Alignment between Text Images and their Transcripts for Handwritten Documents

Alejandro H. Toselli, Verónica Romero and Enrique Vidal

Abstract An alignment method based on the Viterbi algorithm is proposed to find mappings between word images of a given handwritten document and their respective (ASCII) words in its transcription. The approach takes advantage of the underlying segmentation made by Viterbi decoding in handwritten text recognition based on Hidden Markov Models (HMMs). Two levels of alignments are considered: the traditional one at word level and the one at text-line level where pages are transcribed without line break synchronization. According to various metrics used to measure the quality of the alignments, satisfactory results are obtained. Furthermore, the presented alignment approach is tested on two different HMMs modelling schemes: one using 78 HMMs (one HMM per character class) and other using two HMMs (for blank space and no-blank characters respectively).

Key words: handwriting image and transcription alignments, forced recognition, digital libraries, handwritten text recognition, Viterbi algorithm

Dr. Alejandro H. Toselli
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: ahector@iti.upv.es

Dr. Verónica Romero
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: vromero@iti.upv.es

Dr. Enrique Vidal
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: evidal@iti.upv.es

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_2, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

Lately, many on-line digital libraries have been publishing large quantities of digitized handwritten documents, which allows both scholars and the general public to access this kind of cultural heritage resources. This is a new, comfortable way of consulting and querying this material. The *Biblioteca Valenciana Digital* (BiValDi)[1] is an example of such digital libraries, which provides an interesting collection of handwritten documents.

Many of these handwritten documents include both, the handwritten material and its proper transcription (in ASCII format for example). Generally speaking, most documents have transcriptions aligned only at the page level, but not at individual text lines, making it difficult the visualization and consulting of these documents for the paleography experts. In Fig. 1 an example of an original piece of manuscript (left) and its corresponding transcription without any kind of alignment (center) is shown.

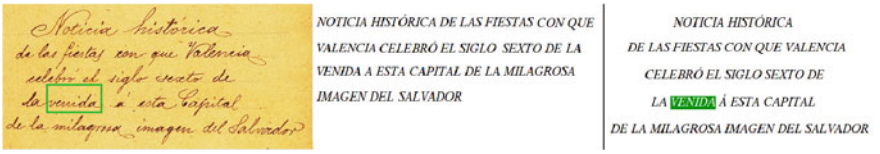

Fig. 1 Left: original image. Center: a corresponding unaligned transcription. Right: the line and word-alignments have been computed. The lines in the manuscript have the same words that the lines at the transcription and an example of a word-alignment is shown with an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word (in reverse video) in the transcript (right)

This fact has motivated the development of techniques to align these documents and their transcripts; i.e. to generate a mapping between each line or word image on a document page with its respective line or word on its electronic transcript. This kind of alignment can help readers to quickly locate image text while reading a transcript, with useful applications to editing, indexing, etc. In the opposite direction, the alignment can also be useful for people trying to read the image text directly, when arriving to complex or damaged parts of the document.

Two different levels of alignment can be defined: line level and word level as shown in Fig 1 (right). Line alignments attempt to obtain beginning and end positions of lines in transcribed pages that do not have synchronized line breaks. This information allows users to easily visualize the page image documents and their corresponding transcriptions. Moreover, using these alignments as segmentation ground truth, large amounts of training and test data for segmentation-free cursive handwriting recognition systems become available. On the other hand, word alignments allow users to easily find the place of a word in the manuscript when reading the corresponding transcript. On a graphical interface properly designed to show word alignments, for example, one can display both the handwritten page

[1] http://bv2.gva.es

image and the transcript and, whenever the mouse is held over a word in the transcript, the corresponding word in the handwritten image would be outlined using a box. In a similar way, whenever the mouse is held over a word in the handwritten image, the corresponding word in the transcript would be highlighted (see Fig. 2).

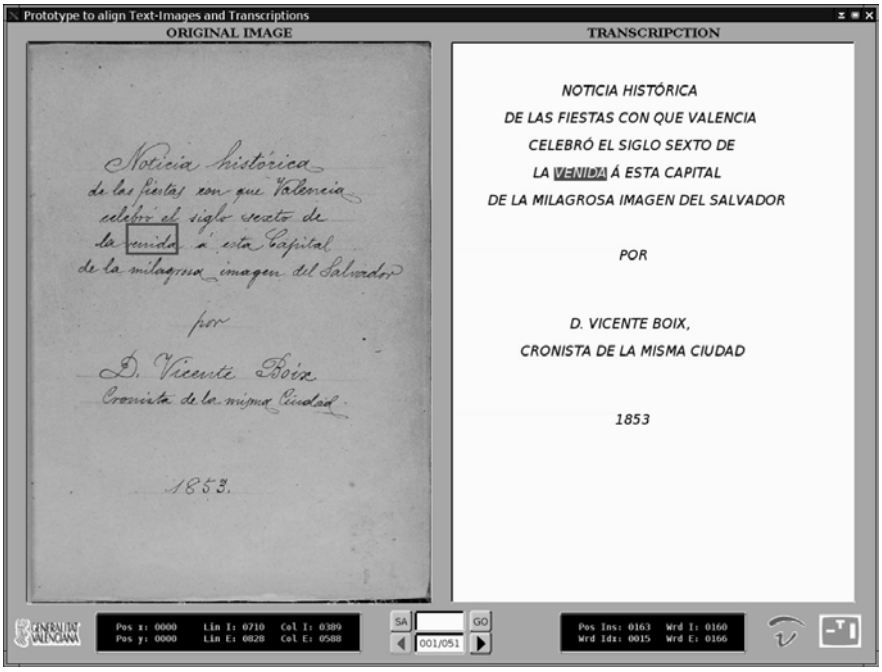

Fig. 2 Screen-shot of the alignment prototype interface displaying an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word in the transcript (right)

It is worth noting that word alignments can be displayed without need of line alignments. However, the visualization of the manuscript and its transcription with both alignments is friendlier for the readers than using only the word alignments.

Creating such alignments is challenging since the transcript is an ASCII text file while the manuscript page is an image. In the case of word alignments, some recent works address this problem by relying on a previous explicit image-processing based word pre-segmentation of the page image, before attempting the transcription alignments. For example, in [6], the set of previously segmented word images and their corresponding transcriptions are transformed into two different times series, which are aligned using *dynamic time warping* (DTW). In this same direction, [3], in addition to the word pre-segmentation, a (rough) recognition of the word images is attempted. The resulting word string is then aligned with the transcription using dynamic programming.

The alignment method presented here (henceforward called Viterbi alignment), relies on the Viterbi decoding approach to handwritten text recognition (HTR)

based on Hidden Markov Models (HMMs) [2, 9]. These techniques are based on methods originally introduced for speech recognition [4]. In such HTR systems, the alignment is actually a byproduct of the proper recognition process, i.e. an implicit segmentation of each text image line is obtained, where each segment successively corresponds to one recognized word. In our case, word recognition is not actually needed, as we do already have the correct transcription. Therefore, to obtain the segmentations for the *given* word sequences, the so-called "forced-recognition" approach is employed (see Sect. 2.2). This idea has been previously explored in [13].

As it has been explained previously, line alignments only make sense in document transcriptions where the beginning and end positions of the image lines are not registered. However, the word alignments can be computed both for line transcriptions or for page transcriptions. In this work, line and word alignment results are reported for a set of 53 pages from a XIX century handwritten document (see Sect. 5.2). To evaluate the quality of the obtained alignments, several metrics were used which give information basically at two different alignment levels: the accuracy of alignment mark placements and the the amount of erroneous assignments produced between word images and transcriptions (see Sect. 4).

The remainder of this paper is organized as follows. First, the alignment framework is introduced and formalized in Sect. 2. Then, an implemented prototype is described in Sect. 3. The alignment evaluation metrics are presented in Sect. 4. The experiments and results are commented in Sect. 5. Finally, some conclusions are drawn in Sect. 6.

2 HMM-based HTR and Viterbi Alignment

HMM-based handwritten text recognition is briefly outlined in this section, followed by a more detailed presentation of the Viterbi alignment approach.

2.1 HMM HTR Basics

The traditional handwritten text recognition problem can be formulated as the problem of finding a most likely word sequence $\hat{\mathbf{w}} = \langle w_1, w_2, \ldots, w_n \rangle$, for a given handwritten sentence (or line) image represented by a feature vector sequence $\mathbf{x} = x_1^p = \langle x_1, x_2, \ldots, x_p \rangle$, that is:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \Pr(\mathbf{w}|\mathbf{x}) \\ &= \arg\max_{\mathbf{w}} \Pr(\mathbf{x}|\mathbf{w}) \cdot \Pr(\mathbf{w}) \end{aligned} \tag{1}$$

where $\Pr(\mathbf{x}|\mathbf{w})$ is usually approximated by concatenated character Hidden Markov Models (HMMs) [2,4], whereas $\Pr(\mathbf{w})$ is approximated typically by an n-gram word language model [4].

Thus, each character class is modeled by a continuous density left-to-right HMM, characterized by a set of states and a Gaussian mixture per state. The Gaussian mixture serves as a probabilistic law to model the emission of feature vectors by each HMM state. Figure 3 shows an example of how a HMM models a feature vector sequence corresponding to character "a". The process to obtain feature vector sequences from text images as well as the training of HMMs are explained in Sect. 3.

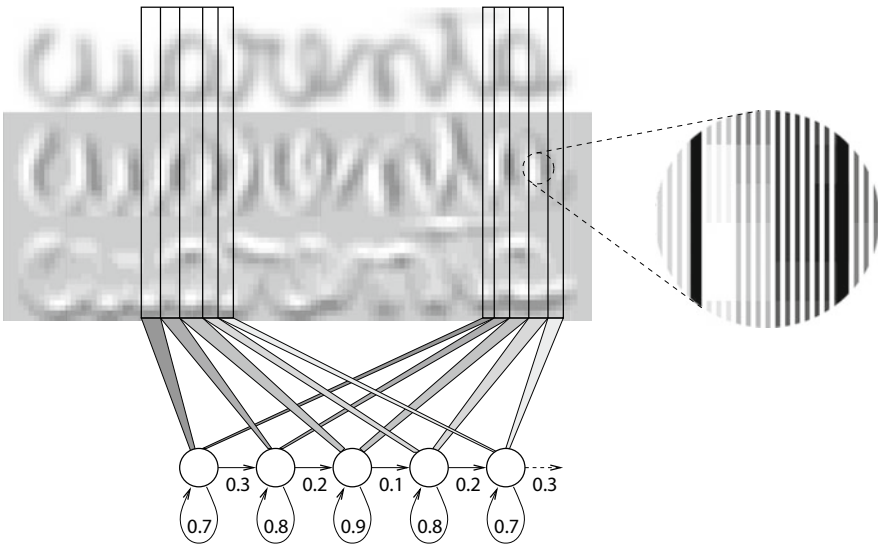

Fig. 3 Example of 5-states HMM modeling (feature vectors sequences) of instances of the character "a" within the Spanish word "cuarenta" (forty). The states are shared among all instances of characters of the same class. The zones modelled by each state show graphically subsequences of feature vectors (see details in the magnifying-glass view) compounded by stacking the normalized grey level and its both derivatives features

HMMs as well as n-grams models can be represented by stochastic finite state networks (SFN), which are integrated into a single global SFN by replacing each word character of the n-gram model by the corresponding HMM. The search involved in the Equ. (1) to decode the input feature vectors sequence $\mathbf{x}$ into the more likely output word sequence $\hat{\mathbf{w}}$, is performed over this global SFN. This search problem is adequately solved by the Viterbi algorithm [4].

2.2 Viterbi Alignment

As a byproduct of the Viterbi solution to (1), the feature vector subsequences of $\mathbf{x}$ aligned with each of the recognized words $w_1, w_2, \ldots, w_n$ can be obtained. This subsequence alignment is implicit or "hidden" in the Equ. (1), which can thus be rewritten as:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w}) \cdot \Pr(\mathbf{w}) \tag{2}$$

where $\mathbf{b}$ is an *alignment*; that is, an ordered sequence of $n+1$ marks $\langle b_0, b_1, \ldots, b_n \rangle$, used to demarcate the subsequences belonging to each recognized word. The marks b_0 and b_n always point out to the first and last components of $\mathbf{x}$ (see Fig. 4).

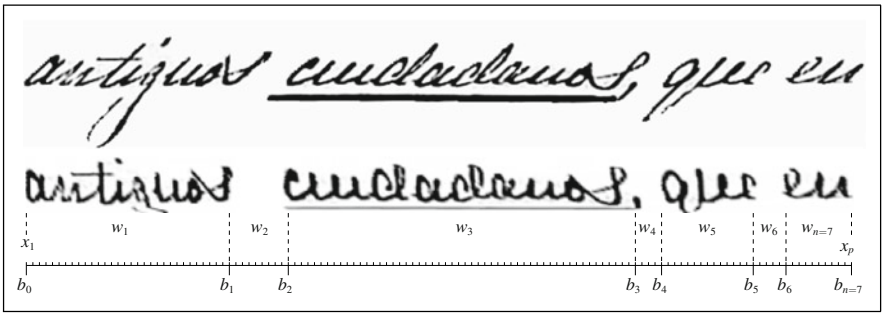

Fig. 4 Example of segmented text line image along with its resulting deslanted and size-normalized image. Moreover, the alignment marks ($b_0 \ldots b_8$) which delimit each of the words (including word-spaces) over the text image feature vectors sequence $\mathbf{x}$.

Now, approximating the sum in (2) by the dominant term, $\max_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w})$:

$$(\hat{\mathbf{b}}, \hat{\mathbf{w}}) \approx \arg\max_{\mathbf{b},\mathbf{w}} \Pr(\mathbf{w}) \cdot \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w}) \tag{3}$$

where $\hat{\mathbf{b}}$ is the optimal alignment. In our case, we are not really interested in text recognition proper, because the transcription is known beforehand. Let $\tilde{\mathbf{w}}$ be the given transcription. Now, $\Pr(\mathbf{w})$ in Equ. (3) is zero for all $\mathbf{w}$ except $\tilde{\mathbf{w}}$, for which $\Pr(\tilde{\mathbf{w}}) = 1$. Therefore,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \tilde{\mathbf{w}}) = \Pr(x_{b_0}^{b_1}, x_{b_1}^{b_2}, \ldots, x_{b_{n-1}}^{b_n} \mid \tilde{\mathbf{w}}) \tag{4}$$

which can be expanded to,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \Pr(x_{b_0}^{b_1} \mid \tilde{\mathbf{w}}) \Pr(x_{b_1}^{b_2} \mid x_{b_0}^{b_1}, \tilde{\mathbf{w}}) \ldots \Pr(x_{b_{n-1}}^{b_n} \mid x_{b_0}^{b_{n-1}}, \tilde{\mathbf{w}}) \tag{5}$$

Assuming that each subsequence $x_{b_{i-1}}^{b_i}$ is independent from that it its predecessors and also depends only of the ith word of $\tilde{w}_i$, Equ. (5) can be rewritten as,

$$\hat{\mathbf{b}} \approx \arg\max_{\mathbf{b}} \Pr(x_{b_0}^{b_1}|\tilde{w}_1) \ldots \Pr(x_{b_{n-1}}^{b_n}|\tilde{w}_n) \tag{6}$$

This problem is optimally solved by using Viterbi search, and this is known as "forced recognition".

2.3 Word and Line Alignments

The word alignments are obtained directly from Equ. (6). If the transcription is given at the line level, the input feature vector sequence $\mathbf{x}$ represents a handwritten text line image and $\tilde{\mathbf{w}}$ its corresponding line transcription. If the transcription is given only at page level, the input is a very long feature vector sequence that represents the whole page. This sequence is obtained by concatenating the feature vector sequences of all the successive text line images of the document page. Accordingly, $\tilde{\mathbf{w}}$ in this case is considered a single, correspondingly long word sequence, without line breaks.

Obviously, when the feature vector sequences of the different text line images are concatenated into the whole-page sequence, we know the positions in $\mathbf{x}$ that identify the joints between lines. Using this information the line alignments are easily computed. Let $\mathbf{l} = \langle l_1, \ldots, l_M \rangle$ be this sequence of positions, where M is the number of lines in the page, and let $x_{b_{i-1}}^{b_i}$ be the feature vector subsequence belonging to the word w_i. This word is considered to belong to the line j if b_{i-1} and b_i are between l_{j-1} and l_j. Sometimes, it may happen that word boundaries (b_{i-1} and b_i) are in two different lines; b_{i-1} is between l_{j-1} and l_j, but b_i is between l_j and l_{j+1}. In this case if $l_j - b_{i-1} \geqslant b_i - l_j$ the word w_i is considered to belong to the line j; otherwise it is considered to belong to the line $j+1$.

This way, once the word alignment has been computed for the whole page, the line alignment is obtained by visiting sequentially each word in the transcript and deciding which line it belongs to.

3 Overview of the Alignment Prototype

The implementation of the alignment prototype involved four different parts: document image preprocessing, line image feature extraction, HMMs training and alignment map generation.

Document image preprocessing encompasses the following steps: first, skew correction is carried out on each document page image; then background removal and noise reduction is performed by applying a bi-dimensional median filter [5] on the whole page image. Next, a text line extraction process based on local minimums of the horizontal projection profile of page image, divides the page into separate line images [7]. In addition connected components has been used to solve the situations

where local minimum values are greater than zero, making it impossible to obtain a clear text line separation. Finally, slant correction and non-linear size normalization are applied [8, 9] on each extracted line image. An example of extracted text line image is shown in the top panel of Fig. 4, along with the resulting deslanted and size-normalized image. Note how non-linear normalization leads to reduced sizes of ascenders and descenders, as well as to a thiner underline of the word "ciudadanos".

As our alignment prototype is based on Hidden Markov Models (HMMs), each preprocessed line image is represented as a sequence of feature vectors. To do this, the feature extraction module applies a grid to divide each line image into $N \times M$ squared cells. In this work, N=40 is chosen empirically (using the corpus described further on) and M must satisfy the condition $M/N = \textit{original image aspect ratio}$. From each cell, three features are calculated: normalized gray level, horizontal gray level derivative and vertical gray level derivative. The way these three features are determined is described in [9]. Columns of cells or *frames* are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in its constituent cells. Hence, at the end of this process, a sequence of M 120-dimensional feature vectors (40 normalized gray-level components, 40 horizontal and 40 vertical derivatives components) is obtained. An example of feature vectors sequence, representing an image of the Spanish word *"cuarenta"* (forty) is shown in Fig. 3.

As it was explained in Sect. 2.1, characters are modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. This topology (number of HMM states and Gaussian densities per state) was determined by tuning empirically the system on the corpus described in Sect. 5.1. Once a HMM "*topology*" has been adopted, the model parameters can be easily trained from images of continuously handwritten text (*without any kind of segmentation*) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called *forward-backward or Baum-Welch re-estimation* [4].

The last phase in the alignment process is the generation of the mapping proper by means of Viterbi "forced recognition", as discussed in Sect. 2.2.

4 Alignment Evaluation Metrics

Several kinds of measures have been adopted to evaluate the quality of alignments at line and at word level.

On the one hand, line alignments are evaluated by means of the line error rate (LER), the average number of words assigned to erroneous lines (AEW) and the maximum number of erroneous words by line (MWE). The LER measures the number of system proposed lines that have different word count than their corresponding reference line-synchronized transcriptions, divided by the total number of lines. The AEW is the number of words that have been assigned to an incorrect line, divided by the total number of words. The MWE, finally, measures the maximum

difference between the word count of the proposed line and that of the correct line-synchronized transcription. Figure 5 shows an example of a line alignment for the image shown in the Fig. 1. The LER, in this example, would be 40%, the AEW 4% and the MWE to 1. A perfect line alignment would have LER=AEW=MWE=0.

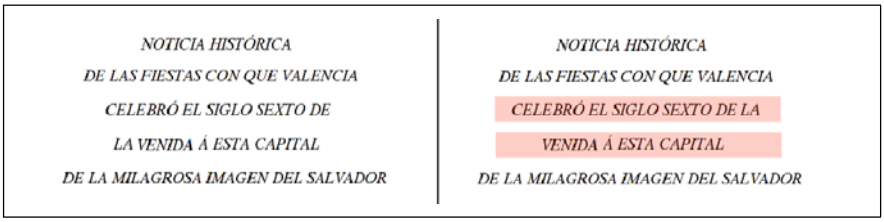

Fig. 5 Example of LER, WAE and MWE computation for the image show in the Fig. 1. On the left we can see the correct lines and in the right the system proposed line alignments. In this case 2 lines have been erroneously aligned (the highlighted lines). The word "LA" has been assigned to the line 3 instead of the line 4. The resulting LER is 40%, the AEW is 4% and the MWE is 1.

On the other hand, quality of word alignments are measured by the alignment error rate (AER) and the average value and standard deviation (henceforward called MEAN-STD) of the absolute differences between the system-proposed word alignment marks and their corresponding (correct) references. The MEAN-STD gives us an idea of the geometrical accuracy of the word alignments obtained, whereas the AER measures the amount of totally erroneous assignments produced between word images and transcriptions.

Given a reference mark sequence $\mathbf{r} = \langle r_0, r_1, \ldots, r_n \rangle$, along with an associated word token sequence $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$, and an automatic segmentation mark sequence $\mathbf{b} = \langle b_0, b_1, \ldots, b_n \rangle$ (with $r_0 = b_0$, $r_n = b_n$), we define the MEAN-STD and AER metrics as follows:

MEAN-STD: The average value and standard deviation of absolute differences between reference and proposed alignment marks, are given by:

$$\mu = \frac{\sum_{i=1}^{n-1} d_i}{n-1} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n-1} (d_i - \mu)^2}{n-1}} \qquad \text{where} \quad d_i = |r_i - b_i| \tag{7}$$

AER: Defined as:

$$\text{AER}(\%) = \frac{100}{N} \sum_{j:w_j \neq s} e_j \qquad \text{with} \quad e_j = \begin{cases} 0 & b_{j-1} < m_j < b_j \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

where s stands for the blank-space token, $N < n$ is the number of real words (i.e., tokens which are not s) and $m_j = (r_{j-1} + r_j)/2$.

A good word alignment would have a μ value close to 0 and small σ. Thus, MEAN-STD gives us an idea of how accurate are the automatically computed word alignment marks. On the other hand, AER assesses word alignments at a

higher level; that is, it measures mismatches between word-images and ASCII transcriptions (tokens), excluding word-space tokens. This is illustrated in Fig. 6, where the AER would be 25%.

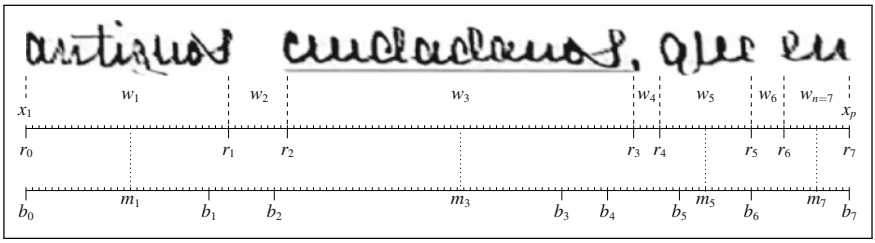

Fig. 6 Example of AER computation. In this case $N=4$ (only no word-space are considered: w_1, w_3, w_5, w_7) and w_5 is erroneously aligned with the subsequence $x_{b_5}^{b_6}$ ($m_5 \notin (b_4, b_5)$). The resulting AER is 25%.

5 Experiments

In order to test the effectiveness of the presented alignment approach, different experiments were carried out. The corpus used, as well as the experimental setup carried out and the obtained results, are explained in the following subsections.

5.1 Corpus Description

The corpus was compiled from the legacy handwriting document identified as *Cristo-Salvador*, which was kindly provided by the *Biblioteca Valenciana Digital* (BiValDi). It is composed of 53 text page images, scanned at 300dpi and written by only one writer. Some of these page images are shown in the Fig. 7.

As has been explained in Sect. 3, the page images have been preprocessed and divided into lines, resulting in a data-set of 1,172 text line images. In this phase, around 4% of the automatically extracted line-separation marks were manually corrected. All the page transcriptions are available with synchronized line breaks containing 10,911 running words with a vocabulary of 3,408 different words.

To test the quality of the computed alignments, 12 pages were randomly chosen from the whole corpus to be used as references. For these pages the true locations of alignment marks were set manually. Table 1 summarized the basic statistics of this corpus and its reference pages.

Fig. 7 Example of page images of the corpus "Cristo-Salvador" (CS), which show backgrounds of big variations and uneven illumination, spots due to the humidity, marks resulting from the ink that goes through the paper (called bleed-through), etc.

Table 1 Basic statistics of the CS corpus (book)

Number of:	References	Total	Lexicon
pages	12	53	–
text lines	312	1,172	–
words	2,955	10,911	3,408
characters	16,893	62,159	78

5.2 Experiments and Results

Since only forced-recognition is needed to obtain line-level word alignments, training and test are carried out on identical data; namely, the whole set of document lines. In principle, a similar process could be applied for the page-level alignments; that is, training the HMMs on the long, whole-page feature vector sequences and the corresponding page transcriptions. However, this process has proved not to be straightforward. On the one hand, training accuracy tends to degrade with the length of the training sequences. On the other hand, training time becomes prohibitive, since the Baum-Welch computing time is proportional to the length of the feature vector sequence times the number of words in the transcription. Therefore, to obtain page-level alignments, HMM models are trained on line data, as in the case of the line-level alignments, but using only a part of the document that has been marked with synchronized line-breaks.

In our experiments, the 41 non-reference pages (860 lines) have been used for training and the test has been carried out with the 12 reference pages. Table 2 shows the line alignment results. The results obtained clearly suggest that, for large documents with hundreds of pages, it would be quite profitable to devote some work

to line-synchronize a (relatively small) part of the document by hand and let the rest be automatically aligned by the system.

Table 2 Line alignment evaluation results: *Line Error Rate* (LER), *Average Number of Words Assigned to Erroneous Lines* (AEW) and *Maximum number of Erroneous Words by line* (MWE).

LER %	MEW	AEW %
6.4	1	0.3

Word alignments can be computed, in the CS corpus, using either line-synchronized transcriptions or whole page transcriptions. In the first case, HMMs were trained with all the 1,172 lines (53 pages) of the corpus and the test has been carried out with the 312 reference lines (12 pages). In the second case, the same HMM models trained for the page-level experiments have been used. Additionally, for the first case (word-alignments at line level), two different HMM modeling schemes were employed: one modeling each of the 78 character classes using a different HMM per class, or other modeling separately the blank "character" class and all the 77 no-blank character classes using two HMMs respectively. Whichever the case, the HMM topology was identical for all HMMs in both schemes: left-to-right with 6 states and 64 Gaussian mixture components per state. Table 3 reports the results of the quality of the obtained word alignments at line level (for the both above-mentioned HMM modeling schemes) and at page level.

Table 3 Word alignment evaluation results at line and page levels: *Alignment Error Rate* (AER) and *Average Value and Standard Deviation* of the absolute differences between the system-proposed word alignment marks and their corresponding (correct) references (MEAN-STD).

	Line Level		Page Level
	78-HMMs	2-HMMs	
AER (%)	7.20	25.98	7.88
μ (mm)	1.14	2.95	1.15
σ (mm)	3.90	6.56	3.43

Several comments about these results are in order. Firth, the system-proposed line alignments are quite acceptable. The results show that from every 100 automatically aligned lines only 6 need to be corrected by the user. Moreover, the error of each incorrect line is due to one word at most. It is worth noting that although the number of erroneous lines is 6%, the number of incorrect system-proposed line breaks is 3%, because each incorrect line break involves two erroneous lines. Overall, the number of words that have been assigned to an erroneous line is just 0.3% of the total number of words.

Furthermore, from the word alignment results using the 78 HMMs scheme, we can see that computing the alignments at line level the best AER is obtained (7.20%). Moreover, the relative low values of μ and σ show that the obtained alignment

marks are quite accurate; that is, they are very close to their respective references. This is illustrated on the left histogram of Fig. 8. The two typical alignment errors

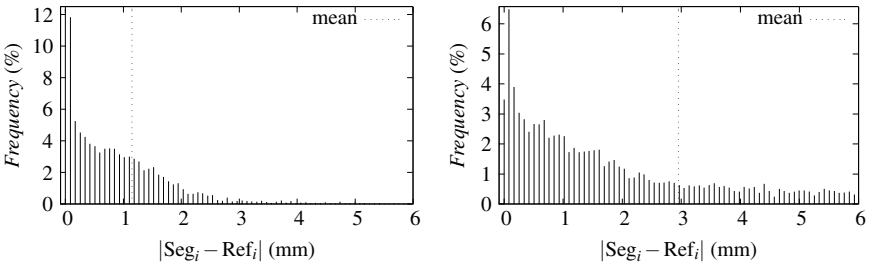

Fig. 8 $|r_i - b_i|$ distribution histograms for 78-HMMs (left) and 2-HMMs (right) modelling schemes.

are known as over-segmentation and under-segmentation respectively. The over-segmentation error is when one word image is separated into two or more fragments. The under-segmentation error occurs when two or more images are grouped together and returned as one word. Figure 9 shows some of them.

Fig. 9 Word alignment for 6 lines of a particularly noisy part of the corpus. The four last words on the second line as well as the last line illustrate some of over-segmentation and under-segmentation error types.

If the word alignments are computed at the page level the results are quite acceptable as well, showing that from every 100 automatically aligned words only 8 are misaligned.

6 Remarks, Conclusions and Future Work

Given a manuscript and its transcription, we propose an alignment method to map every line or word image on the manuscript with its respective line or word on the electronic (ASCII or PDF) transcript. This method takes advantage of the implicit

alignments made by Viterbi decoding used in forced text image recognition with HMMs.

Experiments have been carried out with the CS corpus, which is a legacy handwritten document written in 1853. Despite the difficulty that entails the task, the results achieved in this work are encouraging.

In future works, we plan to carry out the training of the different HMMs models using whole pages, trying to solve the previously explained problems that this training entails. In the experiments carried out in this work, the different HMM models have been trained using text line images, but this way to perform the training is not viable for documents that do not have, at least, a part of their transcriptions synchronized at this line level. Other interesting issue is to follow the new interactive framework presented in [12]. This framework integrates the human activity into the recognition process taking advantage of the user's feedback. This idea has been previously applied to computer assisted translation (CAT) [1], Computer assisted speech transcription (CATS) [12] and computer assisted transcription of handwritten text images (CATTI) [10, 11] with good result.

Acknowledgements Work supported by the EC (FEDER), the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme (CSD2007-00018) and the Spanisg Government (MICINN and "Plan E") under the MITTRAL (TIN2009-14633-C03-01) research project.

References

1. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Ney, A.L.H., Tomás, J., Vidal, E.: Statistical approaches to computer-assisted translation. Computational Linguistics p. In press (2008)
2. Bazzi, I., Schwartz, R., Makhoul, J.: An Omnifont Open-Vocabulary OCR System for English and Arabic. IEEE Trans. on PAMI **21**(6), 495–504 (1999)
3. Huang, C., Srihari, S.N.: Mapping Transcripts to Handwritten Text. In: S. Ltd. (ed.) Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 15–20. La Baule, France (2006)
4. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
5. Kavallieratou, E., Stamatatos, E.: Improving the quality of degraded document images. In: DIAL '06: Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp. 340–349. IEEE Computer Society, Washington DC, USA (2006). DOI http://dx.doi.org/10.1109/DIAL.2006.23
6. Kornfield, E.M., Manmatha, R., Allan, J.: Text Alignment with Handwritten Documents. In: First International Workshop on Document Image Analysis for Libraries (DIAL), pp. 195–209. Palo Alto, CA, USA (2004)
7. Marti, U.V., Bunke, H.: Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. Int. Journal of Pattern Recognition and Artificial In telligence **15**(1), 65–90 (2001)
8. Romero, V., Pastor, M., Toselli, A.H., Vidal, E.: Criteria for handwritten off-line text size normalization. In: Procc. of The Sixth IASTED Int. Conf. on Visualization, Imaging, and Image Processing (VIIP 06). Palma de Mallorca, Spain (2006)
9. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence **18**(4), 519–539 (2004)

10. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. Pattern Recognition **43**(5), 1814–1825 (2009)
11. Toselli, A.H., Romero, V., Rodríguez, L., Vidal, E.: Computer Assisted Transcription of Handwritten Text. In: 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), pp. 944–948. IEEE Computer Society, Curitiba, Paraná (Brazil) (2007)
12. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive pattern recognition. In: Proc. of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, *LNCS*, vol. 4892, pp. 60–71. Springer, Brno, Czech Republic (2007)
13. Zimmermann, M., Bunke, H.: Automatic Segmentation of the IAM Off-Line Database for Handwritten English Text. In: Proc. of the 16 th Int. Conf. on Pattern Recognition (ICPR'02) Volume 4, p. 40035. IEEE Computer Society, Washington, DC, USA (2002)

http://www.springer.com/978-3-642-20226-1

Language Technology for Cultural Heritage
Selected Papers from the LaTeCH Workshop Series
Sporleder, C.; van den Bosch, A.; Zervanou, K. (Eds.)
2011, XXXII, 232 p., Hardcover
ISBN: 978-3-642-20226-1