Wordspotting on Historical Document Images

Thomas Konidaris †

National and Kapodistrian University of Athens
Institute of Informatics and Telecommunications
tkonid@iit.demokritos.gr

Abstract. In this dissertation innovative methods of wordspotting on historical printed documents are presented. In particular, two methods based on document segmentation on word level have been developed. The first method uses a hybrid feature scheme for word matching based on zones and projections. It also uses a process of creating query keyword images for any word using synthetic data. The synthetic words are created using images of individual characters taken from the processed documents. The method also presents a process allowing user feedback in order to improve the final results. The second method uses the Dynamic Time Warping (DTW) algorithm for comparing word images. It assist the transition between the synthetic data and real data comparison. Synthetic data and real data differ and DTW allows a better alignment between the features of the two images. Again, feedback can be applied to improve the results. Furthermore, a method that uses no segmentation on the document images has been also developed. The method overcomes the problem of incorrect segmentation that affect the final results since it detects query keyword images directly on entire document page images. It also allows for partial matching such as detecting word that are included in larger ones. The evaluation of the aforementioned methods showed satisfactory results presenting better performance against competitive methods of wordspotting.

1 Introduction

World wide libraries hold a vast amount of historical documents in terms of books, papers, drawings, journals, etc. These documents are highly valuable items due to the information they contain as well as the historical importance and rarity that characterizes them. The digitization of such archival and historical collections is an ongoing process that results to digital content which allow access to the information without distorting the original material. It is clear that efficient indexing and retrieval are important prerequisites of any system that manipulates such digital content. Optical Character Recognition (OCR) is a standard technology that is widely used in indexing documents with noticeable results in contemporary documents. However, historical documents are prone to a number of difficulties such as typesetting imperfections, document degradations and low print quality which decrease the performance level of OCR systems [14], [23], [10], [9].

† Dissertation Advisor: Sergios Theodoridis, Professor

2 Dissertation Summary

Word spotting is an alternative methodology for document indexing based on spotting words directly on images without the use of any OCR procedures. Thus, a word spotting system attempts to detect words as a whole rather than to exactly recognize the characters as in OCR. In a typical scenario, the query image is selected from a set of predefined keywords of interest or is interactively defined by the user by cropping a rectangular image area that serves as query example. The word spotting system uses the query and detects similar words in document images based on image matching techniques without any conversion of the images into readable text. Extensive studies have shown that indexing terms in documents automatically using a word spotting system makes it possible to use costly human labor more sparingly than a full transcription would require [22]. Several word spotting methods rely on a pre-processing step where the document image is segmented into words. The segmented words are then compared to the query image in order to detect potential matches. Recently, segmentation-free methods have been also proposed that do not require any segmentation and the document image is treated as one entity by-passing any errors that may occur due to poor segmentation results. In this line, we propose a segmentation-free word spotting method for historical printed documents. The method is based on local keypoint correspondences and consists of two distinct steps that determine candidate image areas in order to accurately extract the final bounding boxes which indicate the word instances in the document page. The method is evaluated using two different datasets of different languages and the experimental results show that the proposed method outperformed significantly the competitive approaches.

The word spotting literature can be divided into two main categories depending upon whether segmentation of the document image is applied or not. Indeed, there are several methods that are based on page segmentation as a preprocessing step, while others are applied directly to the document image. Additionally, a variety of features are used in order to describe the query word as well as the document image. These features strive to efficiently express the geometric and local information of the visual content and include projection profiles, Gabor features, zones and gradient-based features, to name a few. Keypoint-based local features have been also successfully used in order to describe document images as a set of local feature vectors that are invariant to scale changes, illumination and distortions. The Scale Invariant Feature Transform (SIFT) [19] is a well known technique in this category that produces an adequate number of distinctive features even for small visual objects. In the following, we summarize some word spotting techniques that rely at least in part on segmentation as well as approaches where no segmentation is required.

2.1 Segmentation based methods

There are three levels of page segmentation that are typically used for detecting words in documents, namely segmentation into lines [11], [20], words [12],

[13], [21] [24], [25], or even characters [2], [8]. Profile features, such as upper or lower word profiles, projection, density or transition profiles have been reported to successfully represent words in a document image that has undergone word level segmentation [22], [24]. Fusion of multiple features is also adopted in several studies in order to improve the word image description. For example, in [26] a multiple feature scheme is used consisting of projection profiles, upper/lower word profiles and background-to-ink transitions. Similarly, Jawahar et al [6] involve word profiles to describe the outline shape of the word, structural features to extract statistical information like moments or variation and finally Fourier coefficients as a compact representation of the features in the frequency domain. In [12] a hybrid feature scheme based on a combination of projection profiles and upper/lower word profiles is used for matching words segmented from document images. In [8], a word spotting method is proposed based on mesh features and in [29] and [33] the feature scheme used is gradient-based binary features. Another feature used for word spotting is based on skeletons and is used in the works of [18] and [7]. Gabor features can also be applied for word spotting as proposed in [2]. In Li et al [17] a word image is decomposed into vertical strokes and a stroke-based coding scheme is built for all the word in the document database. Considering features basedon local keypoints, Ataer et. al. [1] use SIFT features in order to match segmented words from Ottoman documents. Similarly, in [32] a word image matching method is presented using SIFT descriptors on keypoints that are extracted using the Fast-Corner-Detection algorithm [27]. These features are quantized into visual terms (visterms) using hierarchical K-Means algorithm and indexed using an inverted file. In [30] a word spotting method based on line segmentation is presented. The method uses a sliding window over each line. The matching is performed using dynamic programming and slit style HOG features. In the previous methods, the segmented words are presented as feature vectors and Dynamic Time Warping is an algorithm that has been extensively used to match words based on these vectors [25], [26], [6], [13], [11]. Other matching techniques are based on morphological variants [21], voting schemes [18], [1], [32], similarity distances [12], [7], character or string matching [8], [17] and correlation measures [29], [20], [33]. Overall, document segmentation results to higher level structures that are semantically important and can be further explored. On the other hand, detection methods based on segmentation results are intrinsically prone to errors like over- or under- segmentation as well as well as partial occlusion and mis-segmentation.

2.2 Segmentation-free approaches

Although, there is a very large collection of published work concerning the segmentation approach, in the recent years there is a growing research interest concerning segmentation-free methods. There are cases where documents cannot be segmented correctly leading to insufficient results. The segmentation-free approaches overcome the problems associated to bad segmentation results by treating the document image as a whole. In [4] a template matching method

based on pixel densities is used for locating words in documents without segmenting them. Although the method provides rotation and scale invariance, this is applied on limited extent. In [16] an alphabet is used that is manually selected from each document collection processed. The alphabet is used to create word instances that serve as queries. The features extracted are based on gradient values. In [15] gradient information is also used as features for the word images. Word interest points are matched against document images and try to locate zones of interest presenting similar features. Local image features have been also used in segmentation-free methods trying to benefit from the scale and rotation invariance they offer as well as their robustness to noise. Such methods usually involve a voting scheme in order to detect and localize potential word matches in the document image. In this line, Rusinol et al [28] opt SIFT features in a bag-of-visual-words approach. The method is applied on both handwritten and printed documents. The SIFT features are extracted using small predefined squared areas that are assumed to cover most of the font sizes. The search space does not correspond to the entire document image but rather to overlapping local patches of fixed geometry. However, several assumptions concerning the size of the patch and the expected font sizes seriously affect the generalization and the applicability of the method. Furthermore, as the authors mention, the performance of the system is highly related to the length of the queried words.

2.3 Contribution of Research

Our research concentrated into both approaches. In particular we have developed two methods that are based on the segmentation approach and one method that follows the segmentation-free approach. The methods that are based on the segmentation approach segment the document on word level. The first method uses synthetic data to create query keyword images. Each query keyword is created in a synthetic manner using individual character images taken from the processed documents. This way we are able to construct any query word image we like. The feature scheme used combines two different features. The first one is zones and the second is projection profiles. The features are matched using a simple distance metric. The advantage of the method lies on the fact that is very fast in producing an initial set of results. These are further improved through a user's feedback process. The second segmentation-based method uses synthetic data to create query keyword images and a combination of four different features. However, unlike the above method, the features are compared using the Dynamic Time Warping (DTW) algorithm. The DTW algorithms manages to overcome local distortions between the compared feature vectors. This method performs better than the former segmentation method but needs more time to complete as DTW algorithm poses bigger complexity.

On the other hand, the segmentation-free method that was developed comes to solve the problem of incorrect segmentation. There are cases where the documents are not segmented correctly. This error percentage affects the overall performance of the methods. The segmentation-free method does not require the documents to be segmented at any level. Rather, the query keyword images

are compared with the entire document page images. The method uses the SIFT algorithm for the extaction of keypoints and their descriptors. The performance of the method is very satidfactory.

3 Results and Discussion

In this section we will discuss the segmentation-free method that was developed during our research. In the proposed method we adopt a segmentation-free word spotting approach in order to overcome the poor segmentation results that usually characterize historical documents. We are based on SIFT features that have been proved to provide robustness concerning low image quality and image degradation. However, detection of word instances based on direct matching between query and image SIFT keypoints leads to unsatisfactory results. This is due to the spatial scattering of matching correspondences since a query keypoint may be similar to a large number of document page keypoints. These document keypoints do not a priori belong to correct word instances. Furthermore, the existence of multiple word instances in the same document page does not allow the query image to be matched by a sufficient number of correspondences.

The proposed method does not adopt the original matching process described in the SIFT algorithm, but instead a two step approach is followed. In the first step, for every keypoint in the query keyword image, the nearest K points are found in the document page image. These document keypoints are used as indicators in order to create candidate image areas. In the second step, each candidate image area is matched against the query keyword image. The keypoint correspondences are used by the RANSAC algorithm in order to estimate the final bounding boxes indicating the detected word instances. Furthermore, we use the strength of SIFT descriptors in a way that multiple instances of the desired word can be found on the document page.

3.1 Detection of Candidate Image Areas

The first step of the proposed method involves the matching of the query keyword keypoints to the document keypoints. The purpose is to find point correspondences on the document image that will serve as indicators of candidate image areas. For each keypoint of the query keyword we locate the K most similar keypoints on the document image. The value of K is experimentally defined as discussed in section 4. Let f_q and f_d be the SIFT feature vectors of the i^{th} keypoint in the query keyword image and the j^{th} keypoint in the document image, respectively. The distance between these two keypoints is calculated as follows:

$$d(i,j) = cos^{-1}(\langle f_q^i, f_d^j \rangle) \tag{1}$$

where $\langle f_q, f_d \rangle$ denotes the dot product between the two normalized vectors.

Each pair of corresponding keypoints defines a candidate image areas on the document page. Since we know the relative position of the query keyword

keypoint in respect to the edges of the query keyword image we define a bounding box around the keypoint on the document image taking into account the position of the corresponding keypoint of the query keyword image. Let $p_q(x_q, y_q)$ be a point on the query keyword image and $p_d(x_d, y_d)$ be its corresponding point on the document page image. Let dx, dy be the distance of the query keypoint from the left and the top edge of the query keyword image respectively. The bounding box surrounding the candidate image area is defined by its top-left (x_{min}, y_{min}) and bottom-right (x_{max}, y_{max}) corners is given by the following equations:

$$x_{min} = x_d - (\frac{sc_d}{sc_q} \cdot dx \cdot t_s) \tag{2}$$

$$y_{min} = y_d - (\frac{sc_d}{sc_q} \cdot dy \cdot t_s) \tag{3}$$

$$x_{max} = x_d + [(w_q - x_q) \cdot \frac{sc_d}{sc_q} \cdot t_s] \tag{4}$$

$$y_{max} = y_d + [(h_q - y_q) \cdot \frac{sc_d}{sc_q} \cdot t_s] \tag{5}$$

where w_q and h_q are the width and height of the query keyword image, respectively. Parameter t_s is the boundary size factor, which gives extra space to the boundaries of the candidate image areas and has been experimentally set to 1.1. Variables sc_q and sc_d are the scales of the query keypoints p_q and p_d, respectively, as provided by the SIFT algorithm.

3.2 Detection of Word Instances

In the previous section we matched the query keyword image with the entire document page image in order to use the matching keypoints as indicators for creating candidate image areas. These areas cannot guarantee that they contain the query word under consideration. For this reason the keypoints of the query keyword image are matched against the keypoints of each candidate image area. For each keypoint on the query keyword image we find the most similar keypoint on the candidate image area using Eq. 1. In order to estimate a model that describes the efficiency of these keypoint correspondences the RANSAC algorithm [3] is involved. RANSAC is an iterative method that can efficiently estimate the parameters of a model even when the measurements contain outliers. Using RANSAC the number of inliers is calculated, that is, the number of corresponding pairs that are conveniently described by the model. Moreover, the keypoint correspondences are used in order to calculated a homography that serves as a transformation matrix from the query keyword image to the candidate image area plane [5]. There must be at least four point correspondences to calculate the homography matrix. Let $P_q = (x_q, y_q)$ be a point in the query keyword image and $P_c = (x_c, yc)$ be the corresponding point in the candidate image area. The transformation between these two points can be given by the following equation:

$$P_c = H \cdot P_q \tag{6}$$

where H is the homography matrix. The above equation can take the form:

$$\begin{bmatrix} x_q \\ y_q \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11}\ h_{12}\ h_{13} \\ h_{21}\ h_{22}\ h_{23} \\ h_{31}\ h_{32}\ h_{33} \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix}$$

This process is applied for all candidate image areas aiming to produce a set of bounding boxes that are afterwards ranked according to their matching efficiency. The inliers percent provides an indicator of the goodness of fit regarding the RANSAC model. However, there may be a number of detected bounding boxes having equal inliers percent value. In order to distinguish them, we propose to divide the inliers percentage value L by a quantity D which corresponds to the sum of the distances between the query and the candidate image area keypoints. Thus, for a candidate image area the ranking value V is calculated as follows:

$$V = \frac{L}{D} \tag{7}$$

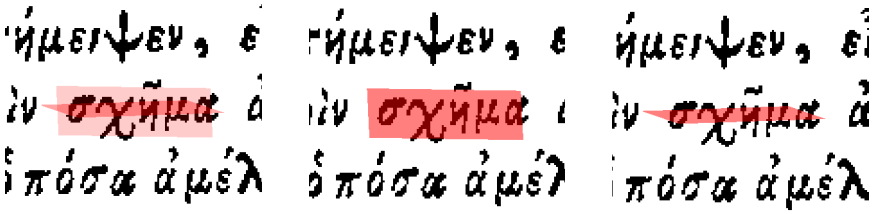
3.3 Removing Overlapping Results

We have seen that the candidate image areas are created using the point correspondences between the query keyword image keypoints and the keypoints of the document page image. There are cases where more than one candidate image areas correspond to the same word in the document image. Therefore, we end up with overlapping bounding boxes, each of them having different ranking values V as calculated by Eq. 7. On the document image, two bounding boxes B_i and B_j are considered overlapping if the following equation holds:

$$IoU = \frac{B_i \cap B_j}{B_i \cup B_j} \geq t_v \tag{8}$$

where t_v has been experimentally defined equal to 0.3. The bounding box that has the larger ranking value V among the overlapping bounding boxes is the one kept while the others are discarded from the list. Figure 6 illustrates an example of resulting bounding boxes concerning the same candidate image area. The two bounding boxes are considered overlapping since their intersection over union ration exceeds the threshold t_v , as shown in Figure 1(a). However, the bounding box in Figure 1(b) has a smaller ranking value V than the bounding box in 6(a) and it is discarded from the list of bounding boxes. The remaining bounding boxes are further filtered out using the word length of the query keyword image. The bounding boxes which the following equation holds are excluded from the list of the results.
where w_b is the length of a bounding box from the results list, w_q is the length of the query keyword image and t_w is the threshold which has been experimentally set to 0.4.

$$|\frac{w_b - w_q}{w_q}| \geq t_w \tag{9}$$



(a) $IoU = 0.3212$ (b) $V_i = 0.0207$ (c) $V_j = 0.0183$

Fig. 1: Resulting bounding boxes B_i and B_j for the same word on a document page image. (a) Their intersection over union ratio, (b) The bounding box B_i with ranking value of 0.0207, (c) The bounding box B_j with ranking value of 0.0183. The bounding box B_j is discarded since it has lower ranking value V.

3.4 Experimental Results

The experiments that were conducted in order to evaluate the proposed method included two different datasets. The first dataset consists of 100 pages from a Greek historical typewritten book from the period of Renaissance and Enlightenment. The second dataset consists of 100 pages of a German historical typewritten book of Eckartshausen which was published in 1788 and is owned by the Bavarian State Library [31]. For the Greek dataset we have used seven (7) query keyword images as queries and for the German dataset we have used ten (10) keyword images as queries.

The proposed method was compared against the method presented in [4] and original SIFT. Figure 2 shows the performance of the proposed method against the competitive ones concerning the Greek dataset. Likewise, the results for the German dataset are shown in Figure 3.

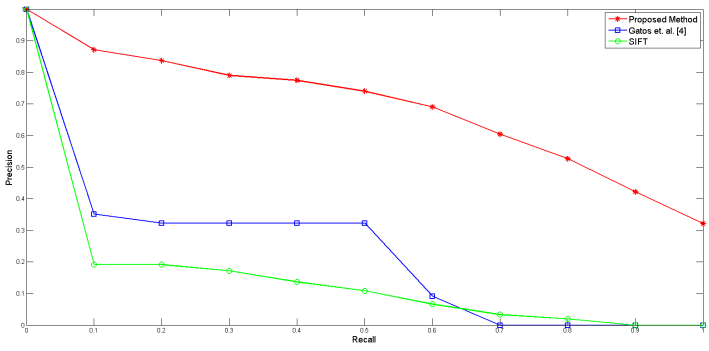


Fig. 2: Precision-Recall curves for the different methods concerning the Greek dataset.

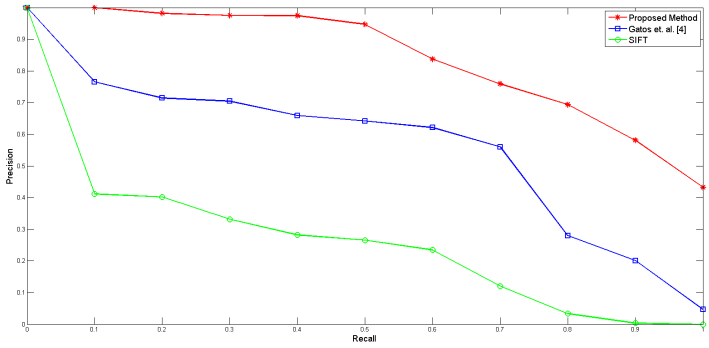


Fig. 3: Precision-Recall curves for the different methods concerning the German dataset.

The proposed method outperforms the competitive methods significantly. Furthermore, there is clear evidence that if we apply original SIFT matching between a query keyword image and a document page image in order to get keypoint correspondences that will serve as indicators to the creation of candidate image areas, the results are very poor.

4 Conclusions

The research aimed at creating methods for wordspotting. Our methods touched both segmentation and segmentation-free approaches. In particular, we have developed two methods that are based on the segmentation approach. These are a fast method that uses synthetic data, user feedback and a feature scheme that combines two different features. The method performed very well and in small amount of time. The second method used the DTW algorithm in order to solve the problem of variations and distortions that is found between words. This method, outperforms the first segmentation based method giving much better results. As far as the segmentation-free method is concerned, we have developed a method that does not require any prior segmentation of the document images. The query keywords are matched against the entire document page image. Such segmentation-free methods starting to gain great attention since they overcome the problems of bad segmentation and can even be used at documents that segmentation fails dramatically. The results of the mehtod are very encouraging as well as satisfactory.

References

1. Esra Ataer and Pinar Duygulu. Matching ottoman words: an image retrieval approach to historical document indexing. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 341-347, 2007.

2. H. Cao and V. Govindaraju. Template-free word spotting in low-quality manuscripts. In *6th International Conference on Advances in Pattern Recognition (ICAPR'07)*, pages 45-53, 2007.
3. M. A. Fishler and R. C. Bolles. A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the Association of Computer Machinery*, 24(6):381-395, 1981.
4. B. Gatos and I. Pratikakis. Segmentation-free word spotting in historical printed documents. In *10th International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 271-275, Barcelona, Spain, 2009.
5. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.
6. C. V. Jawahar, A. Balasubramanian, and M. Meshesha. Word-level access to document image datasets. In *Proceedings of the Workshop on Computer Vision, Graphics and Image Processing (WCVGIP)*, pages 73-76, 2004.
7. P. Keaton, H. Greenspan, and R. Goodman. Keyword spotting for cursive document retrieval. In *Workshop on Document Image Analysis*, pages 74-81, San Juan, Puerto Rico, 1997.
8. S. Kim, S. Park, C. Jeong, J. Kim, H. Park, and G. Lee. Keyword spotting on korean document images by matching the keyword image. In *Digital Libraries: Implementing Strategies and Sharing Experiences*, volume 3815, pages 158-166, 2005.
9. V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos. A complete optical character recognition methodology for historical documents. In *Tenth International Conference on Document Analysis and Recognition (ICDAR)*, pages 501-505, Barcelona, 2009.
10. A. Koerich, R. Sabourin, and C. Y. Suen. Devanagari ocr using a recognition driven segmentation framework and stochastic language models. *International Journal on Document Analysis and Recognition (IJDAR)*, 6(2):126-144, 2003.
11. A. Kolcz, J. Alspector, M. Augusteijn, R. Carlson, and G. Viorel Popescu. A line-oriented approach to word spotting in handwritten documents. *Journal of Pattern Analysis and Applications*, 3(2):153-168, 2000.
12. T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis. Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. *Internation Journal of Document Analysis and Recognition (IJDAR), special issue on historical documents*, 9(24):167-177, 2007.
13. T. Konidaris, B. Gatos, S. J. Perantonis, and A. Kesidis. Keyword matching in historical machine-printed documents using synthetic data, word portions and dynamic time warping. In *The eighth IAPR Workshop on Document Analysis Systems*, pages 539-545, 2008.
14. F. Lebourgeois, J.-L. Henry, and H. Emptoz. An ocr system for printed documents. In *Proceedings of IAPR Workshop on Machine Vision Applications*, pages 83-86, Tokyo, Japan, 1992.
15. Y. Leydier, F. LeBourgeois, and H. Emptoz. Text search for medieval manuscript images. *Pattern Recognition*, 40:3552-3567, 2007.
16. Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz. Towards an omnilingual word retrieval system for ancient manuscripts. *Pattern Recognition*, 42(9):2089-2105, 2009.
17. L. Li, S. J. Lu, and C. L. Tan. A fast keyword-spotting technique. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 68-72, 2007.

18. J. Lladós and G. Sánchez. Indexing historical documents by word shape signatures. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 362-366, 2007.
19. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91-110, 2004.
20. A. Marcolino, V. Ramos, M. Ramalho, and J.R. Caldas Pinto. Line and word matching in old documents. In *Fifth IberoAmerican Symposium on Pattern Recognition (SIAPR)*, pages 123-135, 2000.
21. M. Meshesha and C. V. Jawahar. Matching word images for content-based retrieval from printed document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 11(1):29-38, 2008.
22. A. Murugappan, B. Ramachandran, and P. Dhavachelvan. A survey of keyword spotting techniques for printed document images. *Artificial Intelligence Review*, 35(2):119-136, 2011.
23. P. Natarajan, I. Bazzi, Z. Lu, J. Makhoul, and R. M. Schwartz. Robust ocr of degraded documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 357-361, 1999.
24. T. M. Rath and M. Manmatha. Word spotting for historical documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 9(24):139-152, 2007.
25. T. M. Rath and R. Manmatha. Features for word spotting in historical manuscripts. In *International Conference of Document Analysis and Recognition*, pages 218-222, 2003.
26. T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. *Computer Vision and Pattern Recognition*, (2):521-527, 2003.
27. E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, volume 1, pages 430-443, 2006.
28. M. Rusinol, D. Aldavert, R. Toledo, and J. Lladós. Browsing heterogeneous document collections by a segmentation-free word spotting method. In *11th International Conference on Document Analysis and Recognition (ICDAR'11)*, pages 63-67, China, 2011.
29. S. N. Srihari, Srinivasan H, C. Huang, and S. Shetty. Spotting words in latin, devanagari and arabic scripts. *Indian Journal of Artificial Intelligence*, 16(3):2-9, 2006.
30. Kengo Terasawa and Yuzuru Tanaka. Slit style hog feature for document image word spotting. In *10th International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 116-120, 2009.
31. Carl von Eckartshausen. *Aufschlüsse zur Magie aus geprüften Erfahrungen über verborgene philosophische Wissenschaften und verdeckte Geheimnisse der Natur*. Bavarian State Library, 1778.
32. Ismet Zeki Yalniz and R. Manmatha. An efficient framework for searching text in noisy document images. In *Proceedings of Document Analysis Systems (DAS)*, pages 48-52, 2012.
33. B. Zhang, S. N. Srihari, and C. Huang. Word image retrieval using binary features. In *Document Recognition and Retrieval XI (SPIE)*, pages 45-53, 2004.