Combining Alignment Results for Historical Handwritten Document Analysis

Emanuel Indermühle[1] Marcus Liwicki[2] Horst Bunke[1]
[1]Institute of Computer Science and Applied Mathematics
University of Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
{eindermu, bunke}@iam.unibe.ch
[2]German Research Center for AI (DFKI GmbH)
Knowledge Management Department, Kaiserslautern, Germany
marcus.liwicki@dfki.de

Abstract

In this paper we propose a new strategy for combining the outputs of several alignment systems. Based on the word boundaries retrieved from a number of individual alignment systems, the new boundaries are estimated. We investigate three strategies for this estimation. First, the mean value of the individual boundaries is taken, second the median is selected, and third, confidence values of the alignment systems are considered. We apply the combination strategies on a word mapping system for historical handwritten manuscripts. After some preprocessing and normalizing steps, three differently trained hidden Markov model based handwriting recognizers are applied to the text lines in forced alignment mode. As a result, the positions of the word boundaries are obtained. In in a number of experiments it is shown that a combination strategy based on the median outperforms the others and all individual alignment systems with a word mapping rate of about 95%.

1. Introduction

Alignment of pattern streams to their transcription is a problem which appears in speech and handwriting recognition [12, 16] as well as in the analysis of other data, such as amino acid or nucleotide sequences [4]. Its difficulty depends on the underlying data and on the length of the sequence observed. In handwriting recognition it is still an open issue.

In the emerging application area of historical document analysis [1] the results of forced alignment may be used to make the contents of documents browsable and searchable. In contrast to word spotting [9, 14], which allows one to search for specific words, transcription mapping enables a fast generation of ground truth for word recognition [17] or applications such as browsing within a handwritten manuscript. This is particularly useful for scientists in the fields of literature or graphology.

Alignment of handwritten historical manuscripts to their transcriptions has already been investigated by Tomai et al. [16]. The authors have segmented historical documents into text lines and generated different hypotheses of word segmentation for each text line. The best mapping between the words of the transcription and the result of a word recognizer are used to produce the requested alignment. Kornfield et al. [5] have segmented documents into lines and words. Features are extracted for each word of the segmentation and in the transcription. Dynamic Time Warping is then applied to both word sequences resulting in an alignment. For a line by line matching, the authors report that this method reaches 74.5 out of 100 *box level averaging*[1], while they achieve 75.4 with Tomai's method. Rothfeder et al. [15] describe an alignment of ASCII text to the result of a word segmentation procedure. Since word segmentation is not absolutely reliable, the authors use an HMM based algorithm for the alignment to handle over- and undersegmentation. With this approach similar results as in [16] and [5] were achieved. However, all three methods depend on a preceding word segmentation. According to Kornfield et al. [5] this is a major source of error.

In contrast with previous work, the word segmentation is integrated in the transcription mapping task in this paper. We use an HMM based handwriting recognizer accepting whole text lines as input. The alignment is applied to historic manuscripts that are mapped to their counterparts in the printed edition. Since machine printed editions of the manuscripts do often exist, one can assume that digital versions of the texts are available and can be used for the forced alignment.

In previous work we have investigated different systems

[1] measurement defined in [5]

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.19

to recognize historical documents directly [3]. There we showed that the highest recognition results are achieved if the system is trained with a large writer independent dataset and then adapted to a writer specific training set. In this paper we study similar issues for forced alignment. Furthermore, we combine different alignment systems to improve the performance.

The rest of this paper is organized as follows. In Section 2 we introduce our forced alignment procedure, which is based on an HMM recognizer, and in Section 3 new methods for combining alignments are proposed. Section 4 gives an overview of the underlying data and the preprocessing steps. Experiments and results are presented in Section 5. Finally, Section 6 draws some conclusions and provides an outlook to future work.

2. Forced Alignment with HMM Recognizers

Forced alignment is the task of generating an alignment between the states of an HMM and the observed feature vector sequence. This is done by forcing a recognition result. The alignment can be used to get a mapping between a transcription and the data, e.g. the speech file or the text image. The states in the state sequence that represent the end of a letter HMM are marking a border between two letters. The corresponding feature vectors, given by the alignment, are the keys to the locations of the borders in the text image. This paper is focused to the mapping of whole words to the location of their counterpart in the text image. The forced alignment is applied to single text lines.

In the HMM-recognizer the alignment of feature vector sequences and HMM state sequences is obtained during the recognition process by applying the decoding procedure to a feature vector sequence and additionally specifying its transcription. When HMM-recognizers are applied in recognition mode, the Viterbi algorithm evaluates the optimal alignment of the feature vectors to the HMM states of every word combination given by vocabulary and language model. When the best alignment is found the state sequence can be retrieved from information collected during this process. In the forced alignment mode only the word combination given by the transcription has to be evaluated. The algorithm terminates as soon as the optimal alignment is found.

Since the same system is used for recognition and forced alignment, also the same parameters need to be tuned. In [2] it was shown that optimizing the number of Gaussian components in the observation probability distribution has a positive impact on the performance of a handwriting recognition system. Hence it is reasonable to assume that the same is true for forced alignment.

The forced alignment procedure used in this work is similar to the one proposed in [17]. The HMM based system that performs the forced alignment is identical to the recognizer for handwritten text lines introduced in [10]. In [3] it is described how this recognition system is configured to work with the handwriting data presented in Section 4.1.

3. Combination of Alignments

Multiple classifier combination [6] has been investigated in different areas. It is expected that a combined system can benefit from the strength of the individual classifiers. Having different word mapping systems at hand leads to the idea of combining them. Given the word borders of the n different alignments, sorted in ascending order according to their position $b_1, \ldots, b_n$, we have investigated three methods for combining them, as described below.

For *mean combination* the position of the new word border is identified by taking the mean of the different word border's positions $\overline{b} = \sum_{i=1}^{n} \frac{b_i}{n}$. This method is used in many other multiple classifier systems as well. However, if different word borders have been found, in most cases the correct border is one of them and does not lie in between.

The *median combination* overcomes the problem of the mean combination. We set $\overline{b} = b_{\frac{n+1}{2}}$ if n is odd. Given an even number of classifiers, either one of the central values $b_{\frac{n}{2}}$ or $b_{\frac{n+2}{2}}$ could be taken (or their average $\frac{1}{2}(b_{\frac{n}{2}} + b_{\frac{n+2}{2}})$). The advantage of this method is that outliers have no influence on the result. Additionally, the new position is the position of one of the old word borders which means that this position has already been found to be a word border.

The *confidence combination* is the most advanced method proposed here. It additionally takes the classifiers' confidence into account. The combination procedure consist of two parts. First, outliers are eliminated. Second, the border with most confident alignment is selected. A word border b_i is identified as an outlier if $\frac{d_i}{\overline{d}} > L$ where

$$d_i = \sum_{j=1}^{n} abs(b_i - b_j), \tag{1}$$

$$\overline{d} = \frac{1}{n} \sum_{i=0}^{n} d_i, \tag{2}$$

and L is a constant. In other words, the sum d_i of the distances to the other word borders $b_{j \neq i}$ must be lower than L times the mean of the same value d of the other borders. If L is set to a smaller value, more outliers will be eliminated. In case of three word borders, L decides if one word border will be eliminated. Practically, this is similar to majority voting as proposed in [7] for classifiers. In this work L is chosen to be 1.5. For the second step of the *confidence combination* procedure, a confidence value is needed. It is based on an empirical observation made when analyzing the alignments. We noticed that faulty alignments often have a

larger variance in the width of single characters. This leads to the idea to take this variance – or, more precisely, its inverse – as the confidence value.

4. The System

4.1. Data of the Swiss Literary Archives

The Swiss Literary Archives[2] in the Swiss National Library maintain a large collection of various works from Swiss authors. In particular, this collection includes a huge number of original handwritten documents. In the context of research in literature, there is an interest in automated tools that make all these materials electronically accessible. Systems that are able to automatically align text in ASCII format with images of the manuscript would be a very useful tool in this research.

The handwriting material used in the study described in this paper consists of two volumes of poetry manuscripts from Swiss author Gerhard Meier (1917–2008), including 145 pages and 1,640 lines in total. The pages are provided as digital gray scale images with a size of 4,000 by 3,000 pixels. As ground truth to this handwriting, a digitized version of the machine printed edition is used.

Gerhard Meier used to write on individual sheets of paper, which he collected in a folder. He used rather generous spacing between lines, which is an advantage for the line segmentation. However, there are artifacts such as underlining of titles, arrows, lines indicating the end of a page, lines separating two consecutive paragraphs from each other, punching holes, and page margins (see Figure 1).

4.2. Preprocessing

The proposed alignment system needs individual text lines as input. Since better alignments are achieved if the text lines are normalized, some preprocessing steps are applied. First the images are binarized using Otsu's algorithm [13]. Then the pages are segmented into individual lines with a recently developed line extraction procedure [8]. Finally, skew correction, slant correction, height normalization, and width normalization are applied to the text line images.

The normalized text line images are the input to the alignment system. To apply the HMM based system, features are extracted. For this purpose, a sliding window of one pixel width is moved from left to right over the image. At each position of the window, a vector of nine features is extracted. The features extracted are the number of pixels, the center of gravity of the pixels, the second order moment

[2]Website: `http://www.nb.admin.ch/slb/`

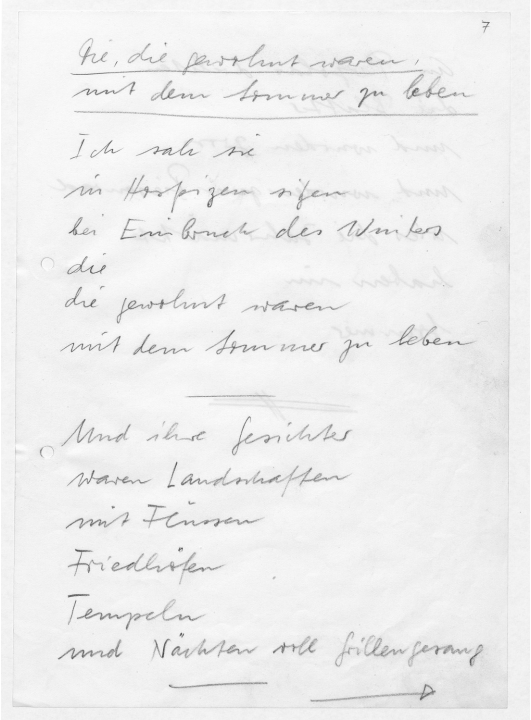

Figure 1. Page of a manuscript by Gerhard Meier.

of the window, the location of the upper-most and lower-most pixel, the orientation of the upper-most and lower-most pixel, the number of black-white transitions, and the number of black pixels divided by the number of all pixels between the upper- and the lower-most pixel. For more details see [10].

5. Experiments and Results

5.1. Experimental Setup

The poems by Gerhard Meier were randomly divided into a training set, a validation set and a test set. To study the influence of the amount of training data on the forced alignment results, the size of the training set is increased from 200 to 500, 1,000, and 2,400 word instances. The validation set and the test sets are fixed. The validation set contains about 950 words and the test set 1,400 words. All three sets are mutually disjoint.

For HMM-based recognizers, training is essential. In this paper three different training strategies are tested, resulting in three different systems. The first system (referred to as *basic system*) is trained directly from scratch with the training set. The second system (referred to as *iamdb system*) is trained on a subset of the IAM-database [11]. This subset contains 6,161 lines and 53,841 words from 283 writers. The third system (referred to as *adaptation system*) is

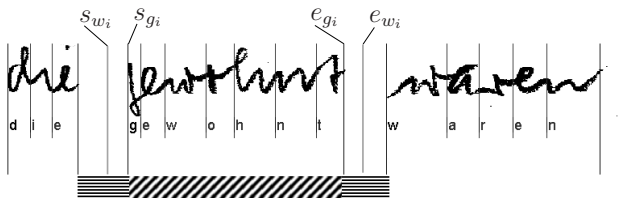

Figure 2. The word "gewohnt" is extracted correctly if the borders are located in the horizontally shaded region.

trained on the IAM-database and then adapted to the training sets used for the basic system. We refer to [3] for more details.

To measure the system performance we define the *word mapping rate*. For this purpose, the alignment is compared with an alignment created by hand, the *alignment ground truth*. To test the word w_i against its counterpart g_i in the alignment ground truth, the initial borders s_{w_i}, s_{g_i} and the final borders e_{w_i}, e_{g_i} are considered. The word mapping rate is the percentage of segmented words that fulfill the following requirements $e_{g_{i-1}} \leq s_{w_i} \leq s_{g_i}$ and $e_{g_i} \leq e_{w_i} \leq s_{g_{i+1}}$ as shown in Figure 2. Note that the correct borders must be used to extract a word. Otherwise the mapping fails. Lines with only one word are ignored, since word mapping is a trivial task there.

5.2. Results

The results of the three differently trained systems are shown in Fig 3. All three systems achieve results within a

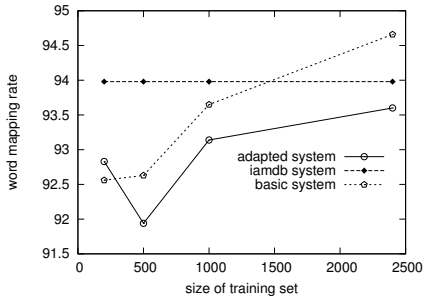

Figure 3. Word mapping rate on the test set vs. the size of the training set. Note that the training set is not part of the iamdb system.

close range. Surprisingly, the adapted system is inferior to the iamdb system. When 500 words are used for adaptation, the difference is statistically significant. This means that adaption has a negative effect on the ability to recognize word boundaries, at least for a small training set. It means also that the task of finding word boundaries is quite independent of the writing style. The basic system, however, slightly outperforms the other systems with a word mapping rate of 94.66% if enough training data is available.

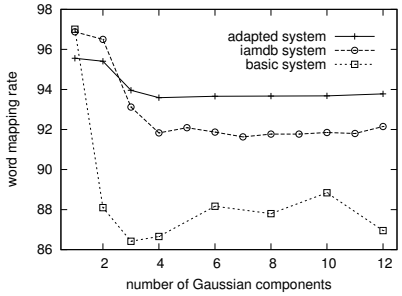

Figure 4. Word mapping rate on the validation set vs. the numbers of Gaussian components in the observation distribution of the HMMs.

The influence of the number of Gaussian components in the observation distribution of the HMMs on the word mapping rate is displayed in Figure 4. The best results are achieved with only one Gaussian component. This is surprising because – according to [2] – for handwriting recognition, the system works best with about 12 Gaussian components. However, there is a reasonable explanation for this observation. Since word mapping is requested, the white-space character is very important. The feature vectors of all white-space character have just one accumulation point, namely, the feature vector of a white column. If more than one Gaussian component is used to describe the distribution, irregularities at the border of the white-space become overemphasized and this leads to the effect that white-spaces are recognized too wide. Hence, a lower word mapping rate results.

When the alignments of the three systems are combined with the methods described in Section 3 the word mapping rate can be improved (see Figure 5). For a training set containing 1,000 words the improvement over the iamdb system is even statistically significant (using a dependent t-test for paired samples with a significance level of $\alpha = 0.05$). The mean combination does not reach the performance of the other combinations and individual systems. This is caused by the influence of outliers. The confidence combination is slightly better than the best individual system. The median combination can improve the word mapping rate significantly.

6. Conclusion

The mapping of historical handwritten documents to their transcriptions has gained some attention in recent years. In this paper we have described a system implemented upon an HMM handwriting recognizer that address this problem. The method has been tested on handwritten historical manuscripts by Swiss author Gerhard Meier. In

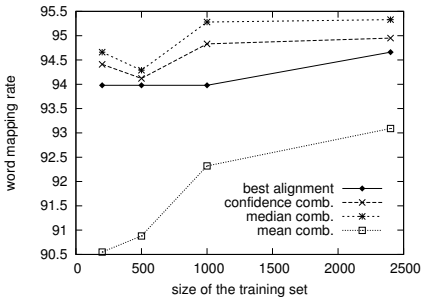

Figure 5. The word mapping rate of the combination methods in comparison to the best individual system on the test set.

this paper we applied an HMM-recognizer in forced alignment mode to generate the mapping.

Our results achieved with forced alignment lead to the conclusion that transcription mapping systems for historical documents can be built upon a handwriting recognizer based on HMM. The best strategy to train the recognizer depends on the size of the training set. If it is small, a writer independent training is preferable, otherwise a system trained on the writer specific training set might be the better choice. The results can be improved if different alignments are combined. Out of the three methods we proposed for combination, the median combination performs best and leads to a statistically significant improvement.

In future it would make sense to test the system described in this paper as a whole, considering the problem of unknown line breaks and differences between the manuscript and the printed edition. Moreover, although the third combination method takes a confidence value into account, it was not able to outperform the combination based on the median values. Therefore, other confidence measures could be an interesting additional subject for future work.

Acknowledgments

We thank Dr. Corinna Jäger-Trees, curator of the Gerhard Meier fonds in the Swiss Literary Archives, and her colleagues for their kind support. This work has been supported by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

References

[1] A. Antonacopoulos and A. C. Downton. Special issue on the Analysis of Historical Documents. *Int. Journal Document Analysis Recognition*, 9(2):75–77, 2007.

[2] S. Günter and H. Bunke. HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and Gaussian components. *Pattern Recognition*, 37:2069–2079, 2004.

[3] E. Indermühle, M. Liwicki, and H. Bunke. Recognition of handwritten historical documents: HMM-adaptiation vs. writer specific training. In *Proc. Int. Conf. on Frontiers in Handriting Recognition*, pages 186–191, 2008.

[4] K. Karplus and B. Hu. Evaluation of protein multiple alignments by sam-t99 using the balibase multiple alignment test set. *Bioinformatics*, 17(8):713–720, August 2001.

[5] E. M. Kornfield, R. Manmatha, and J. Allan. Text Alignment with Handwritten Documents. In *Proc. 1st Int. Workshop on Document Image Analysis for Libraries*, pages 195–205, 2004.

[6] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons Inc, 2004.

[7] L. Lam and S. Y. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man and Cybernetics*, 27(5):553–568, 1997.

[8] M. Liwicki, E. Indermühle, and H. Bunke. Online handwritten text line detection using dynamic programming. In *Proc. 9th Int. Conf. on Document Analysis and Recognition*, volume 1, pages 447–451, 2007.

[9] R. Manmatha, C. Han, and E. M. Riseman. Word spotting: A new approach to indexing handwriting. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 631–637, 1996.

[10] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[11] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for offline handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[12] P. Moreno, C. Joerg, J. van Thong, and O. Glickman. A recursive algorithm for the forced alignment of very long audio segments. In *5th Int. Conf. on Spoken Language Processing*, 1998.

[13] N. Otsu. A threshold selection method from gray-scale histogram. *IEEE Trans. Systems, Man, and Cybernetics*, 8:62–66, 1978.

[14] T. M. Rath and R. Manmatha. Word spotting for historical documents. *Int. Journal Document Analysis Recognition*, 9(2):139–152, 2007.

[15] J. Rothfeder, R. Manmatha, and T. Rath. Aligning transcripts to automatically segmented handwritten manuscripts. In H. Bunke and A. Spitz, editors, *Proc. 7th Int. Workshop Document Analysis Systems VII*, number 3872 in LNCS, pages 84–95. Springer, 2006.

[16] C. I. Tomai, B. Zhang, and V. Govindaraju. Transcript Mapping for Historic Handwritten Document Images. In *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 413–418, Washington, DC, USA, 2002.

[17] M. Zimmermann and H. Bunke. Automatic segmentation of the IAM off-line database for handwritten English text. In *Proc. 16th Int. Conf. on Pattern Recognition*, volume 4, pages 35–39, 2002.