Text-image alignment for historical handwritten documents

S. Zinger[a], J. Nerbonne[b], L. Schomaker[c]

[a]Video Coding and Architectures Research group, Eindhoven University of Technology,
P.O. Box 513, 5600MB Eindhoven, the Netherlands;
[b]Center for Language and Cognition Groningen, University of Groningen
Oude Kijk in 't Jatstr. 26, 9700 AS Groningen, The Netherlands
[c]Artificial Intelligence Department, University of Groningen
Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

ABSTRACT

We describe our work on text-image alignment in context of building a historical document retrieval system. We aim at aligning images of words in handwritten lines with their text transcriptions. The images of handwritten lines are automatically segmented from the scanned pages of historical documents and then manually transcribed. To train automatic routines to detect words in an image of handwritten text, we need a training set - images of words with their transcriptions. We present our results on aligning words from the images of handwritten lines and their corresponding text transcriptions. Alignment based on the longest spaces between portions of handwriting is a baseline. We then show that relative lengths, i.e. proportions of words in their lines, can be used to improve the alignment results considerably. To take into account the relative word length, we define the expressions for the cost function that has to be minimized for aligning text words with their images. We apply right to left alignment as well as alignment based on exhaustive search. The quality assessment of these alignments shows correct results for 69% of words from 100 lines, or 90% of partially correct and correct alignments combined.

Keywords: handwriting, alignment, cost function

1. INTRODUCTION

Our project aims at information retrieval from handwritten documents. In this article we present our work on aligning text words from text transcriptions with their images automatically extracted from handwritten lines.

We work with the archive of the Queen's Office (Kabinet van de Koningin - KdK) collection of the Nationaal Archief in the Hague (the Netherlands). This collection is dated 1903 and contains short paragraphs of 5-10 lines. Each paragraph describes a decision or an order of the Dutch Queen.

Currently the information retrieval requests from users are processed manually, by archivists. We aim at retrieving historical handwritten documents given a textual query. This will considerably facilitate access to information, making it quicker and easier. Figure 1 shows how we would like our system to work: the user enters a word he wants to find in a manuscript, and the search engine finds and highlights this word in the handwritten text. Creating a search engine for handwritten documents is a challenging problem.[1] We aim at creating such a system using a set of transcribed training data so that it will enable us to retrieve untranscribed documents as well.[2] Previous research on historical document retrieval systems led to difficulties in merging handwritten text with its annotations.[3] Word spotting may also be used for indexing handwritten documents,[4] but it is difficult to distinguish words automatically in a handwritten text. We will show an example of word segmentation for our data and describe the difficulties we encounter.

Send correspondence to Svitlana Zinger. E-mail: sveta_zinger@yahoo.com

Figure 1. An ideal case: searching in the handwritten text.

At first we work with entire lines of handwritten text: we divide an image of a handwritten document into lines using its visual features. These lines then can be transcribed and used for training a handwritten document retrieval system. Text composed of transcriptions of handwritten lines is relatively small and has a number of particular aspects: highly redundant context, many administrative terms; lack of redundancy in the dates, numerals and proper names; special abbreviations and terms that cannot be found in other collections.

Having the images of handwritten lines and their transcriptions we can align them so that the text words are aligned to their handwritten equivalents. This alignment is important because we can use it in several ways:

- make easier the reading of the handwritten text - a user points at a word and gets its transcription;
- images of words are directly usable for pattern matching on untranscribed data;
- database of images of words may be used for training machine learning algorithms.

Research on alignment has been done using different types of sequences. Edit distance, longest common subsequence, longest common substring are applied for bio-sequences alignment, animal songs comparison, speech recognition.[5] Sequence alignment is also used for dialect comparison[6] and for statistical machine translation.[7] In all examples mentioned above the sequences are represented by the same type of media: digital text for machine translation, audio signal for speech recognition, etc. In our case we align data represented by different media: digital text and image. This poses additional problems and several solutions have been proposed in the literature. For aligning text with video, text is converted to audio signal and then aligned with that present in the video sequence.[8] For aligning images of handwritten text with their transcripts, a pattern recognition is applied to the images, that leads to text which is subsequently aligned with the transcripts.[9]

The alignment presented in this paper does not rely on pattern recognition techniques: we do not perform OCR on the handwritten text before aligning it. Dynamic time warping for aligning text and images[3] uses the hypothesis that the alignment of several handwritten words to one text word and vice versa is possible. This is the case when we expect that there is no correspondence in the order of elements of the sequences to align. Such problems are solved by dynamic programming. In this article we use another hypothesis: the order of words in the transcriptions and the order of words on the image of a handwritten line correspond to each other. The element of alignment is a word - a sequence of characters separated by a space from other such sequences.

2. TEXT-IMAGE ALIGNMENT

2.1 Data Description and Baseline Alignment

We currently have 1200 scanned pages from the KdK pages. And we expect more than 40000 pages to be scanned during the coming several months. Both the available and the expected data is from the KdK collection and is created by the same person: therefore the handwriting and its properties do not vary.

Figure 2. Tool for annotating handwritten lines. Pressing the Edit button below an image of a line, user can enter new transcriptions or change the existing ones.

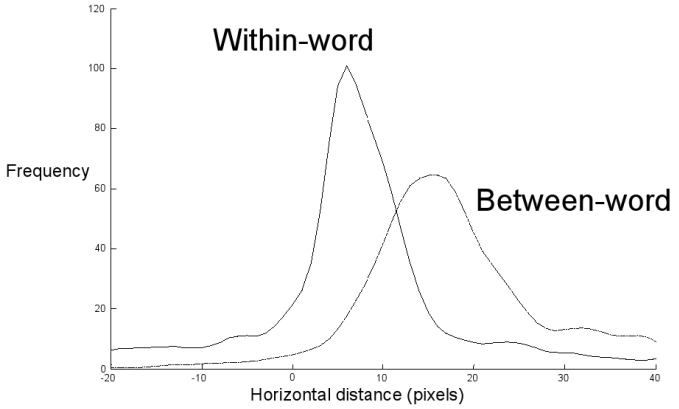


Figure 3. Histogram of distances between and inside words, calculated on 200 lines.

We transcribe lines of handwritten text using a tool developed in the beginning of our project. The lines are extracted automatically from the scanned pages. A screenshot of the tool we use for creating transcriptions in Figure 2 shows lines from a KdK book - our initial data.

At first we explore the performance of word segmentation on our data. Segmenting handwritten words in KdK books is a difficult problem. For some lines it is impossible to set a threshold on the spaces between words so that no word is broken in two or more pieces. The spaces between and inside words are often approximately the same as we can see from Figure 3. We perform experiments on word segmentation on 100 lines (650 words) randomly chosen from our data. Applying a threshold on the length of spaces in the vertical projection of pixels on the horizontal axis, we get 51% of words correctly segmented. The threshold is adapted for every line by taking a median value of all the spaces we found in this line. An example of segmentation is on Figure 4. We have also performed segmentation on words ignoring the regions of the image that lie outside the line with the highest density (less than 2 times smaller than the peak value) of ink when projected on the vertical axis. This eliminates ascenders and descenders of handwritten letters as well as the noise caused by line segmentation. The results of word segmentation did not improve though. These results on word segmentation motivated us to begin with lines and not words as initial elements of our system.

Even though matching images of handwritten lines gives us some information about the unannotated images,[10] it still does not provide information on where the word of interest is located in the line. Using n-grams on the available 25000 lines of text annotations, we can define the indicator-words that precede and succeed the proper names that are the words of interest. For example, the Dutch words equivalent to “in”, “from” and “municipality”

Figure 4. Segmentation of words: above - real case, below - ideal case.

Figure 5. Example of an alignment that we consider to be partially correct: the word "Dusseldorf" is aligned with "te Dusseldorf" (in English - "in Dusseldorf").

often precede a geographical name in the transcriptions of our data. Finding these indicator-words in a line requires them to be learnt in advance: instances of these words are needed for learning.

In order to align images with their transcriptions we cut the image of a handwritten line into pieces that are separated by the longest $N-1$ spaces where n is the number of words in the transcription of this line. The result of this experiment on 100 lines provides 45% of words correctly aligned and 11% - partially correct. We consider a text-image alignment result to be partially correct if the largest part of the aligned image contains the largest part of the text word (Figure 5). This means, for example, that the alignment is partially correct if the first letter of the word is missing on the image of the word, or that the word is on the image but there are also several letters of the neighboring word. In the following section we present our method and results on getting images of handwritten words aligned with their text transcription.

2.2 Alignment as Cost Function Minimization

Since handwritten lines overlap and often cannot be automatically fully separated, there is noise present in most of the segmented lines. This is the reason why we do not use visual features such as detected ascenders and descenders[3] in order to perform or improve the alignment (Figure 6).

We notice that the relative lengths of ASCII and handwritten words (as fractions of the line length in characters and pixels respectively) are highly correlated: correlation coefficient is 0.965 for 14 ideally aligned lines that contain 87 words (Figure 7). We expect that using these relative lengths of words for the alignment will improve its results.

Let us consider an image of a handwritten line and its text transcription. The transcription contains N words, a word $n \in [1:N]$ contains W_n letters. The handwritten line on the image is L pixels long; b is a boundary of a

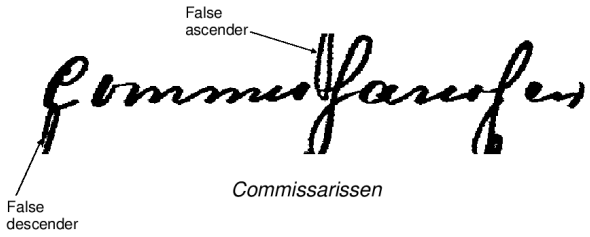


Figure 6. False ascender and descender that occur as a result of line segmentation.

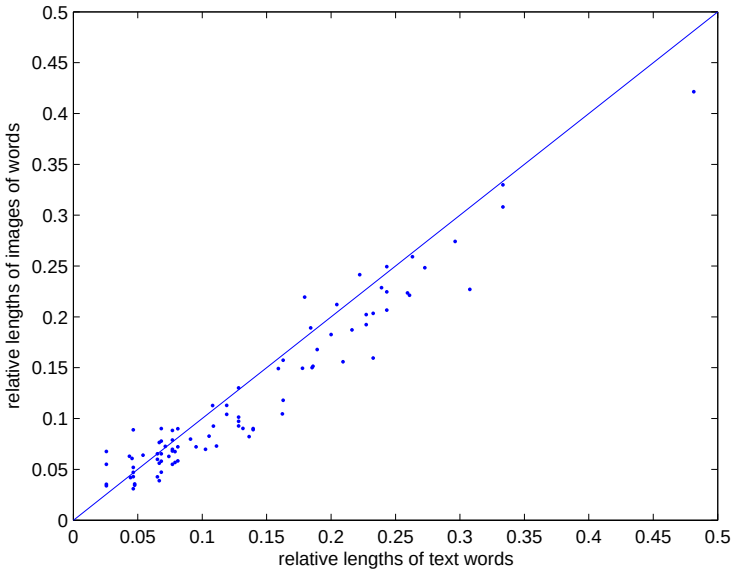


Figure 7. Points indicate relative lengths of words in characters on the horizontal axis and in pixels on the vertical one. In the ideal case all points lie on the solid line.

sought handwritten word, it is located on a space between portions of handwriting. The spaces are found using the ink projection of the unslanted image on the horizontal axis. We define the cost function

$$f_n(i) = |h(i) - t_n|, \tag{1}$$

where n is a word we currently consider; $t_n = \frac{W_n}{\sum_{j=1}^{N} W_j}$ - relative length of this text word, $h(i) = \frac{b-i}{L}$ - relative length of a handwritten word. The solution is the optimal word boundary i' that minimizes the cost function:

$$i' = arg\,min\, f_n(i). \tag{2}$$

The boundary b of the handwritten word is fixed: it can be word's left or right boundary depending on the direction in which the alignment is performed. The second boundary is then found using the cost function minimization. In our case we fix the right boundary and search the left one by minimizing the cost function. We minimize the cost function by finding its values for all possible i and selecting i' that minimizes the function. Once we found i', we fix it and look for the optimal boundary of the next word. The cost function is minimized N times: an optimal boundary is found for each word from the transcription.

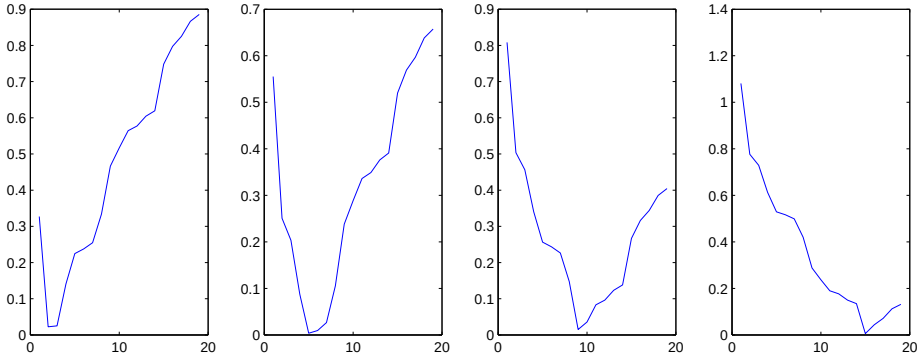
A slight modification of the cost function above can be obtained by replacing the length of a text word in letters by its length as a sum of statistically estimated letter lengths.[11] The letter lengths are estimated using the manually cut letters.

In the case of exhaustive search for the best alignment, the cost function is

$$f(C) = \sum_{n=1}^{N-1} |h(b_n(C)) - t_n|, \tag{3}$$

where C is an ordinal number of a combination of $N-1$ spaces; all possible combinations of spaces are considered so that each combination divides the image on N pieces that are presumably words; $h(b_n(C)) = \frac{b_{n-1}(C) - b_n(C)}{L}$ - relative length of a handwritten word.

(a) Handwritten line consisting of four words



(b) Energy functions are computed for finding each word's boundary: ordinal number of a boundary is on the horizontal axis, energy values - on the vertical axis.

Figure 8. Energy for left to right alignment: energy is minimized for every word.

Table 1. Experimental results on aligning 100 lines.

	Largest spaces (baseline)	Left to right alignment	Left to right using estimated letter lengths	Exhaustive search
Correctly aligned (%)	45	64	64	69
Partially correct (%)	11	19	18	21
Wrong (%)	44	17	18	10

3. RESULTS COMPARISON

Table 1 shows the evaluation of the quality of the alignments on 100 lines. We test four alignment methods described above:

- alignment using the largest spaces between portions of handwritten text,
- left to right alignment using text word length in letters,
- left to right alignment using text word length as a summation of estimated lengths of letters it contains
- and alignment through the exhaustive search using text word length in letters.

Each of these methods is applied to the whole test set - 100 lines. The time performances for the first three alignment methods do not vary much: 0.79, 0.80 and 0.82 seconds respectively for the longest spaces based alignment, left to right alignment and left to right alignment using estimated character lengths. Much more time is needed for the exhaustive search: 23.30 seconds. These results are average time performances in Matlab for the 100 lines. We expect much better time performances for the C implementation of our algorithms. From Table 1 and its graphical representation on Figure 9 we can see that using the exhaustive search alignment through cost function minimization gives the best results: the percentage of wrong alignment is the smallest in comparison with the other three methods. We can also see that using the statistically estimated letter lengths for calculating the length of a text word does not improve the results of alignment.

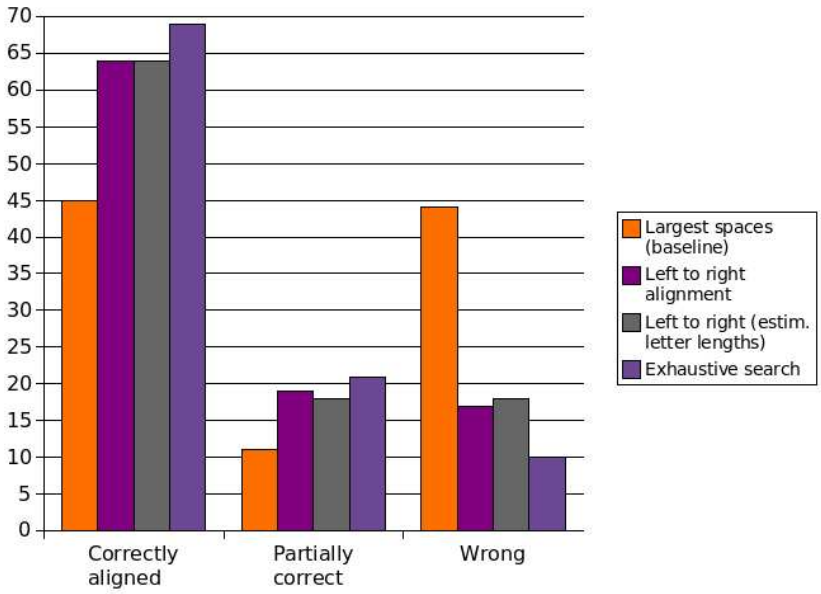


Figure 9. Graph with the evaluation results of alignments.

4. CONCLUSIONS AND FUTURE WORK

We have applied and analysed several text-image alignment techniques that we consider suitable for our historical handwritten documents. The assumption made for the alignments is that the order of words in transcriptions and on images is the same. Our methods are not sensitive to the quality of the line segmentation because we cut the image areas containing ascenders and descenders before finding spaces on a handwritten line. The best results are obtained by minimizing a cost function we introduce for the exhaustive search for the best combination of spaces. Since we use relative lengths of words, our alignment methods do not depend on the actual handwritten word length: some lines contain 2 words and some - 10 words, and the length of the same word in these two lines can vary considerably.

The methods presented in this article avoid pattern recognition. A possible future word is detecting uppercase letters in the handwritten text and incorporating the information on the location of these letters in the cost function. Another option is to render the text of transcriptions and explore the alignment possibilities then. Further one can think about aligning syllables, letters or connected components.

REFERENCES

[1] S. Srihari, C. Huang, and H. Srinivasan, "A search engine for handwritten documents," in *Document Recognition and Retrieval XII, SPIE*, pp. 66–75, (San Jose, CA, USA), 2005.

[2] T. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proceedings of the ACM SIGIR 2004 Conference*, pp. 369–376, (Sheffield, UK), 2004.

[3] E. Kornfield, R. Manmatha, and J. Allan, "Text alignment with handwritten documents," in *Proceedings of Document Image Analysis for Libraries (DIAL)*, pp. 195–209, (Palo Alto, CA, USA), 2004.

[4] S. Uchihashi and L. Wilcox, "Automatic index creation for handwritten notes," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 3453–3456, (Phoenix, AZ, USA), 1999.

[5] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, CSLI Publications, 1999.

[6] M. Wieling, T. Leinonen, and J. Nerbonne, "Inducing sound segment differences using pair hidden markov models," in *Computing and Historical Phonology: 9th Meeting of ACL Special Interest Group for Computational Morphology and Phonology Workshop*, pp. 48–56, (Prague, Czech Republic), 2007.

[7] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, (University of Maryland, College Park, MD, USA), 1999.

[8] S. Maji and R. Bajcsy, "Fast unsupervised alignment of video and text for indexing/names and faces," in *Proceedings of the Workshop on multimedia information retrieval on The many faces of multimedia semantics*, pp. 57–64, (Augsburg, Germany), 2007.

[9] A. Toselli, V. Romero, and E. Vidal, "Viterbi based alignment between text images and their transcripts," in *Proceedings of the Workshop on Language Technology for Cultural Heritage Data*, pp. 9–16, (Prague, Czech Republic), 2007.

[10] L. Schomaker, "Retrieval of handwritten lines in historical documents," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, pp. 594–598, (Curitiba, Brazil), 2007.

[11] J. Mackowiak, L. Schomaker, and L. Vuurpijl, "Semi-automatic determination of allograph duration and position in online handwritten words based on the expected number of strokes," in *Proceedings of the 5th International Workshop on Frontiers in Handwriting Recognition*, pp. 433–436, (Colchester, UK), 1996.