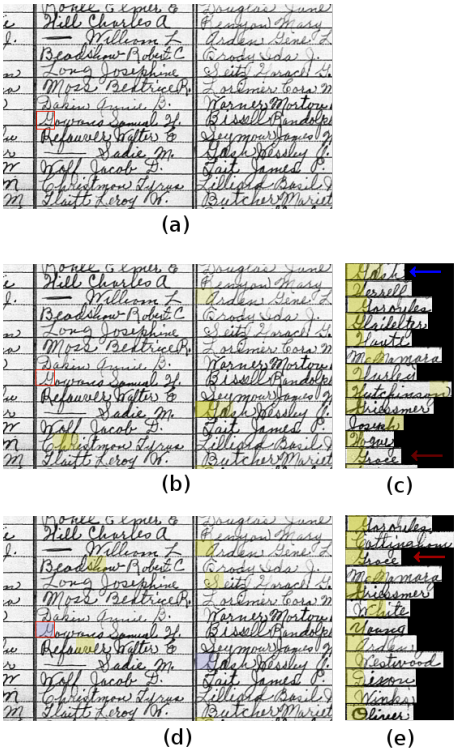

Figure 4: Demonstration of transcription assistant employing subword spotting. First the unrecognized characters are selected (red) in a document **(a)**, in this case a "G", and the bounding box is snapped to one with a precomputed PHOC vector. This PHOC vector is compared against all others of the same window size. The ranked list is shown to the user both as a list **(c)** and by highlighting the results with the color intensity representing the strength of the match **(b)**. Suppose that the top results aren't words the user is familiar with, but the user recognizes the "G" in the first result (blue arrow, "Gash"). By selecting it, the system performs another QbE search, combining the results with that of the original query. **(d)** and **(e)** shows the combine results. In **(e)** the word "Grace" (red arrow), a more recognizable word, has moved to the third spot.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (12):2552–2566, 2014.

[2] R. Manmatha, Chengfeng Han, and E. M. Riseman. Word spotting: a new approach to indexing handwriting. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 631–637, June 1996. doi: 10.1109/CVPR.1996.517139.

[3] U. Marti and H. Bunke. The IAM-database: An english sentence database for off-line handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[4] Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Mauricio Villegas, and Enrique Vidal. Icdar2017 competition on handwritten text recognition on the read dataset. In *Proceedings of ICDAR*, 2017.

[5] S. Sudholt and G. A. Fink. Evaluating word string embeddings and loss functions for cnn-based word spotting. In *Proceedings of ICDAR*, 2017.

[6] J. A. Snchez, V. Romero, A. H. Toselli, and E. Vidal. Icfhr2016 competition on handwritten text recognition on the read dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 630–635, Oct 2016. doi: 10.1109/ICFHR.2016.0120.

[7] Curtis Wigington, Seth Stewart, Brian Davis, and Bill Barrett. Data augmentation for recognition of handwritten words and lines using a cnn-lstm network. In *Proceedings of ICDAR*, 2017.