Spot it! Finding words and patterns in historical documents

Vladislavs Dovgalecs, Alexandre Burnett, Pierrick Tranouez, Stéphane Nicolas, Laurent Heutte

Spot it!
Finding words and patterns in historical documents

Vladislavs Dovgalecs*, Alexandre Burnett, Pierrick Tranouez, Stéphane Nicolas and Laurent Heutte
Université de Rouen, LITIS EA 4108, BP 12
76801 SAINT-ÉTIENNE DU ROUVRAY, France
vladislavs.dovgalecs@univ-rouen.fr*

***Abstract*—We propose a system designed to spot either words or patterns, based on a user made query. Employing a two stage approach, it takes advantage of the descriptive power of the Bag of Visual Words (BOVW) representation and the discriminative power of the proposed Longest Weighted Profile (LWP) algorithm. First, we try to identify the zones of images that share common characteristics with the query as summed up in a BOVW. Then, we filter these zones using the LWP introducing spatial constraints extracted from the query. We have validated our system on the George Washington handwritten document database for word spotting, and medieval manuscripts from the DocExplore project for pattern spotting.**

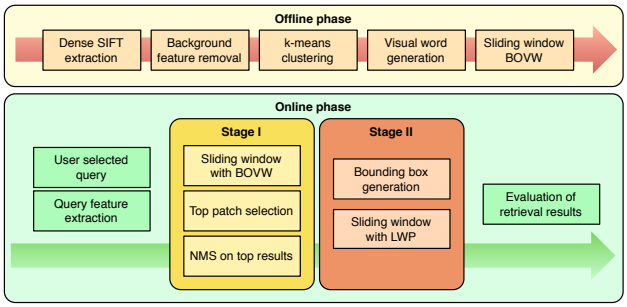

Figure 1. The offline and online processing phases of the proposed method.

I. Introduction

In the last decades a considerable effort has been made to digitize large collections of historical handwritten and printed documents and manuscripts. While the problems of preservation and access to these documents seem to be at least partially solved, there is a need for efficient and seamless search in such digitized documents. It is clear that manual search in more than a handful of scanned document pages becomes quickly unwieldy and automated search solutions are required.

Modern text-based search systems have reached some point of maturity and are efficient at searching very large text corpora. Unfortunately this requires manual transcription and annotation of scanned documents, which implies considerable human resources. Another option are modern optical character recognition systems (OCR) that work relatively reliably with scanned printed documents. These systems fail on handwritten historical documents due to writing irregularities, touching lines, free page layout, varying styles, degradations such as stains and tears, ancient fonts etc. In the last decades techniques called word spotting have emerged, which are attractive by their properties to solve this kind of detection problems.

In this paper we address the problem of unsupervised word spotting and graphical pattern spotting in historical handwritten documents, where the goal is to find regions containing the requested word or visually similar graphical pattern. The presented method is designed to retrieve both words and graphical patterns from a single query.

Indeed, word and pattern spotting using a single query is a very challenging problem due to multiple reasons. Typically the digitized manuscript is the only source of information and no annotation or transcription is available. This seriously limits state-of-the-art object detection methods [1] as their discriminative model training would require at least several positive class samples. It also remains unclear how to select negative class instances when training such two class classifiers. Furthermore, when resorting to matching techniques such as sliding windows, writing visual variability results in a relatively large number of false positives. These false positives are often visually similar to the query but do not necessarily contain the same word.

To answer these two challenges, we propose a segmentation-free method designed to spot both word and graphical patterns. The method consists in two phases: an offline phase feature extraction and an online query search phase. In the online phase, the algorithm proceeds with a two-stage query detection. The first stage performs rapid elimination of large zones with visually dissimilar content, while the second stage concentrates with more precise, feature order sensitive matching and scoring.

This paper is organized as follows: in section II we review the related works, in section III the proposed method is presented, in section IV experimental results are discussed and finally in section V we draw conclusion and sketch future research directions.

II. Related work

From literature one can point out three large groups of works aiming to solve the problem of word spotting, looking from the point of view of required pre-processing and ordering from most to least intensive processing: word segmentation, line segmentation and no prior segmentation.

Historically the first works adopted **word segmentation** pre-processing and used simple pixel-wise comparison approaches

for spotting in printed and handwritten documents [2], [3]. A more sophisticated method [4] uses cohesive elastic matching, which compares only informative parts of the template keyword for omnilingual word retrieval task. A probabilistic method modeling word shape and annotation was used for word recognition in historical documents [5] and implemented as a search engine in [6].

The second group of methods simplify the problem by requiring only successful **line segmentation** which is a simpler task. Hidden Markov Models (HMM) became rapidly very popular for handwritten word recognition by exploiting natural sequence information encoded in segmented text lines. In [7] a comparison of HMM-based word spotting to DTW matching using the same SIFT-like features showed a large improvement in performance giving favor to more sophisticated methods modeling ordered sequences of features. Unfortunately the HMM training requires large amounts of training data and considerable training time. Addressing this issue, an attractive sub-word model, where only a limited set of individual letters are needed for training, was proposed in [8].

Line segmentation in historical handwritten documents can be still a difficult task due to touching text lines and unconstrained layout. The third relatively recent group of methods requires no word or line segmentation and performs in a **segmentation-free** context. A typical approach consists in sliding a window over a document image and performing patch-by-patch comparisons. In [9] every patch is described using a Histogram of Oriented Gradients and color features while the negative effect of many false positives is countered using a boosted two-class large margin classifier. Authors in [10] attempt to discover underlying semantic structure from the BOVW features using an unsupervised Latent Semantic Analysis. A rapid spotting method exploiting adapted low-level features and efficient compact coding techniques was proposed recently in [11]. The HOG features are computed for each basic cell of an image allowing rapid description creation for an arbitrarily sized query. In order to increase the classifier generalization capabilities to unseen patterns, the query is learned using the Exemplar SVM. Finally the Product Quantization (PQ) method is used to further reduce HOG-PCA features down to a single real value enabling very compact storage of descriptors in a computer memory. In a similar spirit, a recent method [12] exploits an inverted file of quantized SIFT descriptors for rapid retrieval of word images and proposes a scoring method taking into account discriminant visual feature ordering information.

The main contribution of this paper is a segmentation-free word spotting method that rapidly eliminates non-matching regions in the first stage and concentrates on spatial ordering sensitive matching in those localized areas. Our method is composed of two stages: (1) initial candidate zone identification using orderless BOVW representations and (2) spatial ordering constraints enforcement using the proposed Longest Weighted Profile (LWP) algorithm. The secondary contribution of this paper is the robust LWP matching algorithm with sensitivity to visual feature location information. In this paper we focus on (1) robust matching techniques given a single query and (2) the system capable to work with both words and graphical patterns.

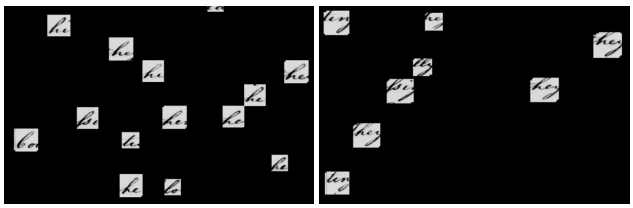

Figure 2. Illustration of some visual words belonging to two different clusters. Notice visually similar patches in each cluster with occasionally different content.

Reliable word spotting with a single query is a challenging task requiring careful feature selection and appropriate design of the decision stage. For instance, performance of a two class classifier may be severely limited as the trained model will usually rely on a single positive example. We argue that a careful selection of visual features and focus on matching techniques robust to false positives can provide meaningful detections.

Handwritten historical documents usually carry more than simple textual information. A manuscript can contain various graphical elements that could also be interesting to retrieve. To this end our method is capable of working not only with handwritten words but also with graphical patterns. Despite the lack of a classic training phase needed in the state-of-the-art object detection methods, our framework performs in a completely unsupervised setup and provides visually consistent retrieval results also on graphical patterns.

III. General presentation of the method

The proposed system consists of offline and online phases as depicted in Fig. 1. The main purpose of the offline phase is to extract dense local features, encode them as visual words using a codebook and produce BOVW representation features. In the online phase the BOVW representations and visual word information are used in a two-stage query search process. The procedure starts with a candidate zone spotting using BOVW representation information followed by another sliding window pass using the LWP algorithm for scoring.

A. Offline feature extraction

The offline processing phase begins with densely sampled SIFT features on a regular grid with a step of 5 pixels. At each grid location $f_{ij} = (x_i, y_i)$ three SIFT descriptors are extracted each covering a patch of size 10x10, 15x15 and finally 20x20 pixels. To lower the computational burden and storage requirements, we remove descriptors whose gradient norm does not exceed a certain threshold of 0.01, which removes most of the background features while keeping zones with non-uniform visual information. It is important to not compute rotation invariant SIFT descriptors since we want to discriminate between horizontal and vertical writing strokes for example. Our evaluation showed that the best performance were obtained if the three SIFT descriptors for a fixed feature f_{ij} are concatenated in a $3 \times 128 = 384$ dimensional vector d_{ij}. Visualization in image space of such concatenated patch

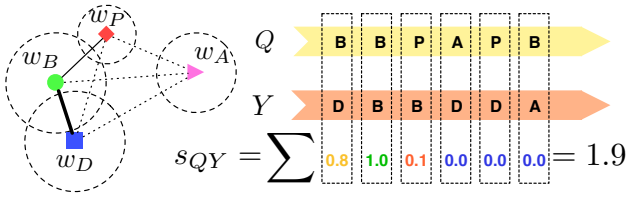

Figure 3. Longest Weighted Profile illustration. Four visual word clusters representing letters "B", "P", "D" and "A" are shown together with their mutual similarities (left). Illustrative example of comparing query (Q) and test (Y) sequences: green - perfect match, orange and red - partial match with variable degrees of similarity and blue - mismatch. Best seen in color.

descriptors showed more coherent and visually similar patches as in Fig. 2. Due to space limitations we do not present more results justifying this choice of features.

The processing proceeds with codebook creation using k-means algorithm for clustering. Empirical experimental results and cluster compactness analysis revealed approximately 1000-1500 natural clusters in the concatenated SIFT descriptor space. In this work we fixed the number of clusters to $k = 1500$. Furthermore, every descriptor d_{ij} is compared to the codebook of labeled cluster centers and is assigned the respective closest cluster label $w_l \in \{1, \ldots, k\}$ also called visual word.

For enhanced speed query comparison, we precompute fixed size patch BOVW representations following the same setup as in [10]. The size of a patch is fixed to 300×75 pixels and is sampled every 25 pixels in each document image of the database. Weak spatial information is encoded by additionally splitting every sliding window into two even cells.

B. Online query search

The online phase starts as a user outlines a query region Q depicting a word or a graphical pattern. The query search proceeds with feature extraction in the selected region and proceeds in two stages: (1) candidate zone detection followed by (2) false positive filtering using spatial constraints. It is important to note that the query Q is represented using weakly ordered BOVW feature Q_{BOVW} and as a sequence Q_{seq} of visual words. In this paper a sequence is constructed by projecting window enclosed visual words on a horizontal axis. Each representation is used in its respective processing stage.

1) Candidate zone detection: In the first stage, candidate zone detection works by comparing query feature Q_{BOVW} with all densely sampled sliding window features $Y^{ij}_{\text{BOVW}}, j = 1, \ldots, n_i$ in the i^{th} document image. The input of the first stage is the feature Q_{BOVW} and the densely sampled sliding windows covering each document image. The output of the first stage are zones of interest containing potential match described by bounding box windows. Every two pairs of BOVW features are compared using the χ^2 distance metric,

$$\chi^2\left(Q_{\text{BOVW}}, Y^{ij}_{\text{BOVW}}\right) = \frac{1}{2}\sum_{l=1}^{k}\frac{\left(Q_{\text{BOVW}}(l) - Y^{ij}_{\text{BOVW}}(l)\right)^2}{Q_{\text{BOVW}}(l) + Y^{ij}_{\text{BOVW}}(l)} \quad (1)$$

which is adapted to histogram comparison. Comparing the query to all sliding windows on a single page produces a vector of distances, where we retain only the top 200 results.

Analysis of results, which we do not show due to limited space, showed that irrespective of the size of the query, the retained sliding windows tend to form compact clusters indicating zones of interest. For each such localized zone of interest we build a bounding box. In order to reduce the number of degenerate cases of a large bounding box containing two or more weakly overlapping zones of interest, the non-maximum suppression algorithm is applied. Such post-processing of the results increases the chance of a zone of interest to be compactly represented while also reducing the search space for the second stage.

2) Introducing spatial information: The second stage of the method uses the detected zones of interest as input for further filtering. The output of the second stage is a set of query size windows with their respective similarity score.

The essence of the second stage is to enforce spatial ordering information characteristic of words and graphical patterns alike. As BOVW features used in the first stage encode only weak spatial information (windows are split vertically in two even size cells), experimental results showed that a large number of visually similar false positives are returned as well. We argue that enforcing a spatial arrangement of local features efficiently filters all zones with mismatching feature ordering.

Processing proceeds with a query size window, which we slide in each bounding box and compute the matching score using the LWP algorithm. The LWP algorithm uses a second representation of the query Q_{seq} and compares it to each sequence $Y^{ij}_{\text{seq}}, j = 1, \ldots, n_i$

$$s^{ij} = \text{LWP}\left(Q_{\text{seq}}, Y^{ij}_{\text{seq}}, M\right) \quad (2)$$

where n_i is the number of sliding windows in the ith bounding box and M is a matrix encoding intra-cluster visual similarities as discussed in subsection III-C. It is important to note that applying the LWP algorithm directly without preliminary first stage filtering would be prohibitively costly.

C. The Longest Weighted Profile (LWP) Algorithm

The proposed LWP algorithm is used to match robustly two selected regions represented as visual word sequences and return a similarity score based on the common information found in the two sequences. Ordering of visual words encodes fine to coarse spatial relationships of local features and is an important cue to distinguish between two similar but not identical regions. This sort of false similarity is due to the orderless nature of BOVW representation. A remedy to this problem is to leverage spatial ordering of visual words in each region. It is clear that, due to a priori suboptimal feature sampling and writing variability, two sequences (generated from two image words) cannot be compared directly element to element and that a noise insensitive matching procedure should be applied instead.

The proposed matching procedure (see Algorithm 1) features the required robustness properties. It takes as input two visual word sequences Q_{seq} and Y_{seq}, an inter-cluster similarity

Algorithm 1 The Longest Weighted Profile (LWP) algorithm.

INPUT:
Q - query sequence composed of m visual words
Y - test sequence composed of n visual words
M - $k \times k$ inter-cluster similarity matrix
OUTPUT: s_{QY} - similarity score
PROCEDURE
$S := \text{array}\,(0 \dots m, 0 \dots n) \leftarrow 0$
for $i := 1 : m$
  for $j := 1 : m$
    if $Q_i = Y_j$
      $S_{ij} := S_{i-1,j-1} + 1$
    else
      $\Lambda_{ij} := S_{i-1,j-1} + M_{Q_i,Y_j}$
      $S_{ij} := \max\left(S_{i,j-1}, S_{i-1,j}, \Lambda_{ij}\right)$
  end
end
return S_{mn}

matrix M and computes a similarity score. The score is higher if common and visually similar (encoded by matrix M) visual words are in the same order. Furthermore, by tolerating the small random variations in the writing of a word that may push a descriptor from one k-means cluster to another, it allows the algorithm to increase the similarity threshold, thus eliminating false positives without losing true ones. It is important to note that the algorithm is insensitive to noisy, mismatching feature insertions that may occur at arbitrary locations in the sequences. The standard substring search algorithm [13] is not robust in this context as it will return the longest contiguous substring which could be easily a subpart of a longer sequence interrupted by random insertions.

As depicted in Fig. 2, features carrying the same visual word label do not necessarily carry the same visual information. Counting only strict match and mismatch cases would reflect the perfect case of local feature clustering which is certainly not the case with real-world data. It is therefore reasonable to account for these perhaps artificially split clusters and assign some positive similarity score (see scoring illustration in Fig. 3) when comparing two different visual words belonging to two highly overlapping clusters. This prior information is encoded in the symmetric matrix M and we propose one possible way to compute it

$$M_{ij} = \max\left(0, \frac{\langle \mu_i, \mu_j \rangle}{\|\mu_i\| \, \|\mu_i\|}\right)^{\tau}, \forall i, j \in \{1, \dots, k\} \quad (3)$$

where $\mu_i, i = 1, \dots, k$ are the concatenated SIFT feature space cluster centers. The parameter $\tau > 0$ controls the trade-off between close and far cluster centers and is empirically set in all our experiments to $\tau = 50$. It is interesting to note that in the limit $\tau \to +\infty$, the matrix M reduces to an identity matrix and the algorithm reduces to the classic Longest Common Subsequence algorithm [13]. Therefore the running time of the proposed algorithm comparing two sequences of m and n elements using dynamic programming is $O\,(mn)$.

In Fig. 4 we show an example comparing the LWP algorithm over the LCS algorithm in the word spotting context.

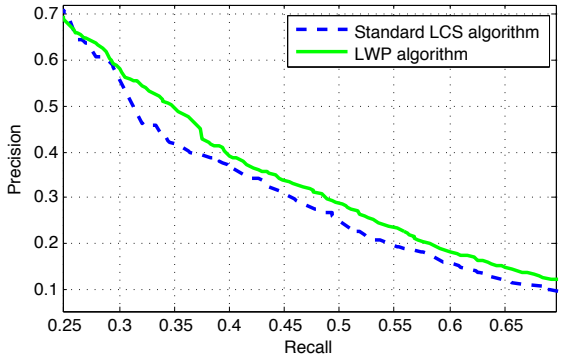

Figure 4. Performance comparison of the standard LCS [13] and proposed LWP algorithms. Evaluation is done exhaustively on all the words from 2 pages of the George Washington database.

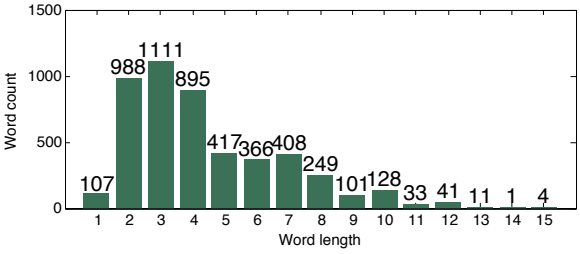

Figure 5. Relative number of annotated words in the George Washington database grouped by the length.

IV. Experimental results

In this paper the evaluation of word spotting performance is done implementing the protocol used in [10] on the George Washington handwritten document dataset. For graphical pattern spotting, to our knowledge there is no publicly available historical document collection with pattern annotation. To this end we provide qualitative results of pattern spotting on our in-house database of historical documents.

The evaluation protocol assumes that both databases are annotated with a unique label assigned to each window carrying the same information (word or pattern). System performance is evaluated by querying each annotated element in turn and measuring Precision and Recall. We consider detection of a word or pattern to be successful if the ground truth window is covered more than 50% of its surface by a result window.

The performance of the proposed spotting method using all 4860 words as query on George Washington handwritten documents is shown in Fig. 6. The results are grouped by the length of query in characters for convenience of comparison. The relative number of annotated words in this database is depicted in Fig. 5. Globally the results reflect the difficulty to retrieve short words and less difficulties for longer words. This is an expected behavior since short words (e.g. 1-3 characters) are described with less visual features and the search results

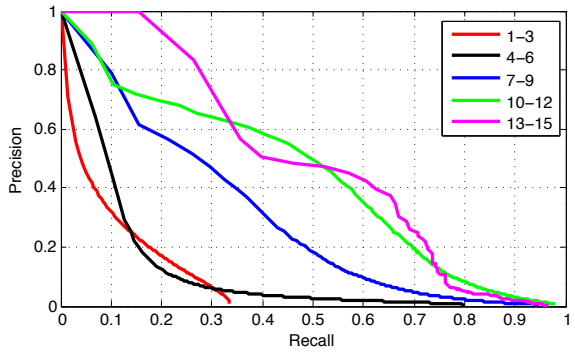

Figure 6. Word spotting performance using all 4860 annotated words.

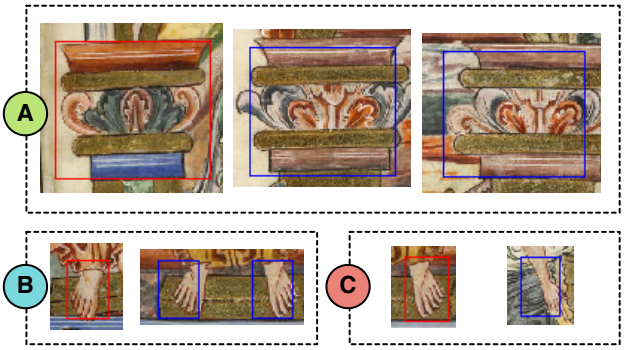

Figure 7. Qualitative graphical pattern spotting performance. The query window is shown in red and the corresponding retrieved results in blue.

contain many false positives. From the use case perspective, these words are often stop words and their retrieval utility in real-life scenario may be limited. The retrieval results are significantly better for queries composed of 7-9 and 10-12 characters. On this particular database, up to 30%-40% recall can be safely attained without sacrificing much in precision. The best results are obtained using the longest words (13-15 characters) and up to 50% of word can be retrieved with approximately the same precision.

Finally, we evaluated the performance of the method in the context of graphical pattern spotting with some results shown in Fig. 7. Overall the results show consistency in structure and visual content of the query and the retrieved regions (examples A and B). Analysis of the results showed that the first stage results provide already good estimates of the visually similar regions with little improvement from the LWP algorithm. The typical false positives (example C) arose from examples carrying high level information that are not leveraged in this spotting method. Finally, we encountered difficulties to select the best threshold score, which is characteristic of systems using no classifiers. Practical implementation in the Docexplore platform [14] allows this threshold to be controlled by the user and can be tuned manually as a sensitivity parameter.

V. Conclusions and Future Work

The proposed segmentation-free spotting method has been shown to work with both word and graphical patterns. The system works remarkably well if we consider the completely unsupervised learning approach and a single query. The proposed method demonstrates the power of discriminative BOVW features for candidate zone detection and the LWP algorithm for enforcing spatial information. In future work we intend to learn discriminative visual features from the document images and provide deeper performance analysis of the system for graphical pattern spotting. We expect that document-specific visual features could be useful to further improve the retrieval results for medium to long words. In order to reliably detect short words, discriminative prior knowledge (e.g. use the whitespace as a cue for word separation) could be integrated by extending the decision stages of the proposed method.

Acknowledgment

This work has been done in the context of the EU Interreg IVa Docexplore project (http://www.docexplore.eu). The authors would like to thank the Municipal Library of Rouen, France for providing digitized historical manuscripts and Cécile Capot for the great effort she put in graphical pattern annotation.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on PAMI*, vol. 32, no. 9, 2010.

[2] S.-S. Kuo and O. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models," *IEEE Transactions on PAMI*, vol. 16, no. 8, pp. 842–848, 1994.

[3] R. Manmatha, C. Han, and E. M. Riseman, "Word Spotting: A New Approach to Indexing Handwriting," in *Proc. of CVPR*. IEEE Computer Society, Jun. 1996, pp. 631–637.

[4] Y. Leydier, A. Ouji, F. Lebourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009.

[5] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proc. of DIAL*. IEEE Computer Society, 2004, pp. 278–287.

[6] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proc. of Intl. ACM SIGIR Conf. on Research and development in IR*. New York, NY, USA: ACM, 2004, pp. 369–376.

[7] J. A. Rodriguez and F. Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," *In Proc. of ICFHR*, pp. 7–12, August 2008.

[8] S. Thomas, C. Chatelain, L. Heutte, and T. Paquet, "An Information Extraction model for unconstrained handwritten documents," in *ICPR, Istanbul, Turkey*, 2010, pp. 3412–3415.

[9] P. Yarlagadda, A. Monroy, B. Carque, and B. Ommer, "Recognition and analysis of objects in medieval images," in *Proc. of ICCV*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 296–305.

[10] M. Rusiñol, D. Aldavert, R. Toledo, and J. L. o. s, "Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method," in *In Proc. of ICDAR*, 2011, pp. 63–67.

[11] J. Almazan, A. Gordo, A. Fornés, and E. Valveny, "Efficient Exemplar Word Spotting," in *British Machine Vision Conference*, 2012.

[12] I. Z. Yalniz and R. Manmatha, "An Efficient Framework for Searching Text in Noisy Document Images," *10th IAPR Intl. Workshop on DAS*, pp. 48–52, 2012.

[13] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Higher Education, 2001.

[14] P. Tranouez, S. Nicolas, V. Dovgalecs, A. Burnett, L. Heutte, Y. Liang, R. Guest, and M. Fairhurst, "DocExplore: overcoming cultural and physical barriers to access ancient documents," in *Proc. of ACM symposium on Document engineering*. Paris, France: ACM, 2012, pp. 205–208.