Sequential Word Spotting in Historical Handwritten Documents

D. Fernández, J. Lladós & A. Fornés
Computer Vision Center – Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
08193 Bellaterra (Cerdanyola) Barcelona, Spain
{dfernandez, josep, afornes}@cvc.uab.es

R. Manmatha
Department of Computer Science
University of Massachusetts
Amherst MA 01003, USA
manmatha@cs.umass.edu

***Abstract*—In this work we present a handwritten word spotting approach that takes advantage of the a priori known order of appearance of the query words. Given an ordered sequence of query word instances, the proposed approach performs a sequence alignment with the words in the target collection. Although the alignment is quite sparse, i.e. the number of words in the database is higher than the query set, the improvement in the overall performance is sensitively higher than isolated word spotting. As application dataset, we use a collection of handwritten marriage licenses taking advantage of the ordered index pages of family names.**

I. Introduction

With the increase of massive digitization of documents residing in historical archives, the extraction of information has become a central task among the Document Analysis researchers and practitioners. In many cases, in particular in historical manuscripts, the full transcription of these documents is extremely difficult due the inherent deficiencies: poor physical preservation, different writing styles, obsolete languages, need of knowledge on the domain, etc. In certain applications, word spotting has become a popular and efficient alternative to full transcription. For example, word spotting is especially useful when the purpose is to retrieve information on people or events in historical or genealogical research from archives residing in municipalities or parishes.

The classical flow of a word spotting system starts by a word segmentation step, although recent contributions propose segmentation free approaches. Then, the extracted image word candidates are represented by a feature descriptor. Different types of descriptors have been proposed [1], [2], [3], [4]. Some methods describe the word image with a global representation, e.g., gradient, contextual, and convexity features [5], [6], [7], [8], features based on moments of binary images [9] or features based on the spatial distribution of shape pixels in a set of predefined image sub-regions [10]. Finally, a similarity function – e.g. Dynamic Time Warping (DTW) [11] , Hidden Markov Models (HMM) [12], [13] or Neural Networks [14] – is used to compare the query word with the candidate ones and a ranked list is generated as output accordingly. In some cases, to allow coping with large scale datasets, indexing strategies are proposed.

Word spotting inherently involves a high level of degradation in the images. Hence, instead of recognising the input text and searching the query word with an ascii string comparison, the problem is formulated holistically as a visual search of a given query shape in a larger image. Stated as a retrieval problem, and depending on the degradation level of images and the objective precision and recall, a number of false positive or true negative responses can be obtained. This level of error may be unacceptable, depending on the functional requirements of the application. There is an emerging trend of using contextual/semantic information to improve the performance of classical word spotting. Informally speaking, contextual word spotting [15] can be described as taking advantage of the lexical or syntactical structure of the sentence, i.e. the position of the query word in a sentence and the words before or after it (its context). If we integrate the joint probabilities of appearance of different words, we can overcome the individual misrecognition of one of them. Semantic word spotting [16] can be seen as a variant where the semantic categories of words reinforce the position where they can appear.

In complex frameworks, parsing tools allow to analyse the input according to a pre-defined context structure. In this work, we exploit a simpler structure referred as *sequential context* that can be useful in a certain cases. Some types of documents, e.g. form-like records (birth, marriage, death), present a repetitive structure in the words within text lines along the pages. According to the presence of such patterns consisting of sequences of non-consecutive words, by sequential context we mean the knowledge about the order of appearance of a certain set of words, among which is the query. Given this assumption, we benefit from the order, searching for the query word between the preceding and following words according to the known sequence. Therefore, *sequential word spotting* can be defined as an alignment between two sequences of word images. For highly noisy or distorted query words, the search within a sequence allows to fix anchor words in the document (preceding and following), and hence to restrict the search position.

Our proposed approach is inspired in previous works where the alignment of text sequences is used to correct errors between different editions of the same book [20], or to align original [19] and translated editions [21]. In other cases the semantic information is used to retrieval purposes [16] – words related semantically are given as similar, although the transcription and the shape of the word is different–. Other works adapts classical approaches, as Dynamic Time Warping [17] or Hidden Markov Models [18] to align the transcriptions of several documents.

As experimental framework, we use the *Llibre d'Esposalles* database [27]. It is a collection of books stored in the Barcelona Cathedral Archive, containing handwritten marriage licences. Each book contains the marriage licenses of one or two years. For each book, there is a separate set of index pages. The indices list the groom surnames (in some cases, the bride surnames too) in a pseudo-alphabetical order: the names appear in the chronological order of marriage, listed in sections corresponding to the initial letter. The indices were therefore written a posteriori. In some cases, by a different writer many years later. The indices contain errors and in some cases the page numbers do not correspond to the actual position of the surname entries. The objective of sequential word spotting allows to align the indices with the surnames of the licenses, overcoming wrong pagination and improving the performance of individual word spotting reinforced with the context.

The rest of the paper is organized as follows. Firstly, in section II the approach is explained in deep. Secondly, in section III the experiments and results are presented. We finally draw the conclusions in section IV.

II. Sequential Word Spotting

Classical Word Spotting approaches are methods widely accepted to extract the information of Historical Handwritten Documents, but although we can get pretty good results, it seems that they are reaching their ceiling. A quality leap in Word Spotting approaches can be achieved by introducing contextual information. The objective is to use this kind of information to improve the results obtained in the classical Word Spotting approaches. Concretely, our approach is inspired in time series alignment algorithms [24], [25], [26], to search instances of handwritten words in the same order of a given input list of words.

Starting with the premise that we have a *list A* and a second *list B* (longer than *A*) of words, and from the premise that *A* is included in *B* (in the same order of appearance), the objective of Sequential Word Spotting is to spot every word of the list *A* in the list *B*, taking into account the ordered sequence of these words.

Our approach follows the following steps: given the index and the licence documents, we first segment the words. Once the words are extracted, we compute the features for each word image. Next, the words in the index are aligned with the words in the marriage licences. A visual scheme of the model is shown in Figure 1. Next, we describe this process in detail.

A. Preprocessing and Word Segmentation

Since Historical documents can be affected by degradations, a preprocessing step is applied before segmenting words. First, the document is binarized, and the margins are removed (for further details, see [6]). The page is then segmented into blocks using the approach developed by Cruz et al. [22]. In the index pages we select the blocks that contains the surnames of the husband of the marriage license, and removed the blocks with the page numbers. In the marriage licence pages the page is segmented in three columns. The left column contains the surname of the husband of each marriage licence, the central column is the marriage licence and the right column is the fee that they paid for the wedding. For the purpose of this work,

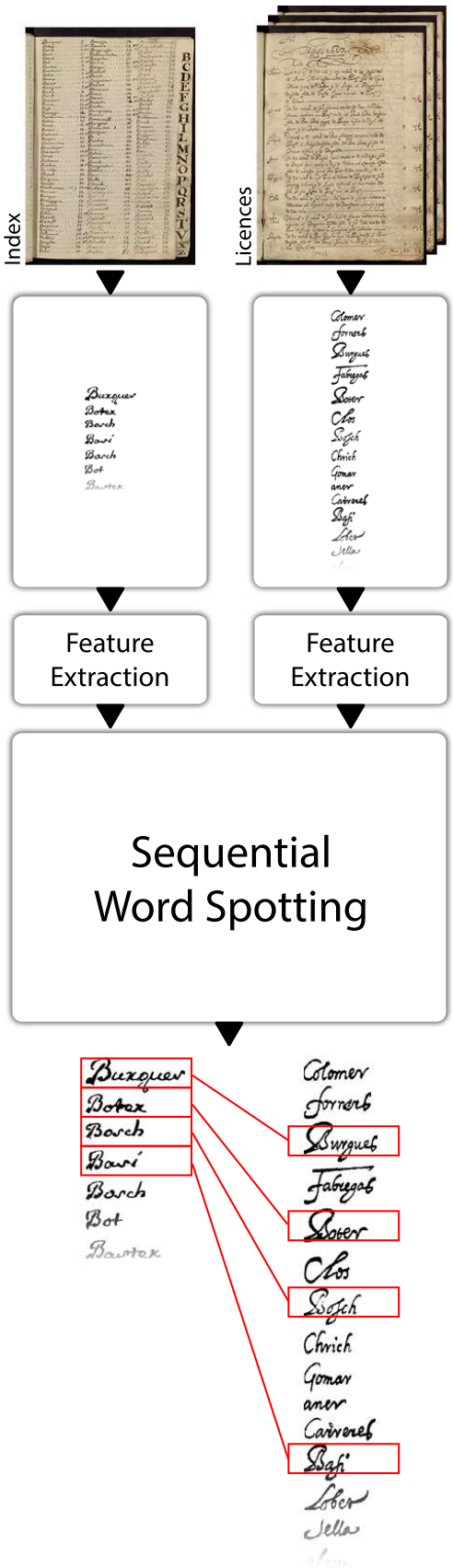


Fig. 1: Illustration of the entire process.

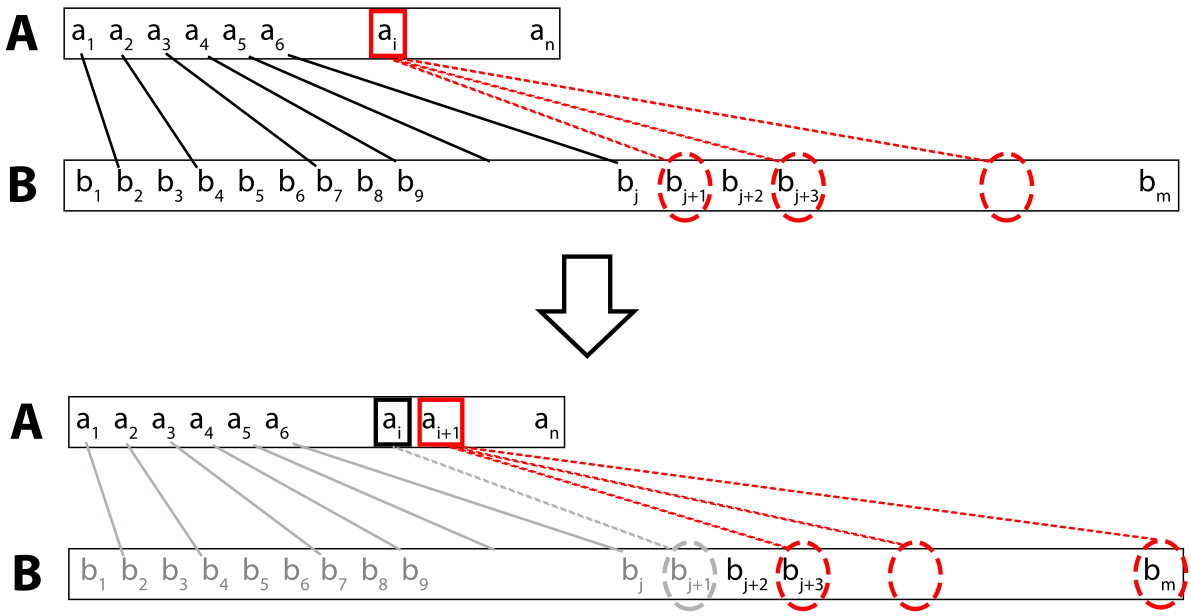


Fig. 2: Illustrative example of the mapping function f. The mapping function generates some candidates for the input query a_i. From the candidates, the query word is mapped to the word b_{j+1}. Then, the candidates for the query a_{i+1} are generated starting from the position of the previous mapped word b_{j+1}, taking into account the similarity to the input query a_{i+1}.

we select the left column and discard the rest. Once the blocks are selected, the words are extracted using a projection function which is smoothed with an Anisotropic Gaussian Filter [23].

B. Feature extraction

The objective of this work is to show the importance of using contextual information in classical Word Spotting approaches. Hence, it is a step in the flowchart that can be easily interchanged by any other descriptor. In this work we have used several descriptors (described section III). Some of them are suitable for multi-writer databases and others only for single-writer ones. In any case, the output of this step is a feature vector describing each selected word.

C. Sequential Word Spotting

The objective of Sequential Word Spotting is to find similar instances of words from the list of input queries, but taking into account the ordered sequence previously established.

Formally, given a sequence of handwritten query words $A = a_1, a_2, a_3, \ldots, a_n$ and a list of target words $B = b_1, b_2, b_3, \ldots, b_m$, where $n < m$ and $A \subset B$, the alignment is formulated by a mapping function f. Thus, let f be defined as a mapping function between the words in sequence A to some words in sequence B. Hence the mapping $j = f(i)$, where $i \in [1, \ldots, n]$ and $j \in [1, \ldots, m]$, represents that $a_i = b_j$. An order constraint $f(i) < f(i+1)$ is imposed in the mapping, in other words the the words in A appear in B following the same order. The alignment algorithm returns a valid mapping $f(j)$ where $j = 1, \ldots, m$ with the minimum cost.

The mapping function f selects the instances of B taking into account the similarity of the words and the sequence order of A. For each input instance a_i, the r most similar instances $\{b_{f^1(i)}, b_{f^2(i)}, \ldots, b_j, \ldots, b_{f^r(i)}\}$ from B are selected. Each candidate b_j is a possible solution of f and therefore it is explored to find the optimal solution. Then, for the next input instance a_{i+1} the instance candidates of B have to appear after b_j. In other words, the instances that appear before the position $j+1$ are not taken in account. The candidate chosen for a_i is the b_j with the minimum cost. The final cost is the cumulative sum of the distances between the a_i words with the corresponding $b_{f(i)}$ candidates. A visual illustration of the mapping function f is shown in Figure 2.

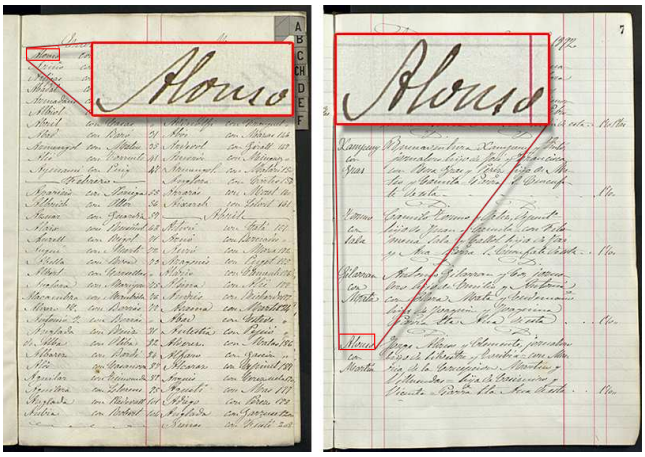
III. Experiments and Results

The experiments have been performed using two volumes of the Marriage Licenses Books from the Cathedral of Barcelona. The *Llibres d'Esposalles* [27] are a set of 244 books, written between 1451 and 1905, that contain information on approximately 550.000 marriages celebrated in over 250 parishes. Each book is written by a different writer, and is composed of the index and the marriage licenses. For the experiments we have used the following two volumes:

- *Volume A*: Multi-writer. This volume (see Fig.3a) is from the 17th century. The indices and the marriage licences were written by a different author. We have used all the words from the index that begin by the letter *A*. Thus, there are 186 words in the indices, and 1302 words in the licenses.
- *Volume B*: Single-writer. This volume (see Fig. 3b) is from the 19th century. Both the indices and the marriage licences are written by the same author. We have used the words that begin with the letter *B*. There are 182 word indices, and 1489 words in the licences.

We have evaluated the performance of the approach comparing our method with a classical Word Spotting approach

(a)



(b)

Fig. 3: Samples of the datasets used in this work: (a) *Volume A*: Multi-writer. (b) *Volume B*: Single-writer. The left column corresponds to indices, and the right column corresponds to the marriage licenses. We observe in the detail an example of words that belong to the same licence. Notice the handwriting variability of the word "Basi" in Volume A.

using several descriptors. The proposed descriptors can be categorized in two groups: single-writer and multi-writer. The first group contains descriptors that obtain very good results in a single-writer scenario, but do not seem robust enough for a multi-writer scenario. The Blurred Shape Model (BSM) descriptor [10] is based on computing the spatial distribution of shape pixels in a set of pre-defined image sub-regions. The Histogram of Oriented Gradients (HOG) descriptor [7] takes the pixel gradient information as the basis to extract features. The approach developed by Rath and Manmatha [28] uses the sequential information of graphemes to extract the features and Dynamic Time Warping (DTW) to compute the distance between them. The deformable HOG-based descriptor (nrHOG) [8] is an extension of the HOG descriptor for the specific case of handwriting, combining gradient features and a flexible and adaptable grid. The second group contains the work of [29], where an attribute-based approach learns how to embed the word images in a more discriminative space, where the similarity between words is independent of the handwriting style.

The metrics used to evaluate the performance of the experiments are:

- Number of True Positives (TP). A TP is considered when a word of the index is aligned with the correct license.
- TP TOP 5. Instead of evaluating when the searched word is returned in the first position of the vector of (closer) distances, we compute when the word is returned in the top 5 closer positions of the vector (TP TOP 5).
- Mean Distance. We evaluate the mean distance (Mean Dist.) between the computed position and the correct position of the query. This measures how far is the license selected compared to the correct position. So, a small value means that, although the alignment fails, the position of the returned licence is close to the correct position. Formally, let a_i be the word in A assigned to b_j through the mapping f, and $b_{j'}$ the correct instance of a_i in B (i.e. $b_{j'} = a_i$). The position distance between the computed position and the right position is defined as $pos\_dist(j, j') = |j - j'|$. The mean distance is computed as the average of the pos_dist of the failed aligned words.

Table I shows the comparison of several single-writer descriptors in the single-writer volume. It can be seen that we have outperformed the original word spotting method. In all cases, the accuracy is increased and the mean distance is reduced. Table II shows the results computed using a single (nrHOG) and a multi-writer descriptor (attributes). Both descriptors are evaluated using a single and a multi-writer dataset. We can observe that we outperform the original word spotting method again. The accuracy is increased and the mean distance is reduced, specially in the case of the single-writer descriptor over the multi-writer dataset. In such a case, one may conclude that the alignment in context is more helpful when the word shapes are dissimilar, and therefore, the shape descriptors less reliable.

IV. Conclusion and Future Work

The objective of this work is to demonstrate that contextual information improves the performance of a classical Word Spotting approach. We have proved that, using the ordered sequence of the words, we increase the accuracy and the false positives are closer to the searched word. Accordingly, we have proved that, using the spatial information, which relates the words of the documents, the result of Word Spotting approaches has been improved, specially in multi-writer databases.

One of the most difficulties of this work is to spot a specific instance between a big collection of words. Usually, a classical Word Spotting approach computes all the similar instances to the input query, but in this case we have to find only one instance. In classical Word Spotting approaches, this task becomes difficult to achieve. Hence we have evaluated the Word Spotting results using the top 5. But even in that case, the results has been improved in all the experiments.

Descriptor	TP (Align.)	Mean Dist. (Align.)	TP (WS)	Mean Dist. (WS)	TP TOP 5 (WS)
nrHOG	79 (42,47%)	20,47	18 (9,68%)	210,62	51 (27,42%)
BSM	64 (34,41%)	13,52	10 (5,38%)	168,68	27 (14,52%)
HOG	70 (37,63%)	14,32	20 (10,75%)	258,95	40 (21,51%)
DTW	50 (26,88%)	18,31	7 (3,76%)	214,92	25 (13,44%)

TABLE I: Evaluating single-writer descriptors using a single-writer dataset (Volume B). *WS* means a classical Word Spotting approach. *Align.* means our proposed alignment for word spotting.

Volume	Descriptor	TP (Align.)	Mean Dist. (Align.)	TP (WS)	Mean Dist. (WS)	TP TOP 5 (WS)
Vol. A (multi-writer)	nrHOG	10 (5,49%)	39,13	2 (1,10%)	683,80	7 (3,85%)
	attributes	68 (37,36%)	11,16	9(4,95%)	238,48	27 (14,84%)
Vol. B (single-writer)	nrHOG	60 (32,26%)	15,74	12 (6,45%)	254,26	31 (16,67%)
	attributes	79 (42,47%)	20,47	18 (9,68%)	210,62	51 (27,42%)

TABLE II: Evaluating a single-writer descriptor (nrHOG) and a multi-writer (attributes) descriptor using a single and a multi-writer dataset. *WS* means a classical Word Spotting approach. *Align.* means our proposed alignment for word spotting.

This work has been tested in a specific application and with small datasets. As future work, we plan to apply this approach in other kind of structured documents, showing that context information can be the quality leap in word spotting applications.

ACKNOWLEDGMENT

D. Fernández, J. Lladós and A. Fornés are partially supported by the Spanish projects TIN2011-24631 and TIN2012-37475-C02-02, by the EU project ERC-2010-AdG-20100407-269796 and by a research grant of the UAB (471-01-8/09). R. Manmatha is supported by the Center for Intelligent Information Retrieval and by NSF grant #IIS-0910884.

REFERENCES

[1] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," in *CVPR*, 1996, pp. 631–637.

[2] K. Terasawa and Y. Tanaka, "Locality sensitive pseudo-code for document images," in *ICDAR 2007*, vol. 1, 2007, pp. 73–77.

[3] J. A. Rodriguez and F. Perronnin, "Local Gradient Histogram Features for Word Spotting in Unconstrained Handwritten Documents," in *ICFHR*, 2008.

[4] R. Vinciarelli, S. Bengio, and H. Bunke, "Offline recognition of unconstrained handwritten texts using hmms and statistical language models," *PAMI*, vol. 26, pp. 709–720, 2004.

[5] B. Zhang, S. Srihari, and C. Huang, "Word image retrieval using binary features," in *Document Recognition and Retrieval XI*, vol. 5296, 2004, pp. 45–53.

[6] D. Fernández, J. Lladós, and A. Fornés, "Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure," in *Pattern Recognition and Image Analysis*, 2011, vol. 6669, pp. 628–635.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893 vol. 1.

[8] J. Almazán, A. Fornés, and E. Valveny, "Deformable hog-based shape descriptor," in *ICDAR*, 2013.

[9] A. Bhardwaj, D. Jose, and V. Govindaraju, "Script independent word spotting in multilingual documents," in *IJCNLP*, 2008, pp. 48–54.

[10] S. Escalera, A. Fornés, O. Pujol, P. Radeva, G. Sánchez, and J. Lladós, "Blurred Shape Model for binary and grey-level symbol recognition," *PR*, vol. 30, pp. 1424–1433, 2009.

[11] T. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, pp. 139–152, 2007.

[12] J. A. Rodriguez and F. Perronnin, "A model-based sequence similarity with application to handwritten word spotting," *PAMI*, vol. 34, no. 11, pp. 2108–2120, 2012.

[13] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM-based word spotting in handwritten documents using subword models," in *ICPR*, 2010, pp. 3416–3419.

[14] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *PAMI*, vol. 34, no. 2, pp. 211–224, 2012.

[15] D. Fernández, S. Marinai, J. Lladós, and A. Fornés, "Contextual word spotting in historical manuscripts using markov logic networks," in *HIP*, 2013, pp. 36–43.

[16] P. Krishnan and C. Jawahar, "Bringing semantics in word image retrieval," in *ICDAR*, 2013.

[17] E. Kornfield, R. Manmatha, and J. Allan, "Text alignment with handwritten documents," in *First International Workshop on Document Image Analysis for Libraries*, 2004, pp. 195–209.

[18] S. Feng and R. Manmatha, "A hierarchical, hmm-based automatic evaluation of ocr accuracy for a digital library of books," in *JDCL*, 2006, pp. 109–118.

[19] I. Yalniz, E. Can, and R. Manmatha, "Partial duplicate detection for large book collections," in *CIKM*, 2011, pp. 469–474.

[20] I. Yalniz and R. Manmatha, "A fast alignment scheme for automatic ocr evaluation of books," in *ICDAR*, 2011, pp. 754–758.

[21] I. Yalniz and R. Manmatha, "Finding translations in scanned book collections categories and subject descriptors," in *SIGIR*, 2012, pp. 465–474.

[22] F. Cruz and O. Ramos, "Handwritten line detection via an em algorithm," *12th International Conference on Document Analysis and Recognition*, 2013.

[23] D. Fernandez, J. Lladós, A. Fornés, and R. Manmatha, "On influence of line segmentation in efficient word segmentation in old manuscripts," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, pp. 763–768.

[24] E. Lawler and D. Wood, "Branch-and-bound methods: A survey," in *Operations Research*, vol. 14, no. 4, 1966, pp. 699–719.

[25] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," in *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, 2000, pp. 39–48.

[26] G. Navarro and R. Baeza-yates, "Very fast and simple approximate string matching," in *Information Processing Letters*, 1999, pp. 65–70.

[27] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The *Esposalles* database: An ancient marriage license corpus for off-line handwriting recognition," *PR*, vol. 46, no. 6, pp. 1658 – 1669, 2013.

[28] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," *CVPR*, vol. 2, pp. 521–527, 2003.

[29] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Handwritten word spotting with embedded attributes," in *ICCV*, 2013.