2011 International Conference on Document Analysis and Recognition

HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents

Andreas Fischer, Emanuel Indermühle, Volkmar Frinken, and Horst Bunke
Institute of Computer Science and Applied Mathematics
University of Bern, Neubrückstrasse 10, 3012 Bern, Switzerland
{*afischer,eindermu,frinken,bunke*}*@iam.unibe.ch*

***Abstract*—For historical documents, available transcriptions typically are inaccurate when compared with the scanned document images. Not only the position of the words and sentences are unknown, but also the correct image transcription may not be matched exactly. An error-tolerant alignment is needed to make the document images amenable to browsing and searching in digital libraries. In this paper, we propose a novel multi-pass alignment method based on Hidden Markov Models (HMM) that combines text line recognition, string alignment, and keyword spotting to cope with word substitutions, deletions, and insertions in the transcription. In a segmentation-free approach, transcriptions of complete pages are aligned with sequences of text line images. On the Parzival data set, results are reported for several degrees of artificial distortions. Both the accuracy and the efficiency of the proposed system are promising for real-world applications.**

***Keywords*-Handwriting recognition; Hidden Markov models**

I. Introduction

Worldwide, there is a huge repository of scanned or photographed valuable old documents including, e.g., Old Greek manuscripts from Early Christianity, Old German manuscripts from the Middle Ages, and important handwritings from the Modern Ages, such as George Washington's papers at the Library of Congress. In order to make the manuscript images amenable to searching and browsing in digital libraries, automatic handwriting recognition is needed for accessing the content of the images [1].

The automatic transcription of handwriting images with a large vocabulary and multiple writing styles is still far from being perfect [2]. However, computer-readable transcriptions are often available for important historical documents. By means of transcription alignment [3], a correspondence between the words in the transcription and the words in the document image can be established in order to make the image browsable and searchable. Besides an integration into digital libraries, transcription alignment allows an accurate and efficient generation of word image data sets that can be used for the training of handwriting recognition systems [4].

In case of a *perfect transcription*, a one-to-one mapping between the words in the image and the words in the transcription exists. Hence, an alignment can be found by segmenting the document image into words. This scenario has been extensively studied in the literature, e.g., based on the generation of several segmentation hypotheses and image matching using Dynamic Time Warping (DTW) [5], and based on Hidden Markov Models (HMM) [4], [6], [7].

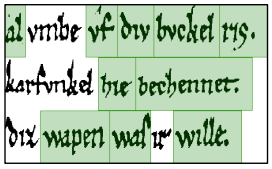
However, available transcriptions for historical documents typically do not match the image content exactly. If a human expert is requested to provide a manual transcription, he or she is often not interested in exactly transcribing each word and each character in the handwriting. Abbreviations are frequently written out in full, mistakes are corrected, and sentences may even be rephrased. In such a real-world scenario, the mapping between transcription and image words is not one-to-one and thus, the alignment becomes more challenging. This problem of *inaccurate transcriptions* has been mentioned, e.g., in [5], but to the knowledge of the authors it has not been studied so far. The basic assumption of current systems to map each word in the transcription with a word image [3] needs to be relaxed to deal with word substitutions, deletions, and insertions in case of an inaccurate transcription.

In this paper, we present an alignment system for inaccurate transcriptions that is based on trained character HMM and performs alignment at the page-level. The input is given by the text line images of a given page together with a computer-readable, inaccurate transcription of the page. The output consists of the location of each word in the image matched with the transcription. In order to cope with errors in the transcription, a multi-pass approach is employed that combines text line recognition, string alignment, and keyword spotting. The procedure is segmentation-free in the sense that no segmentation of the input text line images into words is needed. Neither are line breaks required in the transcription. Because text line recognition is based on the small number of word classes present in the page transcription, it is magnitudes faster than general large vocabulary recognition.

In an experimental evaluation on the historical Parzival data set [8], the proposed system's performance is measured with respect to the newly introduced measure of *alignment accuracy* that takes into account wrong word locations, missing words, and wrongly returned words. For different degrees of artificial transcription distortion, the performance and behavior of the system are analyzed. The reported results are promising for real-world applications.

The remainder of the paper is structured as follows. In Section II, the proposed HMM-based alignment system for

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.20



al trv̊g solhem v̂f div bvckel rîs. hie bechennet. kǒkesas heiden wapen was wille.

Figure 1. Alignment of inaccurate transcriptions. Corresponding words are highlighted in the image and underlined in the transcription. An optimal alignment returns the labels and locations of all highlighted words.

inaccurate transcriptions is described in detail. Next, Section III presents the experimental results, and in Section IV we draw conclusions.

II. HMM-Based Transcription Alignment

In this section, we present an alignment system for inaccurate transcriptions that is based on Hidden Markov Models (HMM) and performs alignment at page-level. The input is given by the text line images of a document page together with a computer-readable, inaccurate transcription of the page. The output consists of word locations in the image for each match with the transcription as illustrated in Figure 1 for part of a Middle High German handwritten text belonging to the Parzival data set [8]. In order to focus on the alignment, we do not take line image extraction errors into account and assume page breaks to be present in the transcription.

The alignment is performed in several steps. After image preprocessing (Section II-A), HMM-based text line recognition is performed in a *first pass* with respect to the page lexicon and language model (Sections II-B and II-C). In a *second pass*, the string edit distance is used to align the recognition output with the transcription (Section II-D). In the resulting recognition gaps (the non-highlighted image parts in Figure 1), lost words are finally recovered in a *third pass* by means of keyword spotting (Section II-E). The performance of the system is measured by the alignment accuracy taking into account wrong word locations, missing words, and wrongly returned words (Section II-F).

A. Preprocessing

For HMM-based recognition, we need to obtain a linear signal from the two-dimensional data. Therefore, we employ the popular sliding window approach to extract a sequence of feature vectors $\mathbf{x} = x_1, \ldots, x_N$ with $x_i \in \mathbb{R}^n$ from the handwriting images.

First, normalized binary text line images are generated. After local enhancement with a Difference of Gaussian (DoG) edge detection operator, the text foreground is retrieved with a global luminosity threshold [8]. As proposed in [9], the handwriting images are then normalized in order to cope with different writing styles. The skew, i.e., the inclination of the text line, is removed by rotation. Next, a vertical scaling procedure is applied to normalize the height with respect to the lower and upper baseline. Finally, horizontal scaling normalizes the width of the text line with respect to the number of black-white transitions.

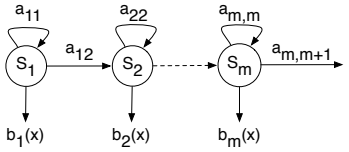


Figure 2. Character HMM

For feature extraction, a window of one pixel width is moved from left to right over the normalized binary text line images. At each of the N positions of the sliding window, $n = 9$ geometrical features are extracted. These features were originally proposed in [9]. Three global features capture the fraction of black pixels, the center of gravity, and the second order moment. The remaining six local features consist of the position of the upper and lower contour, the gradient of the upper and lower contour, the number of black-white transitions, and the fraction of black pixels between the contours.

For more details on text line normalization and sliding window feature extraction, we refer to [8], [9].

B. Hidden Markov Models

The proposed transcription alignment system is based on Hidden Markov Models (HMM) of individual characters. Each character model has a certain number m of hidden states $S_1, \ldots, S_m$ arranged in a linear topology as shown in Figure 2. The states S_i emit observable feature vectors $x \in \mathbb{R}^n$ with output probability distributions $b_i(x)$, each given by a mixture of Gaussians. Starting from the first state S_1, the model either rests in a state or changes to the next state with transition probabilities $a_{i,i}$ and $a_{i,i+1}$, respectively.

The character models are trained using labeled text line images. First, a text line model is created as a sequence of character models according to the labels. Then, the probability of this text line model to emit the observed feature vector sequence $\mathbf{x} = x_1, \ldots, x_N$ is maximized by iteratively adapting the initial output probability distributions $b_i(x)$ and the transition probabilities $a_{i,j}$ with the Baum-Welch algorithm [10]. The trained character models can then be used for text line recognition (Section II-C) and for keyword spotting (Section II-E).

C. First Pass – Text Line Recognition

Although the given page transcription is inaccurate, it provides rich information for handwriting recognition, including

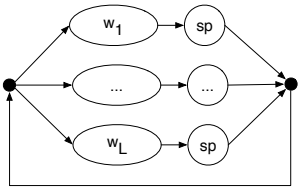


Figure 3. Text Line HMM

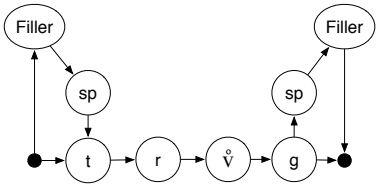


Figure 4. Keyword HMM

the lexicon of possible words and the word relations that can be captured by a language model. This information is used in the first pass of the alignment system that consists of HMM-based text line recognition.

For recognition, the trained character models are first concatenated to words $w_i \in \mathcal{L}$ of the page lexicon $\mathcal{L}$ that contains all word classes occurring in the page transcription. Then, a text line HMM is created that contains a loop of arbitrary length over all lexicon words $w_1, \ldots, w_L$ separated by the space character "sp". An illustration is shown in Figure 3. The optimal word sequence $\mathbf{w}^* = w_{i_1}, \ldots, w_{i_k}$ for the feature vector sequence $\mathbf{x} = x_1, \ldots, x_N$ is given by

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{x}) = \arg\max_{\mathbf{w}} P(\mathbf{x}|\mathbf{w})P(\mathbf{w})$$

with respect to the observation likelihood $P(\mathbf{x}|\mathbf{w})$ and the a priori probability $P(\mathbf{w})$ of the word sequence. The likelihood is given by the HMM probability densities $b_i(x)$ and transition probabilities $a_{i,j}$, while the a priori probability can be expressed with word n-gram language models. In this paper, we use a back-off smoothed bigram language model [9] estimated directly from the given page transcription. That is, the recognition process is closely guided by the word relations among the transcription words.

The optimal word sequence $\mathbf{w}^*$ can be calculated efficiently by means of the Viterbi algorithm [10] in $O(L^2N)$ time with respect to the lexicon size L and the observation sequence length N. As a byproduct of the recognition process, the optimal word locations are returned. In contrast to large vocabulary tasks, where a lexicon size of several ten thousand words is not uncommon, only a small page lexicon is needed. Thus, the text line recognition procedure considered here is magnitudes faster than general text line recognition.

As a variant of single text line recognition, we concatenate the observation sequences of all text lines in order to retain language model information at the line breaks. If a word extends over two text lines after Viterbi decoding it is assigned, in a post-processing step, to the line that covers most of it.

D. Second Pass – String Alignment

The recognition result obtained in the first pass provides a label for each word in the document image. Since we are only interested in the transcription words, an alignment between the recognition result and the transcription is performed in a second pass based on string edit distance [11], taking only insertion and deletion edit operations with cost one into account. Thereby, words are discarded from the recognition result if their labels do not occur at the correct position with respect to the transcription. This can be seen as a second language model constraint imposed on the recognition result that is more strict than the bigram models used in the first pass.

After the second pass, the recognition result does no longer provide word labels for the entire document image. For each discarded word, a gap results in the image (see Figure 1) that reflects a missing word in the transcription. On the other hand, we obtain unassigned words in the transcription for which no correspondence to an image part could be established. These words are assumed not to be present in the image.

E. Third Pass – Keyword Spotting

For each gap in the recognition result after the second pass, the string edit distance alignment provides a list of candidate words from the transcription that could be matched to the gap. For example, the words "trv̊g" and "solhem" are candidates for the first gap in Figure 1. If the candidate list is not empty, it might contain one or several words that are, in fact, present in the image but were not correctly recognized in the first pass, e.g., due to image noise or because of a handwriting style that is not well represented by the character models. Therefore, confidence-based recognition is performed in a third and last pass in order to recover gap words by means of keyword spotting.

For each candidate word, a special keyword HMM is constructed for keyword spotting as recently proposed in [12]. In Figure 4, an example is shown for the word "trv̊g". The keyword is required to be present at the beginning, in the middle, or at the end of the gap observation sequence. The rest of the observation is modeled with a so-called *filler model* that is given by a loop over all characters of the alphabet. Viterbi decoding with respect to the keyword HMM results in the most likely position of the keyword alongside with the keyword gap likelihood $P(\mathbf{x}|K)$. In a second recognition step the filler gap likelihood $P(\mathbf{x}|F)$ is obtained with respect

to the filler model only. The ratio between these likelihoods, normalized with the gap observation length N,

$$\frac{P(\mathbf{x}|K)}{N \cdot P(\mathbf{x}|F)} > T$$

is then used as a confidence measure for keyword spotting. That is, the gap word is added to the alignment system result if the normalized likelihood ratio exceeds a certain threshold T. Note that an overlap of spotting results is allowed. For more details on keyword spotting, we refer to [12].

F. Alignment Accuracy

For evaluating handwriting alignment with inaccurate transcriptions, we introduce the measure of *alignment accuracy* that is very similar to the word accuracy for transcription evaluation. The alignment accuracy is given by

$$Acc = \frac{N - S - D - I}{N}$$

with respect to the number of words N that appear in both the image and the transcription, the number of substitutions S given by wrong word locations, the number of deletions D given by words that were not found in the image, although they are present, and the number of insertions I given by words that were wrongly returned, although they are not present in the image. The values of S, D, and I are found by means of string edit distance alignment with the ground truth taking the word locations into account.

For evaluating the correctness of the word locations, we employ the measure proposed in [7]. Here, the word start and end positions are required to be in within the space character before and after the word. The ground truth is created with HMM-based forced alignment and manual correction [7].

III. Experimental Results

The proposed handwriting alignment system for inaccurate transcriptions is tested on the Parzival data set [8]. This data set includes digital images of a medieval manuscript originating in the 13th century. It contains the epic poem *Parzival* by Wolfram von Eschenbach, one of the most significant epics of the European Middle Ages. The manuscript was written by several monks in the Middle High German language with ink on parchment. 47 pages are considered for experimental evaluation that contain $4,478$ text line images, $4,937$ word classes, and 96 characters.

In order to obtain inaccurate transcriptions in a controlled manner, artificial distortions are applied to the ground truth. For each word in the transcription, an error is added randomly with probability $0 \leq \delta \leq 1$. Possible errors include word substitution, word deletion, and word insertion. For substitution and insertion, out-of-vocabulary words are used. E.g., for a distortion degree of $\delta = 0.5$, half of the words in the transcription are modified, either by substitution, deletion, or insertion.

System	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$
TLR	95.45	94.63	93.46	92.54	91.42
TLR^*	96.29	95.53	94.28	93.46	92.34
KWS	96.64	96.09	95.33	94.60	93.69

Table I
MEAN TEST SET ALIGNMENT ACCURACY. ALL IMPROVEMENTS ARE STATISTICALLY SIGNIFICANT (T-TEST, $\alpha = 0.05$).

System	S	D	I
TLR	121.4	260.2	22.8
TLR^*	127.6	207.2	26.2
KWS	137.4	126.6	33.4

Table II
MEAN TEST SET ERRORS FOR $\delta = 0.5$ ($N = 4,714.8$)

Three variants of the alignment system discussed in Section II are tested and compared. First, standard text line recognition TLR in the first pass, followed by string alignment in the second pass. Secondly, modified text line recognition TLR^* that concatenates all observation sequences in the first pass to retain language model information at the line breaks. And thirdly, the keyword spotting system KWS that includes the final third pass for retrieving words in gaps in addition to the TLR^* system.

A. Setup

First, the $4,478$ text line images and transcriptions are divided into three disjoint sets for training ($\sim 50\%$), validation ($\sim 20\%$), and testing ($\sim 30\%$). Then, in each set, text lines from the same page are collected and used for page-level alignment as described in Section II. Note that the line break information in the transcription is discarded and is not taken into account for alignment. Each of the 47 test pages contains about 28 text lines on average.

For each distortion degree $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, the test set transcription is randomly distorted five times. The reported results are given by the mean alignment accuracy. For the validation set, which is used for parameter optimization, the transcription is distorted only once for each distortion degree δ.

Parameters that are optimized with respect to the validation alignment accuracy include the number of Gaussian mixtures G used for the HMM, the grammar scale factor GSF and word insertion penalty WIP that regulate the language model integration during Viterbi decoding [9], and the threshold T that is used for keyword spotting. The number of HMM states is adopted from previous work [8].

B. Results

The mean alignment accuracy achieved on the test set is given in Table I for the three systems TLR, TLR^*,

and KWS for different transcription distortion degrees δ. The achieved improvements of TLR^* over TLR as well as the improvements of KWS over both other systems are statistically significant for all distortions (t-test, $\alpha = 0.05$).

The error analysis given in Table II shows the behavior of the systems for the distortion degree $\delta = 0.5$. After distortion, $N = 4,714.8$ words are expected to be returned on average together with the correct image location. Taking into account language model information at the line breaks, the TLR^* system returns more words than TLR. By filling recognition gaps by means of keyword spotting, KWS returns even more words. Although these additional words slightly increase the number of substitution errors S (wrong word locations) and insertion errors I (wrongly returned words), the number of deletion errors D (missing words) can be reduced significantly. About 50% of the missing words are retrieved by KWS when compared to TLR.

IV. Conclusions

In this paper, a novel transcription alignment scenario is investigated for handwritten historical documents. In contrast to former studies that take a perfect transcription into account with a label for each word in the document image, the current work considers *inaccurate transcriptions* that do not reflect the image content perfectly. This is a typical scenario for historical documents. An alignment of inaccurate transcriptions allows document indexing for digital libraries and the extraction of training samples for recognition systems.

A multi-pass alignment system based on HMM is presented that combines text line recognition, string alignment, and keyword spotting to align an inaccurate transcription with the document image. The proposed method can be applied at page-level and is segmentation-free in the sense that neither segmentation of text line images into words nor line break information in the transcription are needed. Because the system is based on a limited page lexicon, it is magnitudes faster than general text line recognition.

For measuring the system performance, the *alignment accuracy* is introduced that takes into account wrong word locations, missing words, and word insertions. On the medieval Parzival data set it is demonstrated that the proposed system can achieve an alignment accuracy of 93.69%, even if half of the transcription words are artificially distorted by word substitution, insertion, and deletion. The accuracy is remarkable if the general recognition difficulty of the data set is taken into account. In [8], a text line transcription accuracy of 73.00% was reported for the same data set.

Both the accuracy and the efficiency of the proposed system are promising for real-world applications. Future work includes the application of multiple alignment iterations and the investigation of the amount of training data needed for robust alignment. An alignment of inaccurate transcriptions is furthermore interesting in the context of interactive transcription systems and semi-supervised learning.

Acknowledgments

This work has been supported by the Swiss National Science Foundation (Project CRSI22_125220).

References

[1] A. Antonacopoulos and A. Downton, "Special issue on the analysis of historical documents," *Int. Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 75–77, 2007.

[2] A. L. Koerich, R. Sabourin, and C. Suen, "Large vocabulary off-line handwriting recognition: A survey," *Pattern Analysis and Applications*, vol. 6, pp. 97–121, 2003.

[3] C. Tomai, B. Zhang, and G. Govindaraju, "Transcript mapping for historic handwriting document images," in *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 413–418.

[4] 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*, vol. 4, 2002, pp. 35–39.

[5] E. M. Kornfield, R. Manmatha, and J. Allan, "Further explorations in text alignment with handwritten documents," *Int. Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 39–52, 2007.

[6] J. L. Rothfeder, R. Manmatha, and T. M. Rath, "Aligning transcripts to automatically segmented handwritten manuscripts," in *Proc. 7th Int. Workshop on Document Analysis Systems*, 2006, pp. 84–95.

[7] E. Indermühle, M. Liwicki, and H. Bunke, "Combining alignment results for historical handwritten document analysis," in *Proc. 10th Int. Conf. on Document Analysis and Recognition*, 2009, pp. 1186–1190.

[8] M. Wüthrich, M. Liwicki, A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz, "Language model integration for the recognition of handwritten medieval documents," in *Proc. 10th Int. Conf. on Document Analysis and Recognition*, vol. 1, 2009, pp. 211–215.

[9] 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*, vol. 15, pp. 65–90, 2001.

[10] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[11] R. Wagner and M. Fischer, "The string-to-string correction problem," *Journal of the Association for Computing Machinery*, vol. 21, no. 1, pp. 168–173, 1974.

[12] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM-based word spotting in handwritten documents using subword models," in *Proc. 20th Int. Conf. on Pattern Recognition*, 2010, pp. 3416–3419.