Transcription Alignment of Latin Manuscripts using Hidden Markov Models

Andreas Fischer
Institute of Computer Science and Applied Mathematics
Neubrückstrasse 10
3012 Bern, Switzerland
afischer@iam.unibe.ch

Volkmar Frinken
Institute of Computer Science and Applied Mathematics
Neubrückstrasse 10
3012 Bern, Switzerland
frinken@iam.unibe.ch

Alicia Fornés
Computer Vision Center - Dept. of Computer Science
Universitat Autònoma de Barcelona, Edifici O, 08193
Bellaterra, Spain
afornes@cvc.uab.es

Horst Bunke
Institute of Computer Science and Applied Mathematics
Neubrückstrasse 10
3012 Bern, Switzerland
bunke@iam.unibe.ch

ABSTRACT

Available transcriptions of historical documents are a valuable source for extracting labeled handwriting images that can be used for training recognition systems. In this paper, we introduce the Saint Gall database that includes images as well as the transcription of a Latin manuscript from the 9th century written in Carolingian script. Although the available transcription is of high quality for a human reader, the spelling of the words is not accurate when compared with the handwriting image. Hence, the transcription poses several challenges for alignment regarding, e.g., line breaks, abbreviations, and capitalization. We propose an alignment system based on character Hidden Markov Models that can cope with these challenges and efficiently aligns complete document pages. On the Saint Gall database, we demonstrate that a considerable alignment accuracy can be achieved, even with very weakly trained character models.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*text processing*

1. INTRODUCTION

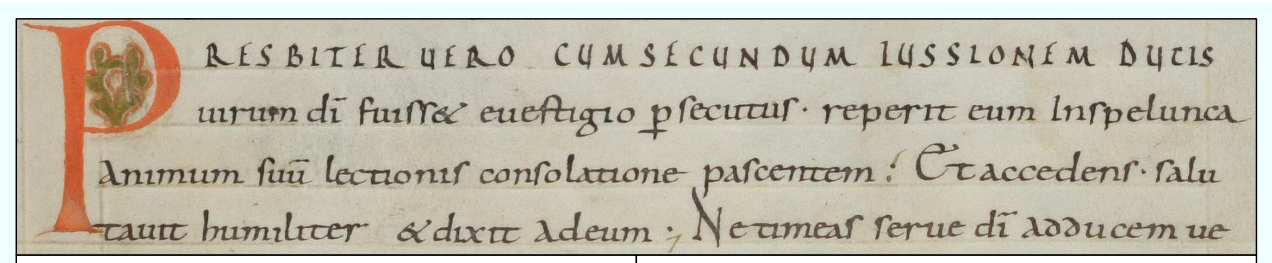
In the context of cultural heritage preservation, many libraries all around the world have digitized a vast amount of handwritten historical documents. However, the scanned or photographed manuscript images often are inaccessible to the world in general because the content of the documents is not available electronically. Hence, the interest in automatic handwriting recognition for historical documents has grown strongly in recent years [1]. The overall aim is to create automatic recognition systems that can replace or at least alleviate a costly manual transcription of the manuscripts.

While highly accurate commercial handwriting recognition systems exist for restricted vocabulary tasks in modern documents, such as bank cheque reading [2], the task of unconstrained handwriting recognition with large vocabularies underlying natural language is still considered unsolved [3]. Besides large vocabularies, there are other major challenges to consider. Unlike for printed documents, a high variability of writing styles has to be taken into account for handwritten documents, even in case of a single writer. Also, while for modern handwriting recognition it is possible to use special input devices to record *on-line* information about the handwriting process [4], this information is not available for *off-line* recognition of digital images of handwritings.

Despite the difficulties mentioned, encouraging reports of relatively high word recognition accuracies, e.g., about 74% for unconstrained modern English handwritings [5], can be found in recent literature. Such a high recognition accuracy could already be very helpful for indexing historical documents in digital libraries or support humans in an interactive transcription [6]. However, the prerequisites to achieve this recognition accuracy are, firstly, the availability of a large amount of training samples and, secondly, a large text corpus to extract language model information that is essential for successful recognition systems [7].

Both prerequisites typically are hard to match in case of historical documents, where vast differences in writing style and language exist for different epochs and geographic locations. Instead, a recognition system often has to be built from scratch for a new kind of historical manuscript. That is, a costly manual transcription is needed to provide training samples before even a weakly trained recognition system is available.

A cheaply accessible source for acquiring training data is



Presbyter vero cum, secundum iussionem ducis, virum Dei fuisset e vestigio prosecutus, reperit eum in spelunca animum suum lectionis consolatione pascentem; et accedens salu- tavit humiliter, et dixit ad eum: Ne timeas, serve Dei, ad ducem ve-	***RESBITER VERO CUM SECUNDUM IUSSIONEM DUCIS*** virum ***di fuiss&*** e vestigio ***psecutus.*** reperit eum in spelunca animum ***suu*** lectionis consolatione pascentem***. Et*** accedens***. salu*** ***tavit*** humiliter ***&*** dixit ad eum***.*** Ne timeas serve ***di*** ad ducem ***ve***

Figure 1: Cod. Sang. 562, p. 25. Original image (top), available transcription (bottom left), actual word spelling in the handwriting image (bottom right). Note that line breaks are not given in the transcription, they are displayed for improving the readability. Word spelling deviations are marked in bold.

given by historical manuscripts for which a transcription already is available. An alignment between the available transcription and the handwriting image can be applied to extract useful training samples automatically. Also, such a transcription alignment provides means to browse electronic manuscript editions at text line or word level, which is more convenient for reading in general, and highly beneficial for paleographic studies in particular.

The traditional use case of transcription alignment is the segmentation of text line images into word images in order to create a word database from an existing text line database [8, 9, 10, 11]. In this use case, a perfect text line transcription typically is assumed, i.e., that the exact character sequence of the text line is given in the transcription. However, this condition often is not satisfied for real-world transcriptions of historical documents. Instead of transcribing each word and each character in the handwriting image, human experts often write out abbreviations, correct mistakes, insert punctuation marks, change the capitalization, or even rephrase whole sentences for better readability of the transcription. All these deviations from the actual word spelling in the handwriting image pose challenges for automatic transcription alignment systems. Also, transcriptions often are given at page or even at document level, which requires alignment algorithms to detect line breaks additionally.

In this paper, we present a new historical handwriting data set that incorporates the aforementioned challenges and propose a system for transcription alignment under these real-world conditions. The introduced Saint Gall database [1] consists of 60 page images of a 9th century Latin manuscript written in Carolingian letters with ink on parchment by a single experienced hand. The images come together with the transcription from a high quality edition, which is aligned at page level. In general, the script is regularly written and only few parchment background artifacts are present. Together with the given transcription, the manuscript is an ideal candidate for automatic extraction of sample images for training a Carolingian handwriting recognition system.

[1]The authors aim at making the database available soon at `http://www.iam.unibe.ch/fki/databases`.

There are several challenges to overcome for transcription alignment. Some examples are shown in Figure 1. First, the line breaks are not given in the transcription. Line break detection is complicated by the fact that no hyphens are used for breaking words at the line end. Secondly, words often are abbreviated [2] and do not correspond with their spelling in the transcription, where they are written out in full. Last but not least, the capitalization and punctuation in the transcription does not match the handwriting image.

The proposed transcription alignment system, which is able to cope with the conditions mentioned above, is based on character Hidden Markov Models (HMM). For alignment, a complete page HMM is created according to the given transcription and is matched against a sequence of analytical features extracted from all text line images of the page. Word spelling variants known from the training set are included in the page model. Also, word deletions are allowed in the model to take into account additional words in the transcription. HMM-based recognition then returns word images as a result together with the assigned word label and spelling. On the Saint Gall database, we demonstrate that a considerable alignment accuracy can be achieved, even with very weakly trained character models.

Note that no spelling rules specific to Carolingian scripts are used in the proposed system. The system is principally designed for the general purpose of aligning manuscripts written in any alphabetic language. However, specific rules can be added in a straight-forward manner to the spelling variants of the individual words. Such additional rules are expected to further increase the accuracy of the system.

The remainder of this paper is structured as follows. In Section 2, the Saint Gall database is introduced. Next, the proposed transcription alignment system is presented in Section 3. Experimental results are then given in Section 4 and some conclusions are finally drawn in Section 5.

[2]Note that parchment was expensive in the Middle Ages and hence the use of abbreviations made perfectly sense in order to save some space.

2. SAINT GALL DATABASE

The newly introduced Saint Gall database consists of digital manuscript images that contain the hagiography *Vita sancti Galli* by Walafrid Strabo. In his work, Strabo describes the life of Saint Gall who gave his name to the Abbey of Saint Gall, Switzerland. The Abbey Library holds a manual copy of the work within the Cod. Sang. 562 that was created at the end of the 9th century. The Latin manuscript is written by a (probably) single experienced hand in Carolingian script with ink on parchment. Carolingian minuscules are predominant, but there are also some upper script letters that emphasize the structure of the text and some richly ornamented initials. Each page is written in a single column that contains 24 text lines. An example page is shown in Figure 2. Altogether, the Saint Gall database includes 60 manuscript pages.

The digital images of the manuscript were made available online by the e-codices project, a virtual manuscript library from the Medieval Institute of the University of Fribourg, Switzerland. [3] The documents were digitized with a resolution of 300dpi and are available as high quality JPEG images. The document images have already been used in [12] in the context of layout modeling. The manuscript transcription was attached at page level to the e-codices images by the affiliated Monumenta project. [4] The transcription is taken from the Patrologia Latina edition. [5]

Both images and transcriptions were downloaded as JPEG and HTML, respectively, and have been processed in several steps into a handwriting recognition database. The resulting ground truth consists of the word positions, word labels, and word spellings in text line images. This information allows for training as well as for evaluating the proposed transcription alignment system (see Section 3).

For creating the ground truth, we followed in large parts the semi-automatic procedure proposed in [13]. The different steps of the ground truth creation are described in the following. In Section 2.1, text line image extraction is discussed. Next, computer-aided transcription alignment is addressed in Section 2.2. Finally, word segmentation is presented in Section 2.3. The resulting ground truth is summarized in Section 2.4.

2.1 Text Line Extraction

In a first step, the main text area is selected manually in the manuscript images using the GIMP [6] software as illustrated in Figure 2. Hereby, the user was asked to select only text regions without ornamented initial letters and without marginal notes that were added later to the manuscript. In some cases, upper case letters written left of the main text column and headings written in upper case letters were left out as well. This resulted in a loss of about two percent of the text line images, which imposes an additional challenge for transcription alignment (see Section 2.4).

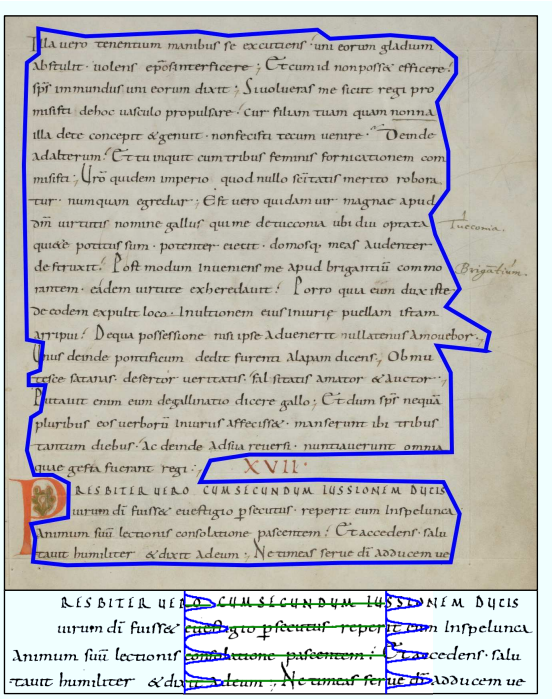


Figure 2: Cod. Sang. 562, p. 25. Text area selection (top), binarization, and skew detection (bottom).

Next, the text foreground is detected automatically using Sauvola's method [14]. Each pixel is assigned to the text foreground if its intensity exceeds a threshold that is dependent on the local pixel intensity distribution. We have used a local window of 20×20 pixels and applied a median filter after binarization in order to remove some remaining noise. Taking into account the text selection, an exemplary result is shown in Figure 2.

Finally, the text column is split into individual lines by means of a semi-automatic procedure originally proposed in [15]. The separating path between two text lines is determined as follows. First, starting points are found at local minima of the horizontal projection profile of the black pixels calculated in the left part of the text image. Next, the skew, i.e., the inclination of the text line, is determined based on two additional histograms that are calculated in the middle of the text line as illustrated in Figure 2. The skew is then given by the slope of the connection between two local maxima. From the starting points, the separating path is then found from left to right in the direction of the skew avoiding text line crossings. For more details on text line segmentation, we refer to [13].

The resulting text line separation is finally corrected manually in a graphical user interface (GUI) displayed in Figure 3. Since the spacing between two text lines is rather large for the Saint Gall database, almost no manual corrections were needed.

[3] http://www.e-codices.unifr.ch
[4] http://www.monumenta.ch
[5] J.-P. Migne PL114, 1852
[6] http://www.gimp.org/

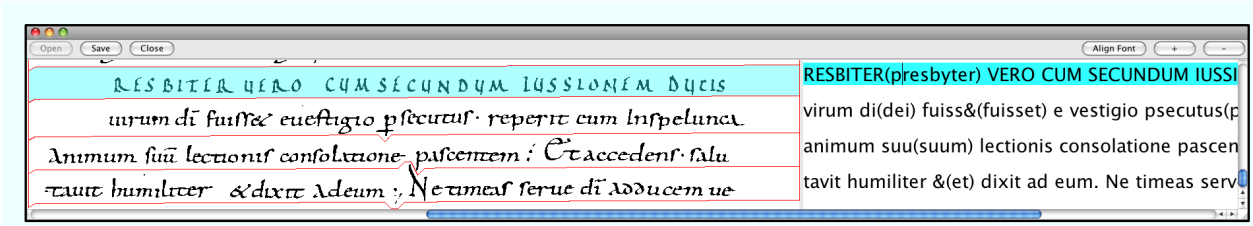


Figure 3: GUI for line segmentation correction (left) and transcription alignment (right).

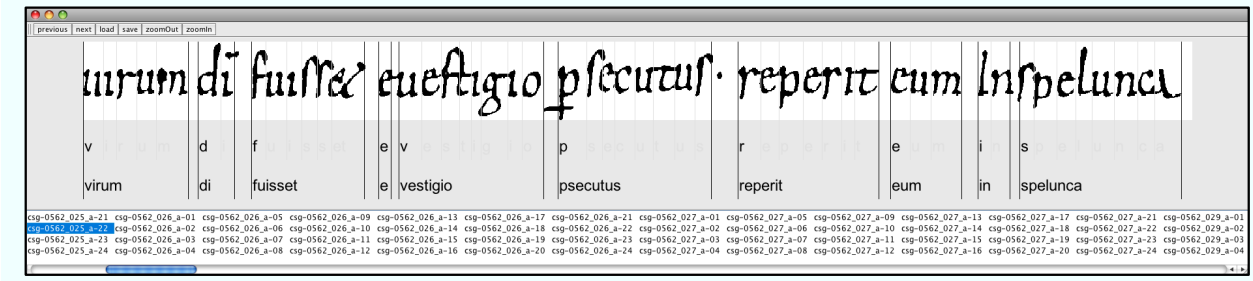


Figure 4: GUI for word segmentation correction.

2.2 Transcription Alignment

In the next step, the correct sequence of characters is assigned to each text line. After parsing the transcription text from the HTML page, the text is added in a text editor to the same GUI that was used for correcting the text line separation (see Section 2.1). A screenshot is shown in Figure 3.

The text is initially inserted in a single line, because the text line breaks in the transcription do not correspond with the image. The user is then requested to insert the correct line breaks manually. The transcription words are displayed in lower case and all punctuation marks are removed, since neither capitalization nor punctuation corresponds with the image. The main task for the user now is to insert the correct word spelling into the transcription. If the spelling deviates from the lower case word label it is typed in manually, followed by the word label in brackets. For example, the second line shown in Figure 1 is encoded as follows:

virum di(dei) fuiss&(fuisset) e vestigio psecutus(prosecutus). reperit eum in spelunca

Besides the standard lower and upper case letters, only one special character was used, i.e., the ampersand "&" for the very frequent abbreviation of *et*. A period is used for any punctuation in the image and a dash to indicate a word break at the line end. Special marks above or below the letters, e.g., the bar below *psecutus* in the example above, were ignored to speed up the alignment process. In future work, they could be taken into account for detecting abbreviations in Carolingian scripts.

Overall, the manual transcription alignment was the most time-consuming part of the ground truth creation process. Fortunately, the Carolingian script is very close to our modern Latin lower case letters. Hence, no expert knowledge in linguistics was needed to perform the alignment.

2.3 Word Segmentation

Once the exact sequence of characters is known for each text line image, it is segmented into individual words by means of a standard forced alignment method using Hidden Markov Models [9]. Hereby, character models can be trained on the complete database and are then used to recognize each text line according to its known character sequence. This procedure will be explained in greater detail in Section 3.

Prior to alignment, the text lines are normalized as suggested in [7]. First, the skew angle is determined by a regression analysis of the bottom-most black pixels of each pixel column. Then, the skew of the text line is removed by rotation. Next, a vertical scaling is applied to obtain three writing zones of the same height, i.e., lower, middle, and upper zone separated by the lower and upper baseline. To determine the lower baseline, the regression result from the skew correction is used, and the upper baseline is found by vertical histogram analysis. Finally, the width of the text line is normalized by scaling based on the number of horizontal black-white transitions in the middle writing zone. No correction of the slant, i.e., the inclination of the letters, is performed, because almost no slant is present in the Carolingian script. An example of a normalized text line image can be seen in Figure 4. For more details on the text line normalization operations, we refer to [7].

After aligning the normalized text line images automatically (see Section 3), the resulting word positions are presented to the user in the GUI shown in Figure 4 for manual correction.

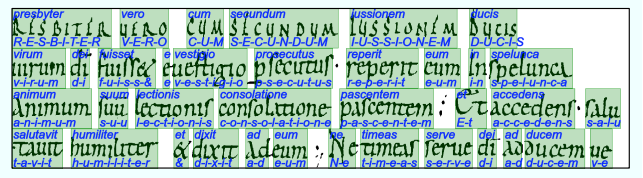


Figure 5: Ground truth consisting of word positions, word labels, and word spellings.

Pages	Lines	Words	Labels	Spellings	Letters
60	1,410	11,602	4,890	5,437	49

Table 1: Saint Gall database statistics.

Hereby, a word boundary is considered correct if the large part of the first and the last character is contained. Note that the punctuation marks are considered to be part of the space between two words and are not displayed in the GUI.

Since the whole database is used for training the alignment system and the exact sequence of characters is known for each text line, the word segmentation was nearly perfect and almost no manual corrections were needed.

2.4 Ground Truth

An example of the resulting ground truth is shown in Figure 5. The ground truth consists of the word positions, labels, and spellings of each text line. In case of a word break at the end of the line, the word label is assigned to the next line. In Table 1, the statistics of the Saint Gall database are summarized. Besides 25 lower case letters and 22 upper case letters, the special letter "&" and the punctuation character "." are used in the database. Note that no punctuation is shown in Figure 5 in order to improve the readability.

3. TRANSCRIPTION ALIGNMENT

In the considered scenario for transcription alignment, the input of the system is given by page transcriptions and text line images of several manuscript pages. As a result, the detected word positions, labels, and spellings are returned for each text line. A perfect alignment system would return the result shown in Figure 5. The page transcriptions are of high quality, i.e., the amount and ordering of the words are near perfect. However, the Saint Gall database pose several challenges for transcription alignment. They are listed in the following.

Line breaks: The line breaks are unknown in the page transcription. Furthermore, 33.90% of the text lines contain a word break at the line end. No hyphen is present in the image at the word breaks.

Abbreviations: Abbreviations are written out in full in the page transcription. Hence, the spelling of the abbreviated words is unknown. Overall, 21.54% of the words are abbreviated.

Capitalization: The capitalization of the words in the page transcription does not correspond with the image, i.e., the capitalization is unknown. The large part of the text is written in Carolingian minuscules, yet 7.06% of the words contain capital letters.

Missing Text Lines: 2.08% of the text line images are missing (see Section 2.1). Because of these errors in the layout analysis, 1.17% additional words are present in the page transcription. Note that there are no line segmentation errors for the remaining lines.

Other Issues: The punctuation marks are unknown in the page transcription. Also, the width of the word spacing shows considerable variations. Figure 5 gives an impression of this issue.

In the following, we present an alignment system based on Hidden Markov Models (HMM) that is able to cope with these challenges. The system is learning-based, i.e., it depends on the ground truth of some training samples. However, it is much less demanding in terms of training material than a system for automatic transcription. In fact, reasonable results can be achieved even when only the first manuscript page is used for training as it will be demonstrated in Section 4. Hence, the proposed alignment system can be used at an early stage in the process of content extraction from historical manuscripts.

The remainder of this section is organized as follows. Feature extraction from normalized text line images is addressed in Section 3.1. Next, character HMMs and their training is described in Section 3.2. Then, in Section 3.3, HMM-based transcription alignment is presented. Finally, measures for evaluating the alignment performance are discussed in Section 3.4.

3.1 Feature Extraction

For HMM-based recognition, the two-dimensional information of the normalized binary images (see Section 2.3) needs to be transformed into a one-dimensional signal. A sequence $\mathbf{x} = x_1, \ldots, x_T$ of feature vectors with $x_i \in \mathbb{R}^n$ is extracted by moving an analysis window with a width of one pixel from left to right over the word image. At each of the T positions of the sliding window, $n = 9$ analytical features are extracted from the foreground pixels. 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 a more detailed description of the features, we refer to [7].

3.2 Hidden Markov Models

The basic modeling unit of the handwritten text is given by

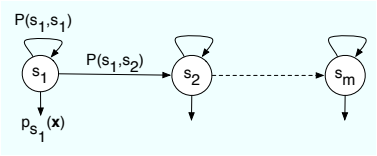


Figure 6: Character HMM

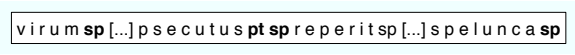


Figure 7: Text Line HMM

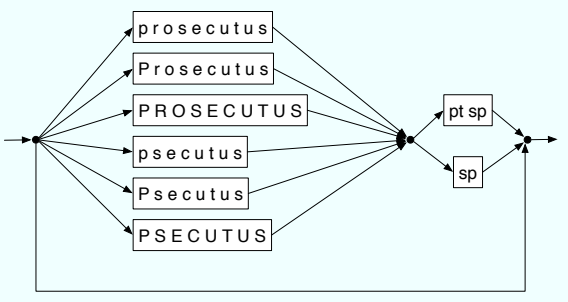


Figure 8: Page HMM

character HMMs shown in Figure 6. Each character model has a certain number m of hidden states $s_1, \ldots, s_m$ arranged in a linear topology. The states s_i emit observable feature vectors $x \in \mathbb{R}^n$ with output probability distributions $p_{s_i}(x)$ given by a Gaussian Mixture Model (GMM). We employ diagonal covariance matrices in order to reduce the number of model parameters that need to be trained. Starting from the first state s_1, the model either rests in a state or changes to the next state with transition probabilities $P(s_i, s_i)$ and $P(s_i, s_{i+1})$, respectively, thus taking into account variable character lengths.

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 transcription as illustrated in Figure 7. The word spacing is either given by the special white space character "sp" or as a combination of the punctuation character "pt" and the space character. Then, the probability of this text line model to emit the observed feature vector sequence $\mathbf{x} = x_1, \ldots, x_T$ is maximized by iteratively adapting the initial output probability distributions $p_{s_i}(x)$ and the transition probabilities $P(s_i, s_i)$ and $P(s_i, s_{i+1})$ with the Baum-Welch algorithm [16].

We also train two special characters "UC" and "lc" that represent general upper case and lower case letters. They can be used during transcription alignment if a letter from the page transcription is unknown. For training, each text line is encoded prior to Baum-Welch training using only these two letter models as well as "sp" and "pt".

3.3 HMM-Based Alignment

If the exact character sequence of a text line is known HMM-based forced alignment can be employed to segment the text line into words [9]. Using the same text line model as for training (see Figure 7), the optimal likelihood $P(\mathbf{x}|\mathbf{c})$ of the feature vector sequence $\mathbf{x}$ for the transcription character sequence $\mathbf{c} = c_1, \ldots, c_N$ is calculated using the Viterbi algorithm [16]. As a byproduct, the optimal word boundaries are returned. This approach was used for creating the ground truth of the Saint Gall database (see Section 2.3).

For aligning page transcriptions, we propose the use of a page HMM that is constructed as follows. First, all letters are converted to lower case and all punctuation marks are removed from the transcription, since neither capitalization nor punctuation are expected to match the manuscript image. Next, several spelling HMMs are created for each word label in the transcription. Besides the lower case character sequence, the first letter can be capitalized and, for headings, the whole word as well. Also, we add known spelling variants from the training set. An example is shown in Figure 8 for the label *prosecutus* and the variant *psecutus* that is assumed to be known. The page HMM is then given by the concatenation of all transcription words according to Figure 8. The word spacing is modeled with and without punctuation using the character models "sp" for white space and "pt" for punctuation. We also allow for word deletion to take into account additional words in the transcription due to missing text line images. If an unknown letter arises in the page transcription it is replaced with the general upper case and lower case characters "UC" and "lc".

For HMM-based recognition, the feature vector sequences of all text lines are first concatenated into a single signal. Then, the optimal sequence of characters is found by means of the Viterbi algorithm with respect to the page HMM. The resulting word boundaries are finally assigned back to the individual text lines based on the signal length of the text lines. If a word extends over two text lines it is considered as a word break at the line end.

Viterbi recognition with the proposed page HMM is very efficient. The computational complexity is given by $O(N^2T)$ with respect to the number of spelling variants N and the feature vector sequence length T. Hence, the alignment is magnitudes faster than automatic transcription with large vocabularies, where N is given by the size of the vocabulary.

3.4 Alignment Performance

The performance of the proposed alignment system is measured with the alignment accuracy

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

with respect to the number of words in the ground truth N, the number of wrong word positions S, word deletions D, and word insertions I. These values are obtained by means of string edit distance [17] between the alignment result and the ground truth. Hereby, only the word labels are taken into account while the recognized word spelling is not required to match the spelling in the ground truth.

Set	%Letters	%Spellings	Train_{lc}	Train_{uc}
T-1	54.2	1.0	28.5	1.0
T-20	97.9	15.4	837.9	31.3

Table 2: Training Sets

System	Acc	R_L	P_L	R_S	P_S
SYS-1	83.37	98.35	84.61	97.88	65.51
SYS-20	92.07	96.40	95.35	95.97	84.81

Table 3: Test Results

Word boundaries are considered correct if they lie within the word spacing area, i.e., the unmarked areas in Figure 5, taking into account a tolerance of 15 pixels, which is about half the width of a character.

For evaluating the ability of the proposed system to return labeled word images for training recognition systems, we also calculate the recall, i.e., the fraction of retrieved word images, and precision, i.e., the correctness of the returned word images. They are given by

$$R = \frac{C}{C+D},\ P = \frac{C}{C+S+I}$$

with respect to the number of correct words $C = N - S - D$. While for some handwriting recognition systems, such as keyword spotting, the correct word label may be sufficient for obtaining training samples, a correct word spelling is needed for training character-based systems. Hence, recall and precision are not only evaluated for position and label correctness, but also for spelling correctness.

4. EXPERIMENTAL EVALUATION

For evaluating the proposed transcription alignment system, the Saint Gall database is split into 20 pages for training, 10 pages for validation, and 30 pages for testing. In addition to the 20 pages training set (T-20), a small training set of only one page, i.e., the first page of the manuscript, is considered as well (T-1). Statistics of the training sets are given in Table 2 including the percentage of known letters and known special spellings with respect to the test set, and the average amount of training samples per lower case letter (Train_{lc}) as well as per upper case letter (Train_{uc}).

Several system variants are compared that differ in the construction of the page HMM (see Section 3.3). The reference system (REF) takes only the lower case letters of each word label into account, then capitalization (CAP), known spelling variants (SPL), and the possibility to delete words (SYS) are added gradually. Figure 9 shows the alignment accuracy results on the test set for different sizes of the training set. While capitalization does not improve the result, adding known spelling variants and the option for deletion both improve the accuracy significantly.

Figure 9: Alignment results for different system configurations.

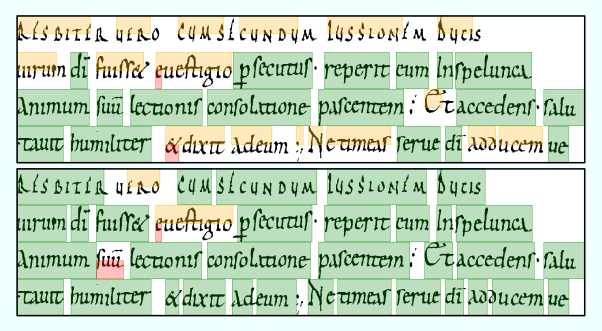


Figure 10: Alignment results for SYS-1 (top) and SYS-20 (bottom).

The optimum number of HMM states and Gaussian mixtures found on the validation set was (5 states, 1 mixture) for T-1 and (15 states, 13 mixtures) for T-20. I.e., not surprisingly, weak character models are preferable for small training sets, and strong models for larger sets. The accuracy of SYS-1 and SYS-20 is listed in Table 3 as well as the word recall and precision with respect to label correctness (R_L, P_L) and spelling correctness (R_S, P_S).

Surprisingly, an alignment accuracy of 83.37 is achieved with the proposed system even if only the first page of the document is used for training the character models. 98.35 percent of the words are retrieved by this system and 84.61 percent of the returned words are correct in terms of word position and label. When the correct spelling of the words is considered, the system's precision drops to 65.51. This is not surprising considering the fact that only 54.2 percent of the letters from the test set are known (see Table 2).

The results of SYS-1 can be improved significantly if 20 pages are used for training. With respect to label correctness, an alignment accuracy of 92.07 is reported. Because more words are deleted during alignment, the recall of 96.40 is smaller than with SYS-1, yet a significant increase in precision up to 95.35 is achieved. Since almost all letters from the test set are known as well as a considerable amount of

special spellings, a high precision of 84.81 is achieved with respect to spelling correctness.

The results demonstrate a great potential of the proposed alignment system to automatically extract new sample images for training handwriting recognition systems. In Figure 10, some alignment results are illustrated for both systems. Position errors are displayed in the upper half of the text line, deletions in the lower half. While word breaks at the line end are handled very accurately throughout the test data, two error cases are predominant. First, since only few samples are available for upper case letters, headings written in upper case pose severe difficulties. Also in case of unknown special spellings, e.g., *suu* for the word *suum* in Figure 10, errors are frequent.

5. CONCLUSIONS

In this paper, we have introduced the Saint Gall database that includes document images as well as the transcription of a Latin manuscript from the 9th century written in Carolingian script. The transcription is taken from a high quality edition, yet there is a gap between the transcription and the handwriting image that is typical for historical documents. E.g., line breaks, abbreviations, capitalization, and punctuation do not match with the handwriting image.

For transcription alignment under these real-world conditions, we propose an HMM-based system that integrates different spellings, capitalization, and punctuation into a page HMM that is able to efficiently perform alignment at page-level. Not taking specific rules for Carolingian scripts into account, the system is principally applicable to any alphabetic language.

On the Saint Gall database it is demonstrated that the proposed system has a great potential to use available transcriptions for extracting new training samples for handwriting recognition systems.

An important challenge for future work is given by the derivation of confidence measures for the returned word images that allow for rejection of incorrect results. The resulting training samples can then be used, e.g., for interactive transcription systems and semi-supervised learning.

6. ACKNOWLEDGMENTS

This work has been partially supported by the Swiss National Science Foundation (CRSI22_125220), by the European project FP7-PEOPLE-2008-IAPP: 230653 as well as by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03, CONSOLIDER-INGENIO 2010 (CSD2007-00018) and the José Castillejo mobility research grant JC2010-0112.

7. REFERENCES

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

[2] N. Gorski, V. Anisimov, E. Augustin, O. Baret, and S. Maximor. Industrial bank check processing: The A2iA check reader. *Int. Journal on Document Analysis and Recognition*, 3:196–206, 2001.

[3] H. Bunke and T. Varga. Off-line Roman cursive handwriting recognition. In B. Chaudhuri, editor, *Digital Document Processing: Major Directions and Recent Advances*, volume 20, pages 165–173. Springer, 2007.

[4] R. Plamondon and S. Srihari. Online and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. PAMI*, 22(1):63–84, 2000.

[5] S. Espana-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez. Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Trans. PAMI*, 33(4):767–779, 2011.

[6] Alejandro H. Toselli, Verónica Romero, Moisés Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825, 2010.

[7] 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.

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

[9] 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.

[10] Jamie L. Rothfeder, R. Manmatha, and Toni M. Rath. Aligning transcripts to automatically segmented handwritten manuscripts. In *Proc. 7th Int. Workshop on Document Analysis Systems*, pages 84–95, 2006.

[11] E. Micah Kornfield, R. Manmatha, and James Allan. Further explorations in text alignment with handwritten documents. *Int. Journal on Document Analysis and Recognition*, 10(1):39–52, 2007.

[12] Micheal Baechler and Rolf Ingold. Medieval manuscript layout model. In *Proc. 10th ACM Symposium on Document Engineering*, pages 275–278, 2010.

[13] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz. Ground truth creation for handwriting recognition in historical documents. In *Proc. 9th Int. Workshop on Document Analysis Systems*, pages 3–10, 2010.

[14] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33:225–236, 2000.

[15] 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.

[16] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[17] R.A. Wagner and M.J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173, 1974.