Efficient Learning-Free Keyword Spotting

George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos

Abstract—In this article, a method for segmentation-based learning-free Query by Example (QbE) keyword spotting on handwritten documents is proposed. The method consists of three steps, namely preprocessing, feature extraction and matching, which address critical variations of text images (e.g., skew, translation, different writing styles). During the feature extraction step, a sequence of descriptors is generated using a combination of a zoning scheme and a novel appearance descriptor, referred as modified Projections of Oriented Gradients. The preprocessing step, which includes contrast normalization and main-zone detection, aims to overcome the shortcomings of the appearance descriptor. Moreover, an uneven zoning scheme is introduced by applying a denser zoning only on query images for a more detailed representation. This leads to a significant reduction in storage requirements of a document collection. The distance between the query and word sequences is efficiently computed by the proposed Selective Matching algorithm. This algorithm is further extended to handle an augmented set of images originating from a single query image. The efficiency of the proposed method is demonstrated by experimentation conducted on seven publicly available datasets. In these experiments, the proposed method significantly outperforms all state-of-the-art learning-free techniques.

Index Terms—Keyword spotting, query by example, learning-free, gradient orientation descriptor, sequence matching

1 Introduction

Digitization and understanding of documents is of great interest to the computer vision as well as the humanities communities. Over the years, several methods have been developed for processing a document image with the aim of acquiring the underlying text, an area of research known as handwritten text recognition (HTR). Although significant progress has been made on HTR, the problem is far from being solved since in several cases the accuracy of such systems is still low. The main challenges that affect the performance of a HTR system include the variability of different writing styles as well as the irregular layouts which lead to imperfect segmentation.

An alternative to the handwritten text recognition approach is keyword spotting (KWS) which can be defined as the task aiming to retrieve specific words of interest (queries) in a document collection without the need of transcribing every single word of the collection [1]. A KWS system returns a ranked list of the word images with respect to the degree of similarity with the query.

Depending on the input format of the query, we can distinguish two scenarios: *query by example (QbE)* [6], [7], [8], [9] and *query by string (QbS)* [2], [3], [4]. In the QbE scenario the format of the query is an image, while in the QbS scenario the query input corresponds to a text string. In this work, we consider only the QbE KWS scenario.

KWS techniques are also divided into two main categories based on the considered search space: *Segmentation-free* and *Segmentation-based* approaches. Segmentation-free approaches aim to locate the query instances on the whole document without the involvement of a segmentation step [5], [10], whereas segmentation-based approaches assume that a segmentation step has preceded. Segmentation-based techniques can be further divided with respect to the level of segmentation. On the one hand, word-based techniques assume that words are already segmented and focus on comparing word images [8], [11], [12]. These techniques are usually applied after an initial segmentation-like step that aims at fast retrieval of candidate regions of words [7]. On the other hand, line-based techniques assume that a text line segmentation step has been previously applied and their goal is to perform word spotting at text line level [2], [4]. An advantage of the segmentation-based compared to the segmentation-free approaches is the reduced computational cost for processing specific regions of the document (located words) rather than the whole document.

In this work, we assume that the segmented words are provided and our focus is on extracting discriminative information from the word images, as this is the case in several recent works [11], [12]. Word segmentation is out of the scope of the presented work. However, it is a well-studied subject and existing techniques, such as [13], [14] and [7], perform sufficiently well. We should note that a word segmentation technique provides candidate word regions, which may be overlapping, and aims to significantly reduce the search space on the document. These candidate regions do not necessarily have a unique correspondence to the existing words in the document since for each actual word

- *G. Retsinas is with the Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", GR-15310 Athens, Greece, and also with the School of Electrical and Computer Engineering, National Technical University of Athens, GR-15773 Athens, Greece. E-mail: georgeretsi@iit.demokritos.gr.*
- *G. Louloudis, N. Stamatopoulos, and B. Gatos are with the Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", GR-15310 Athens, Greece. E-mail: {louloud, nstam, bgat}@iit.demokritos.gr.*

Manuscript received 17 Oct. 2016; revised 4 Apr. 2018; accepted 13 May 2018. Date of publication 10 June 2018; date of current version 12 June 2019.
(Corresponding author: George Retsinas.)
Recommended for acceptance by C. V. Jawahar.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2845880

in the document there could be multiple possible candidate regions [7].

A different taxonomy of KWS methods is related to the existence of a training phase. Techniques that involve a training step belong to the *learning-based* category whereas methods that work directly on the collection without any prior knowledge apart from the query belong to the *learning-free* category. Learning-based techniques are required for QbS approaches in order to pre-train character models, e.g., HMM models [2], and eventually associate the visual information to characters. These character models can be concatenated in order to form the query string, i.e a word model. Learning-free techniques are usually associated with QbE approaches. However, there are also approaches that perform QbE word spotting using pre-trained models [11], [12]. One of the major goals of the proposed work is to explore the capabilities of a learning-free KWS system since the existence of a training set is rare, mainly due to the difficulties for the annotation of document collections.

Keyword spotting in handwritten documents is considered as a challenging task due to the variability of different writing styles. This challenge is even more noticeable in learning-free approaches since it is expected that out of a single word instance (query) the KWS method should be able to simulate all possible variations and eventually return all relevant appearances of the word across the document collection.

A common system for a segmentation-based QbE KWS task usually consists of the following steps. First, a preprocessing (normalization) step is applied to each already segmented word image and several common variations (skew, translation e.t.c.) are absorbed leading to word images of reduced variability. At a next step, feature extraction is performed in order to create a compact representation of each normalized image. The final step involves the calculation of a similarity between the query image and the word images belonging to the document collection based on the features produced at the previous step. The output of such a system is a ranked list of all the words in the collection which reflects the similarity of each word to the provided query, i.e., words very similar to the query appear at the top of the list. It is evident that the selection of a similarity measure is heavily dependent on the nature of the previously extracted features. In this article, segmented words belonging to the document collection will be referred as word images whereas word samples which are used as queries will be referred as query images.

The proposed work is in-line with the aforementioned QbE KWS scheme. Each step has been designed to handle a different set of variations in handwritten word images. These steps collaborate in order to achieve improved accuracy. At the same time, these steps are designed to be cost-effective since one of our major concerns is the creation of a real-time KWS system which requires low retrieval time.

In more detail, a novel preprocessing step is introduced consisting of contrast and main-zone normalization, which cannot be efficiently addressed during the upcoming proposed feature extraction step. The feature extraction step relies upon a modified version of the Projections of Oriented Gradients (POG) descriptor [8], [15], referred as mPOG, and provides a compact representation which is robust to different writing styles. The proposed descriptor is extracted on successive vertical image segments generated by a zoning scheme along the x-axis, resulting in a sequence of descriptors. Finally, a sequence matching algorithm is applied in order to determine the similarity between sequences of descriptors.

One of the main contributions of this work is the introduction of an uneven zoning scheme between word and query images aiming to reduce the storage requirements of a document collection. This is accomplished by considering a denser zoning procedure only for the query images. An appropriate sequence matching algorithm is required to handle uneven sequences in a suitable way that complies with the constraints of the specific problem formulation. To this end, we also propose an efficient algorithm for sequence matching, referred as Selective Matching (SM).

To further improve the robustness of our method, a query image augmentation scheme is explored with respect to main-zone detection, since the preprocessing step may introduce errors. In more detail, multiple instances of the query are created with respect to a parameter of the preprocessing step and subsequently multiple sequences of descriptors are generated. The similarity between word sequences and the augmented query set of sequences is calculated using an extension of the Selective Matching algorithm, denoted as Multi-Instance Selective Matching (MISM). By adopting the described augmentation approach, we manage to resolve a hard assignment problem (selecting a fixed parameter on the preprocessing step) by retaining several possible instances of the normalized image that are used at the final sequence matching step.

In summary, the main contributions of the proposed approach consist of: (i) the introduction of a novel low-cost descriptor, referred as mPOG; (ii) the application of a zoning method in order to produce a sequence of descriptors. The density of the sequence is different for the case of query and word images; (iii) the use of an efficient matching technique (SM algorithm) able to deal with the above mentioned density difference; (iv) the incorporation of query augmentation by making use of different parameters in the preprocessing step together with a novel matching algorithm (MISM) able to handle multiple augmented query images.

The rest of this article is organized as follows. In Section 2, state-of-the-art methods on QbE Word Spotting are highlighted. The proposed method is described in Section 3, organized in sections corresponding to the main contribution as well as the three main steps i.e., preprocessing, feature extraction and matching. The experimental results are reported in Section 4, highlighting the effectiveness of the proposed method on various datasets. Finally, conclusions and future directions are drawn in Section 5.

2 Related Work

This section briefly summarizes recent works which reported notable results in the QbE keyword spotting scenario. A basic remark concerning the existing bibliography on QbE KWS is that experimental results indicate the superior performance of learning-based approaches compared to learning-free methods. This is in-line with the general notion that the existence of a single query image without prior knowledge for the case of learning-free QbE keyword spotting limits the performance of these methods. Despite their lower performance, such learning-free approaches still attract significant interest mainly due to their simplicity as well as their independence on a specific language or on the existence of a training set. The main research interest of learning-free approaches focuses on the extraction of robust descriptors capable of alleviating most of the challenging variations among words belonging to the same class.

A taxonomy of the QbE keyword spotting methods can be defined based on the output of the feature extraction step. Specifically, they can be categorized to methods that produce either *a feature vector of fixed dimensionality (holistic representation)* or *a set of features*. The subsequent matching procedure is directly dependent on the choice of the feature representation (holistic or set). In more detail, methods belonging to the first category use a simple distance/similarity measure at the matching step (e.g., Euclidean distance). On the contrary, methods that are grouped to the second category require a more sophisticated algorithm for matching.

Methods that assume a holistic representation typically utilize an appearance descriptor, usually based on gradient orientation (e.g., HOG), as in [7], [8] and [10]. In [7], HOG and LBP descriptors are extracted over a binarized image and concatenated into one feature vector after dimensionality reduction, while in [8] an alternative of HOG, the POG descriptor, is used either on the whole word image or on vertical image segments. Both methods use the Euclidean distance at the retrieval step for comparing word feature vectors. Almazan et al. [10] also use HOG descriptors. The similarity between representations is computed using an SVM, which considers translations of the initial query image as positive examples and is trained at query time. Another popular approach is the extraction of such appearance descriptors, usually HOG or SIFT, over image patches and the adoption of a Bag of Visual Words (BoVW) scheme. Due to the loss of spatial information, it is imperative to structure the BoVW histograms into a spatial pyramid. This approach is adopted by Aldavert et al. [21] (HOG descriptors and Euclidean distance), Rusiñol et al. [18] (Integral Histogram of Gradients descriptors and Euclidean distance) and Rothacker et al. [18] (SIFT descriptors and Bray-Curtis distance).

A different approach was proposed by Almazan et al. [11] which involves a training phase, introducing a binary embedding of each word's transcription referred as pyramidal histogram of characters (PHOC). According to [11], visual encodings are extracted from each image (using Fisher Vectors) and along with the corresponding PHOC representations (labels) consist the training set of an attribute learning step which uses a SVM. The final descriptor of fixed dimensionality is generated through a common subspace projection, considering both the trained attributes and the PHOC labeling. Due to the involvement of the word's transcription, this method can be applied to both QbE and QbS scenarios. Based on the novelty of Almazan's method, several recent works make use of PHOC embeddings (labels) along with Deep Convolutional Neural Networks in order to create efficient learning-based techniques [12], [29]. Although these methods are not directly comparable with learning-free methods, their performance can be considered as a comparative measure for the capabilities of a learning-free QbE keyword spotting system. A slightly different approach was adopted by Sfikas et al. [9], since this method does not involve a training phase even though a Deep Convolution Network was used to extract features from the word image. The selected network is pre-trained on an independent set of typewritten characters and considered to give discriminative features for character recognition. The final descriptor is the aggregation of the network's responses over different image segments.

Methods which produce a set of features (second category) are more diverse. A coarse subcategorization can be defined by distinguishing sequences of features and graph-structured features. Several approaches have been presented in the literature which extract sequence of features and subsequently use a sequence similarity algorithm, sush as DTW. The most characteristic example is the work of Rath and Manmatha [6], in which DTW is performed on a sequence of simple geometric (statistical) features, computed at each column of the word image. The column-based sequences are usually long and thus DTW is a time-consuming approach. Considering graph-structured features, Howe [20] proposed a flexible inkball model which is a set of connected nodes (corresponding to text strokes) and allows deformable template matching. Due to this formulation, the main computational effort is shifted on the matching procedure performed by an iterative energy minimization algorithm. A different approach which is based on typical keypoint detection is presented by Zagoris et al. [31]. This approach focuses on i) the detection of appropriate keypoints for text images and ii) the extraction of a gradient-based descriptor for each keypoint. The similarity between images is performed by comparing local descriptors as well as their spatial correlation which may cause significant overhead for a large set of keypoints.

3 Proposed Method

3.1 Contribution

The main goal of the proposed method is to handle the majority of challenges encountered in the QbE scenario, such as affine variations between images and different writing styles. At the same time, these variations should be treated in such a way that resource requirements are retained low. The proposed keyword spotting method consists of three main steps, 1) preprocessing, 2) feature extraction and 3) matching, which are described in detail in the following sections. These steps are strongly related and each one is designed to address a different set of common variations in word images that have been identified as crucial to the system's performance.

Specifically, we propose a novel descriptor, referred as mPOG, which encodes gradient orientation and exhibits robustness to small affine distortions. These variations are common when considering collections containing documents of different writing styles. One drawback of appearance descriptors is their sensitivity to rotation and translation. To this end, the proposed preprocessing step provides rotation and vertical translation invariance by detecting the main-zone of a word image. Furthermore, horizontal variation is addressed using a zoning scheme, resulting to a sequence of descriptors followed by an appropriate sequence matching algorithm. Finally, the sequence matching algorithm is modified in order to deal with multiple instances of the query image aiming to assist the error-prone preprocessing step. The proposed preprocessing step is dependent on the selection of specific parameters that may not perform equally well on different datasets. In order to overcome this shortcoming and avoid any errors at an early step, different instances of the normalized query image are generated for different preprocessing parameter values and the aforementioned problem is resolved using the proposed multiple instance sequence matching algorithm.

The proposed method not only achieves outstanding performance in terms of retrieval accuracy, as it will become evident in the experimental results, but it is also cost-effective with respect to time and memory requirements. One of our major concerns was the minimization of the resource requirements without affecting the method's performance aiming towards a real-time KWS application. We distinguish two main directions in resource optimization: 1) storage of word images (document collection storage) and 2) retrieval time. The main computational advantages of the proposed method are highlighted below:

(i) Each word image is described by a short sequence of descriptors using a zoning technique (e.g., 4-8 zones), reducing the storage cost of the image representation. The produced descriptors are further compressed with the use of a dimensionality reduction technique (PCA).

(ii) Possible extra variations are considered only on the query image (denser zoning or multi-instance generation), avoiding the overhead of computing and storing extra information for each segmented word image of the document collection. Therefore, word images are represented with shorter sequences of descriptors compared to query images. To support this decision, an efficient sequence matching algorithm between sequences of uneven length is introduced using dynamic programming.

(iii) Retrieval time is further reduced by a re-ranking scheme on the retrieved words. Assuming a query image and a set of word images, words are first ranked with the use of a holistic descriptor and the Euclidean distance. Then, a subset of possibly relevant words is selected for the sequence-based algorithm.

3.2 Preprocessing

The preprocessing step consists of two main procedures: contrast and main-zone normalization. It is assumed that the input is a gray-scale word image.

3.2.1 Contrast Normalization

Robustness to illumination changes is a key aspect of several descriptors, e.g., Histogram of Oriented Gradients [16], which is commonly addressed locally by performing an appropriate normalization on a group of neighboring descriptors. However, document images are simpler in this aspect since the text is supposed to be black (or dark) colored letters on a white page. Based on this observation, many document image processing tasks (e.g., text-line segmentation, text recognition) rely on binarization as a preprocessing step. Nevertheless, binarization is error prone due to the hard assignment of a pixel to either foreground or background.

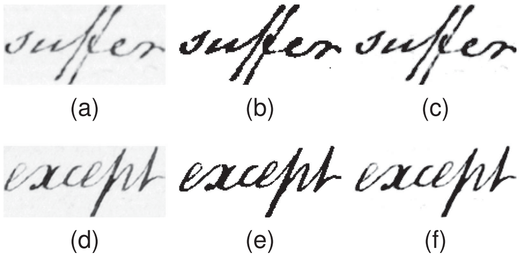


Fig. 1. Contrast-normalization examples: (a),(d) initial grayscale images, (b),(e) Sauvola's binarization, (c),(f) proposed contrast normalization.

The aforementioned problems are surpassed by replacing the hard assignment of the binarization with a soft assignment scheme. In more detail, starting from Sauvola's binarization technique [23], a membership function is defined and applied to each pixel (see Equation (1)) resulting to the contrast-normalized image I_{cn}. Each pixel with illumination value outside the interval $[a(x,y), b(x,y)]$ is considered either foreground or background. We define $a(x,y) = t(x,y) - 1.5s(x,y)$ and $b(x,y) = t(x,y) + 0.3s(x,y)$,[1] where $I(x,y)$ is the initial image, $t(x,y)$ is the Sauvola's method threshold and $s(x,y)$ the standard deviation at each pixel. Mean value, standard deviation and the corresponding threshold are computed as in [23], assuming a neighborhood of interest which is specified by a window of fixed size.

$$I_{cn}(x,y) = \begin{cases} 0, & I(x,y) \leq a(x,y) \\ \frac{I(x,y)-a(x,y)}{b(x,y)-a(x,y)}, & a(x,y) < I(x,y) \leq b(x,y) \\ 1, & I(x,y) > b(x,y). \end{cases} \quad (1)$$

Examples of the contrast-normalization step are presented in Fig. 1 in comparison with Sauvola's binarization, when the same parameters are used. It can be observed that the normalized images retain the useful relative contrast changes on the foreground while at the same time enhance the contrast between foreground and background pixels. Furthermore, despite the fact that binarization introduces boundary discontinuities, it is obvious that the proposed contrast-normalization step enhances the edge contrast without affecting the edge boundaries. This property has proven to be essential for the extraction of informative gradient-based descriptors, which is the next step of the proposed method.

3.2.2 Main-Zone Normalization

The goal of this procedure is to detect the main-zone of the word. Once the main-zone has been detected, the following normalizations can be applied: 1) slope correction (deskew) 2) vertical centralization 3) ascenders and descenders cropping.

Main-zone detection is often a crucial step in text image normalization. The main-zone of a text image (word or line) is defined by an upper and a lower boundary, often referred as upperline and baseline. Under the assumption that these boundaries are linear (or piece-wise linear), regression techniques have been successfully applied [8], [24]. Alternatively, the upper and lower boundaries of the main zone can be arbitrarily detected using trained models as in [25] and [26].

[1] The choice of a, b adjusts an interval $[a, b]$ with its endpoint b close to the Sauvola's threshold t and its length proportional to each pixel's standard deviation s.

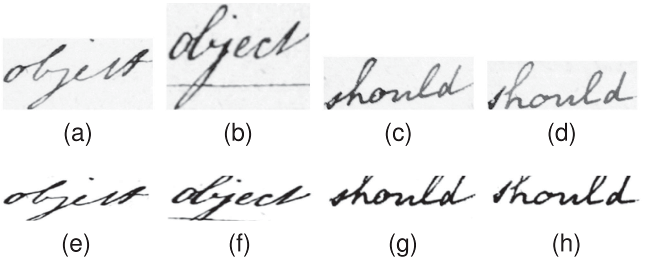


Fig. 2. Different instances of the words "object" and "should", containing vertical size variation, before (a),(b),(c),(d) and after (e),(f),(g),(h).

In this work, the boundaries are considered linear and the main-zone detection is achieved by processing the horizontal projections which correspond to the distinct angles: $\theta \in [-8°, -7°, \ldots, 7°, 8°]$. Horizontal projections are generated using the Radon transform instead of rotating the image, which introduces an extra interpolation step, in order to reduce the computational cost. Given a horizontal projection $P_\theta(i)\,, i = 1, .., L$ corresponding to a specific angle θ, our goal is to find a lower and an upper bound (a and b, respectively) such that the sum of the enclosed projection values is maximized ($\sum_{i=a}^{b} P_\theta(i)$), while the height of the main zone ($b - a$) is minimized. This is achieved by maximizing the criterion of Equation (2), in which the contribution of the two terms is controlled by a regularization parameter r. The problem of finding the main zone is therefore formulated as a problem of finding the maximum contiguous subarray, which can be solved in linear time. The angle with the maximum value J_θ defines the slope of the main zone.

$$J_\theta = \max_{a,b}\left\{\frac{\sum_{i=a}^{b} P_\theta(i)}{\sum_{i=1}^{L} P_\theta(i)} - r\frac{b-a}{L}\right\}. \tag{2}$$

The robustness of the detection is further improved by:

(i) Replacing the length L with L' corresponding to the 95 percent of horizontal projection's information, formulated as in Equation (3). The upper and lower boundary that define L' are referred as u and l, respectively.

$$L' = u - l \quad \text{s.t.} \quad \frac{\sum_{i=l}^{u} P_\theta(i)}{\sum_{i=1}^{L} P_\theta(i)} \approx 0.95. \tag{3}$$

This choice ideally eliminates variations due to the segmentation of the words which may result in the inclusion of parts of neighboring words.

(ii) Introducing an adaptive regularization parameter r which depends on the distribution of the horizontal projection as described in Equation (4). The reasoning behind the definition of r is that words with significantly large ascenders or descenders should obtain a higher value, thus leading to a narrower main zone.

$$r = \frac{\sum_{i=l}^{u} P_\theta^2(i)}{\left(\sum_{i=l}^{u} P_\theta(i)\right)^2}. \tag{4}$$

After the detection of the main-zone, we proceed with skew correction using the slope θ of the detected main-zone, as well as vertical normalization of the image by moving the main zone at the center of the generated normalized image. Additionally, in order to avoid extreme ascenders and descenders, which contain no useful information, the images are cropped (or padded) to a fixed height which is proportional to the main-zone's height. Specifically, given the height of the main zone $h_m = b - a$, the resulting normalized image has a margin of $1.5 \times h_m$ pixels under and over the main zone, i.e., the resulting image has $4 \times h_m$ overall height. The results of the complete preprocessing step are depicted in Fig. 2. It is evident that many of the undesired variations have been reduced. Although appearance descriptors (e.g HOG or mPOG) are robust to small (affine) deformations, they cannot cope with the large variations that are frequently encountered in word images (rotations and translations). Thus, it is imperative to use the proposed preprocessing step before the extraction of an appearance descriptor.

Experimentation indicates that the most critical factor affecting the performance of the proposed keyword spotting method concerns main-zone detection. Even though the proposed preprocessing increases the robustness of main-zone detection, the automatically generated regularization parameter r may result to a poor detection and affect the upcoming steps and consequently the overall system's performance. Instead of expecting to extract precise main-zones, which ideally requires a machine learning algorithm and a training dataset [25], [26], one can assume a set of n_l different regularization parameters and consequently a set of n_l generated query image instances. Starting from the regularization parameter $r_0 = r$ of Equation (4), we define a set of n_l regularization parameters as $r_i = p_i r_0$, where p_i values are uniformly distributed in the interval $[0.6, 1.4]$, i.e $p_i = 0.6 + i\frac{0.8}{n_l}$ $(i = 1, \ldots, n_l)$. In other words, we propose an augmentation scheme in order to cope with possible variations that cannot be addressed effectively at the preprocessing step (see Fig. 5). The retrieval task for the augmentation case (multiple query instances) is performed as it is described in Section 3.4.2.

3.3 Feature Extraction

The main objective of the feature extraction step is to efficiently encode useful and discriminative information of a word image. The proposed feature extraction consists of a zoning scheme along with the extraction of a local descriptor for each zone. The result of the aforementioned step corresponds to a sequence of local descriptors.

A modification of Projections of Oriented Gradients, referred as mPOG, was chosen as the descriptor of the zoned image segments. The POG descriptor was first introduced as a descriptor for the character classification task in our previous work [15] and later applied to keyword spotting with promising results, as described in [8]. The POG descriptor is a projection based encoding and was conceived as an alternative to the HOG descriptor. Essentially, they both encode the same initial information, which is the spatial distribution of strokes or, more precisely, gradient orientation. However, HOG extracts gradient orientation information over image segments, while POG encodes the same information on the whole image using projections.

The main difference of the proposed mPOG descriptor compared to our previous works [8], [15], is the extension of the descriptor in order to be applicable to grayscale images. The gradient information is represented using a set of

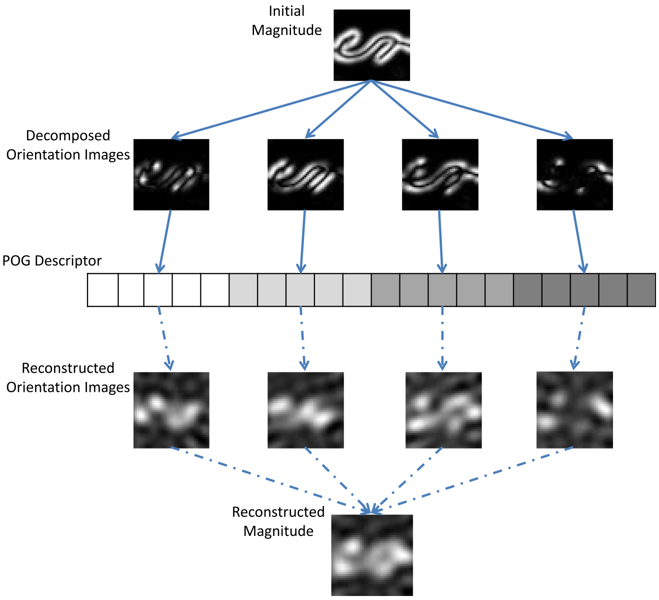

Initial Magnitude
Decomposed Orientation Images
POG Descriptor
Reconstructed Orientation Images
Reconstructed Magnitude

Fig. 3. Overview of the proposed mPOG descriptor extraction applied on an image part together with the corresponding reconstruction. The reconstructed orientation images visualize/highlight the stored information in each orientation.

intensity images, referred as *orientation images*, which correspond to different gradient orientations. This representation, depends not only on the gradient magnitude of each pixel but also on the deviation of the pixel's gradient orientation from the representation's central orientation. Subsequently, the proposed descriptor is constructed by encoding the projections of each orientation image. Furthermore, for the efficient encoding of the generated projections their complex Fourier coefficients are used as features. The Fourier coefficients are represented by their absolute, real and imaginary values, while a $L2$-norm normalization is performed on each projection's encoding. On the contrary, the POG descriptor consists only of unnormalized real and imaginary values of the Fourier coefficients.

In the following sections, the proposed mPOG descriptor is described, emphasizing on the modifications with respect to our previous method [15]. Finally, the zoning procedure is presented in detail.

3.3.1 Modified Projections of Oriented Gradients (mPOG)

The steps of the mPOG feature extraction are briefly described below:

1) Orientation Images. Gradient orientation, which describes edge orientation, has proven to be a very informative feature (e.g., HOG, SIFT e.t.c.); hence the proposed descriptor follows a similar concept by representing gradient orientation information through orientation images. Orientation images are defined as intensity images that represent the gradient information corresponding to a specific range of gradient orientations, as depicted in Fig. 3.

The gradient information is represented by the gradient magnitude M as well as the gradient orientation Φ of the image, using polar coordinates, at each pixel (Equations (5) and (6)). A wrapping is performed so that all orientation values lie on the interval $[0, 180^\circ)$.

$$M(x,y) = \sqrt{(I_x^2(x,y) + I_y^2(x,y))} \tag{5}$$

$$\Phi(x,y) = arctan\left(\frac{I_y(x,y)}{I_x(x,y)}\right). \tag{6}$$

An orientation image I_ϕ, given the central orientation/direction ϕ that characterizes it, is the result of a pixel-based function on the gradient magnitude M and the gradient orientation Φ. The orientation dependency of this function is modeled by a Gaussian weighting function w (see Equation (7)) in order to represent how close an orientation is compared to ϕ. The ϕ value corresponds to the mean value μ while the acceptable orientation interval corresponds to the standard deviation σ of the Gaussian function. The resulting weights for each pixel are referred as orientation weights.

$$w(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{7}$$

Given a number of possible orientations n_ϕ, the orientation range σ_ϕ and the central orientation values ϕ_k are defined as in Equation (8). The resulting orientation images I_{ϕ_k} are generated by multiplying the orientation weights with the gradient magnitude of the image according to Equation (9).

$$\sigma_\phi = \frac{180^\circ}{n_\phi}, \quad \phi_k = k\sigma_\phi, \quad k \in [0, n_\phi - 1] \tag{8}$$

$$I_{\phi_k}(x,y) = w(\Phi(x,y); \phi_k, \sigma_\phi) M(x,y). \tag{9}$$

Orientation images correspond to a decomposition of the magnitude M with respect to the orientation ϕ, i.e., $\sum_{i=1}^{n_\phi} I_\phi \approx M$. The weighting function takes into consideration overlapping orientation ranges, aiming to generate orientation images with smoother strokes in case of edge orientation changes. On the contrary, our previous implementation of POG for binary images involved a hard assignment of pixels in orientation ranges which resulted in discontinuities.

2) Projections. Each orientation image is decomposed into several projections under selected angles by applying Radon transform. This projection-based approach is selected in order to represent (orientation) images in a more holistic manner compared to the HOG encoding (spatial cells). Assuming that the number of projections is n_θ, the projections angles $\{\theta_k\}$ are sampled every $180^\circ/n_\theta$

$$\theta_k = k\frac{180^\circ}{n_\theta}, \quad k \in [0, n_\theta - 1]. \tag{10}$$

3) FFT & Coefficient Selection. Each projection is simplified and eventually encoded by selecting only the low frequency components of the projection, which correspond to a smooth approximation of the projection and retain information about regions with high pixel concentration. Therefore, after computing the Discrete Fourier Transform coefficients $c_j, j \in [0, K-1]$ of the projection, only the first n_c are used to form the projection's descriptor, excluding c_0. Subsequently, each coefficient is normalized by c_0, which is equal to the total number of foreground pixels. The extracted feature vector of a projection is the concatenation of the real,

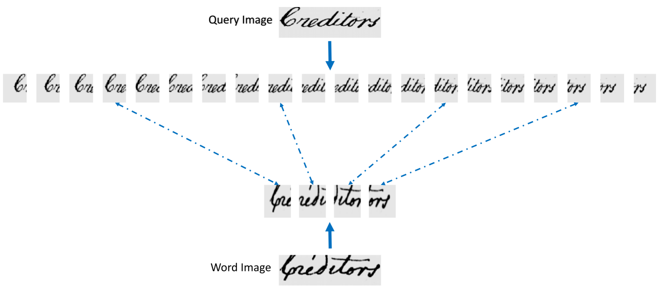


Fig. 4. Visualization of the matching procedure between a query and a word image using different zoning schemes ($n_w = 4$ & $n_d = 5$).

imaginary and absolute values of the complex feature vector $f_j = c_j/c_0, j = 1, \ldots, n_c$.

4) *Final Descriptor*. The final descriptor is the concatenation of the generated coefficients of every projection and every orientation image. Overall, the length of the image descriptor is: n_ϕ (images) × n_θ (projections) × $3n_c$ (Fourier coefficients). Prior to the concatenation, a $L2$-norm normalization is performed to each projection's feature vector of coefficients.

One interesting property of the proposed descriptor is that each orientation image can be approximately reconstructed via the inverse Radon transform, after interpolating each projection to a specific length using the inverse Fourier transform. As a result, a normalized approximation of the image is constructed which provides an informative and helpful visualization of the remaining information. The initial decomposed orientation images and the reconstructed orientation images from the extracted descriptor are presented in Fig. 3. The reconstructed magnitude is visualized as the sum of reconstructed orientation images. We can observe that the descriptor preserves the useful edge information by displaying higher illumination values in the reconstructed images on stroke regions.

3.3.2 Zoning

It is observed (see the experimental results section) that the application of the mPOG descriptor on the entire word image (holistic mPOG) has encouraging performance. However, by keeping a fixed number of coefficients in each projection, we can only retain a uniform distribution of the edge information on each projected direction. Therefore, the usage of a holistic mPOG descriptor results in loss of useful information (e.g., the same amount of information is considered for both vertical and horizontal projections).

Motivated by the successful application of zoning techniques on KWS [8], [9], we propose to uniformly split the image along the x-axis into overlapping segments. In this way, a sequence of descriptors is created. As it was mentioned before, one of the main contributions of this work is the application of a different zoning scheme for query and word images. We assume that possible translation variations on the horizontal axis (e.g., different spacing between letters or a horizontal displacement of the word) could be captured by denser sampling either the query or the word image and not necessarily both. Therefore, a simple zoning procedure is used on word images, assuming n_w zones, while, on the contrary, a denser zoning is performed on query images, assuming $n_w \times n_d$ zones, where n_d is an integer denoting the number of dense samples for each zone. The width of the zones is the same for both cases. The proposed uneven zoning aims to reduce processing and storage requirements for document collections by shifting the extra computational cost to the query. An example of the proposed uneven zoning procedure is shown in Fig. 4.

In practice, simple zoning of the word images should be scarce, i.e $n_w \in [4, 8]$, which results in a cost-effective matching. This observation is also in line with the fact that the mPOG descriptor performs very well on character level [15] and therefore n_w should take values close to the average number of characters in a word.

3.4 Matching of Descriptor Sequences

3.4.1 Single Query Matching

Given the word sequence (reference sequence) $t_i,\ i = 1, , n_w$ and the query sequence $x_j,\ j = 1, , n_w n_d$, our goal is to assign each t_i descriptor to a single x_j descriptor of the query sequence, as shown in Fig. 4. The matching procedure results in a subset $S_m = \{x_{j_1}, x_{j_2}, \ldots, x_{j_{n_w}} : j_{i-1} < j_i\}$ of the query sequence descriptors that best match the reference sequence descriptors. The aforementioned matching approximates an ideal zoning scheme, where each selected query zone is optimally located with respect to the corresponding word zone alleviating possible horizontal translations.

Assuming that no inner (local) horizontal variations exist, the difference of the indexes of neighboring matches should be equal to the denser zoning parameter n_d. This strict assumption allows only global horizontal translation invariance. However, in order to capture local horizontal variations, a constraint on relative distance between neighboring matches (at the query sequence) is imposed, i.e., between j_{i-1} and j_i. This constraint penalizes large variations in x-axis, assuming that the "proper" distance of consecutive matches is $j_i - j_{i-1} \approx n_d$. To this end, the quadratic penalty function is defined as

$$w(k, l; a) = \begin{cases} 1 + p(k - l - a)^2, & \text{if} \quad k > l \\ \infty, & \text{otherwise} \end{cases}, \tag{11}$$

where a penalty, which depends on a scaling parameter p and has values over 1, is determined by comparing the distance $k - l$ to the pre-specified reference distance a.

Due to the quadratic form of the penalty function the meaningful displacement is restricted to a neighborhood of radius b around the "proper" translation, as it is defined in Equation (12).

$$w(k, l; a, b) = \begin{cases} 1 + p(k - l - a)^2, & |k - l - a| < b \\ \infty, & \text{otherwise} \end{cases}. \tag{12}$$

In practice, we assume $b = a/2$ and the penalty function will be used as $w(k, l; a) = w(k, l; a, a/2)$. The scaling parameter p is selected in such a way that the maximum penalty value is 1.2, i.e., $w(k, l; a, b) = 1.2$ when $|k - l - a| = b = a/2$. Thus, p is defined as: $p = 0.2(1/b^2) = 0.8(1/a^2)$.

The best matching subsequence problem can be formulated as a minimizing function over all possible ordered subsets S_m, where $|S_m| = n_w$, as shown in Equation (13).

$$score = \min_{S_m} \Big\{ \|x_{j_1} - t_1\|_2 + \sum_{i=2}^{n_w} w(j_i, j_{i-1}; n_d) \|x_{j_i} - t_i\|_2 \Big\}. \tag{13}$$

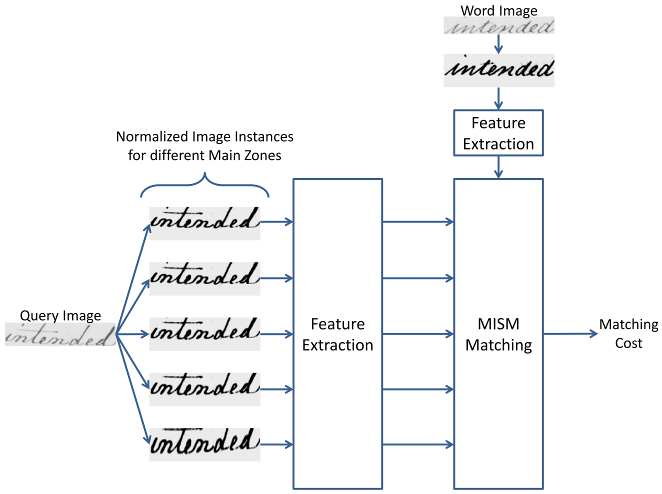


Fig. 5. Overview of the multi-instance selective matching for the case of multiple main zone detection.

The above formulation indicates that the matching procedure can be efficiently implemented using dynamic programming, avoiding to explore all possible ordered subsets S_m which has an exponential complexity. In more detail, given the distance matrix D, where $D_{ij} = \|x_j - t_i\|_2$ (D is of size $n_w \times n_w n_d$), the overall matching cost is computed by iteratively adding the best match of t_i to the best score so far based on Equation (13). A matching matrix M is used to store the solutions (best matchings) of the intermediate subproblems. Each M_{ij} value denotes the best score considering the subproblem of matching the subsequence $t_1, t_2, \ldots, t_i$, while t_i is constrained to match with x_j. The matrix M is constructed step by step according to the update function of Equation (14), utilizing incrementally the score of smaller subsequences. The last row of M consists of possible matching scores for the whole sequence $\{t_i\}$ and thus the overall best matching score is obtained as the minimum value of this row. We refer to this algorithm, which is described in Algorithm 1, as Selective Matching. It can be easily deduced by the description of the proposed algorithm that its time complexity is $O(n_w^2 n_d^2)$.

$$M[i,j] = \min_{\Omega} \left\{ M[i-1,k] + w(j,k;n_d) D[i,j] \right\}$$
$$\Omega = \left\{ k \in [1, n_w n_d] \, , \, |k - j - n_d| \, < \, \frac{n_d}{2} \right\}. \tag{14}$$

Algorithm 1. Description of Selective Matching Algorithm

```
procedure SELECTIVEMATCHING($D$)
  Input: Distance Matrix, $D$ ($n_w \times n_w n_d$)
  Output: Best Matching Score, $score$
  for $j = 1$ to $n_w n_d$ do
    $M[1,j] = D[1,j]$
  for $i = 2$ to $n_w$ do
    for $j = 1$ to $n_w n_d$ do
      Update $M[i,j]$ using Equation (14)
  return $score = \min_j\{M(n_w, j)\}$
```

3.4.2 Matching of Multiple Instances

As it was mentioned on the preprocessing Section 3.2.2, an augmentation scheme is proposed with respect to the main-zone detection in order to cope with variations that cannot be efficiently addressed at the preprocessing step. Therefore, we decided to propagate the possible variations generated by the main-zone detection step to the matching step. According to the proposed augmentation scheme, n_l different query instances are created through the use of a set of n_l different values for the regularization parameter for the main-zone normalization step (see Section 3.2.2).

A modified version of the SM algorithm, namely Multi-Instance Selective Matching, is proposed in order to properly handle the augmented set. The final matching set of descriptors may originate from different query instances in order to increase the robustness of the proposed method. The concept is essentially the same with the simple SM algorithm and therefore has a straightforward implementation using dynamic programming according to the Equation (15). Contrary to the SM algorithm, the matching matrix M as well as the distance matrix D have three dimensions ($n_w \times n_l \times n_w n_d$) since the n_l different image instances form the second dimension of the matrices. The complexity of the new matching procedure is the multiplication of the simple SM algorithm's complexity with the number of different instances n_l, i.e., $O(n_l n_w^2 n_d^2)$. An overview of the MISM is depicted in Fig. 5.

$$M[i,l,j] = \min_{\Omega} \left\{ M[i-1,m,k] + w(k,j;n_d) D[i,l,j] \right\}$$
$$\Omega = \left\{ k \in [1, n_w n_d] \, , \, m \in [1, n_l] \, , \, |k - j - n_d| \, < \, \frac{n_d}{2} \right\}. \tag{15}$$

It should be noted that the idea of multi-instance query generation/augmentation is not limited to main-zone detection and could be applied to any possible variation which cannot be addressed in a typical and efficient way at the preprocessing or the feature extraction step, e.g., local affine transformations on each segmented zone.

4 Experimental Results

4.1 Experimental Setup

The proposed method is developed for the task of segmentation-based learning-free QbE KWS. The implementation of the proposed method is publicly available.[2] Initially, a parameter tuning is performed on a simple setup of the George Washington (GW) dataset.[3] Subsequently, the proposed method is evaluated for different setups of the GW dataset as well as the IAM dataset.[4] Finally, the proposed method is compared to the state-of-the-art methods that either participated or reported results on the KWS competitions of ICFHR 2014 (Bentham14 and Modern14 datasets) [17], ICDAR 2015 (Bentham15 dataset) [18] and ICFHR 2016 (Botany16 and Konzils16 datasets) [19]. The performance of the word spotting methods is recorded in terms of the Precision at Top 5 Retrieved words (P@5) as well as the Mean Average Precision (MAP). The experiments were performed on an 8-core Intel i7-4770 K at 3.50 GHz with 16 Gb of RAM.

In order to be comparable with the results reported in the bibliography, we follow the same evaluation protocol for

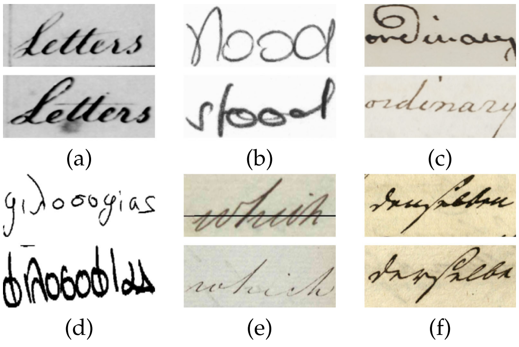
[2] https://github.com/georgeretsi/Learning-Free-KWS
[3] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database
[4] http://www.fki.inf.unibe.ch/databases/iam-handwriting-database

TABLE 1
Properties of the Datasets/Setups Used for Evaluation

Dataset/Setup	#words	#queries	qr
GW1 [9]	4,860	306	yes
GW2 [21]	4,860	1,847	no
GW3 [11]	1,215	901	no
IAM [11]	13,752	4,030	no
Bentham14 [17]	10,370	320	yes
Modern14 [17]	14,754	300	yes
Bentham15 [18]	13,657	1,421	no
Botany16 [19]	3,230	150	yes
Konzils16 [19]	3,534	200	yes

each dataset/setup. Besides the selection of the evaluation metrics for each dataset (e.g., only MAP or both MAP and P@5), we also take into account whether the query image is considered relevant to itself. This information (denoted as the *qr* property), along with the number of queries and segmented words for each dataset are summarized in Table 1. Examples of segmented words from each collection are shown in Fig. 6. Two different instances of the same word are presented in order to highlight the variations of the aforementioned collections. The GW dataset is the only one which is described in detail since three different setups are used for evaluation.

George Washington Dataset. This dataset is the well-known collection of writings of George Washington, consisting of 20 pages segmented into 4,860 words. A standard partition or query selection is not available for GW dataset and subsequently the majority of the reported results are not comparable. To this end, we distinguish three different setups. The first one is used only for parameter selection whereas the other two for comparison. *GW1.* Fifteen words selected as in [9] and all instances of each of the 15 words were considered as queries (306 queries in total). *GW2.* Words with ten or more instances and three or more characters are selected as queries as in [21], resulting in 1847 queries. Several recent works on learning-free QbE KWS report results on this setup [7], [21], [31]. *GW3.* A fourfold cross validation setup is employed, using the exact same partitions as in [11]. Words with at least two instances in the test sets are selected as queries. This setup is adopted by the majority of the recent learning-based methods, since it is split to training and testing partitions. Table 1 reports the average number of queries and words over the four different partitions.



(a) (b) (c)

(d) (e) (f)

Fig. 6. Two indicative examples of the same word in each dataset: (a) George Washington, (b) IAM, (c) Bentham, (d) Modern, (e) Botany and (f) Konzilsprotokolle.

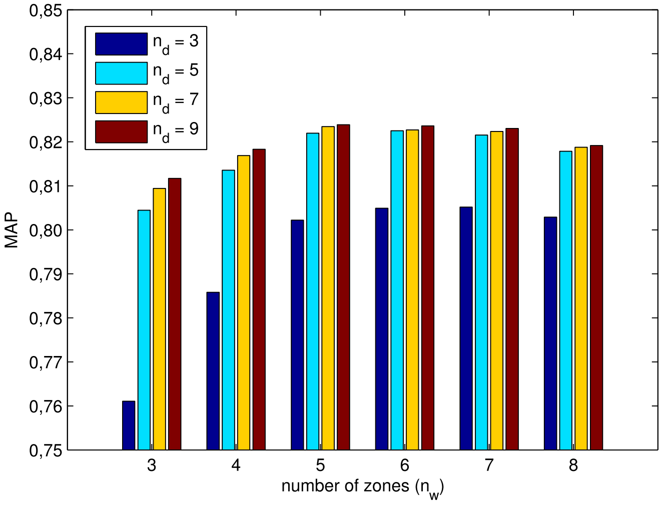


Fig. 7. Performance evaluation of the mPOG descriptor and the selective matching procedure on GW1 for different zoning parameters.

We should highlight the fact that no further parameter exploration has been performed on the datasets. The values of the parameters have been fixed using the GW1 setup and were kept constant throughout the evaluation stage of the method on all other datasets/setups. Such an experimental scheme not only enables the proposed method to be directly comparable to the methods that participated on the KWS competitions but also confirms the generalization of the proposed method.

4.2 Zoning Parameter Selection

We experimented on the impact of the zoning procedure to the system's performance by changing the number of word segments n_w and dense query segments n_d using the GW1 setup. The parameters of the mPOG descriptor are fixed to $n_\phi = 4$, $n_\theta = 6$ and $n_c = 7$. For this experiment, only the case of SM is considered. The performance results are presented in Fig. 7. It can be observed that increasing parameter n_d leads to better performance. This complies with our perception that the denser the sampling, the higher the probability of correctly localizing the corresponding segment. Since both feature extraction and matching/retrieval time depend on the product $n_w \times n_d$, we concluded that the choice $n_w = 6$ and $n_d = 5$ provides results close to the best performance, while retaining low time requirements. These parameter values are considered as the default values for the remaining experimental section both for SM and MISM (augmentation scheme) approaches.

4.3 Comparing Variations of the Proposed Method

In order to highlight the effectiveness of the mPOG descriptor as well as the zoning and sequence matching procedures, we evaluate the performance of the following variations of the proposed pipeline using the GW1 setup.

Holistic Descriptor. The simplest version of the proposed *mPOG* descriptor is to apply it on the whole image as a global descriptor. An alternative descriptor is the widely-used Felzenwalb's variation of Histograms of Oriented Gradients (*fHOG*) [27] which serves as a competitive descriptor to the mPOG. These two descriptors along with the Euclidean distance (denoted as *Holistic-mPOG+Eucl* and *Holistic-*

TABLE 2
Evaluation Metrics on GW1 Dataset

Method	P@5	MAP
Almazan et al. [10]	88.6	61.8
Retsinas et al. [8]	93.0	70.2
Sfikas et al. [9]	92.4	71.1
Holistic-fHOG + Eucl	90.3	65.8
Seq-fHOG + Eucl	89.5	64.3
DSeq-fHOG + Eucl	90.3	64.8
PSeq-fHOG + DTW	91.7	71.6
DSeq-fHOG + DTW	96.4	81.1
PSeq-fHOG + SM	96.3	81.5
PSeq-fHOG + MISM	96.1	83.2
Holistic-mPOG + Eucl	93.0	71.7
Seq-mPOG + Eucl	94.0	74.9
DSeq-mPOG + Eucl	91.7	69.3
PSeq-mPOG + DTW	89.3	71.6
DSeq-mPOG + DTW	96.5	82.2
PSeq-mPOG + SM	97.3	83.7
PSeq-mPOG + MISM	**98.0**	**87.7**

fHOG+Eucl, respectively) provide a baseline performance to be compared with the proposed sequence-based approach.

Sequence of Descriptors. Sequence of descriptors, either *fHOG* or *mPOG*, is extracted from each word image using a zoning scheme. In order to further clarify the extraction process of descriptor sequences, we distinguish three zoning schemes. *(a) Seq*: standard zoning on both word and query images (n_w zones) *(b) DSeq*: dense zoning on both word and image queries ($n_w \times n_d$ zones) *(c) PSeq*: standard zoning on word images and dense zoning on query images (proposed). The proposed method (*PSeq-mPOG+MISM*) belongs to the third category and the compared variations based on uneven zoning are denoted as *PSeq-fHOG+SM*, *PSeq-mPOG+SM*, *PSeq-fHOG+MISM*. Due to the sequential nature of the proposed method, the widely used DTW matching algorithm is also evaluated on both uneven-zoned (*PSeq-fHOG+DTW* and *PSeq-mPOG+DTW*) and dense-zoned sequences (*DSeq-fHOG+DTW* and *DSeq-mPOG+DTW*) for comparison. Finally, the *Seq* and *DSeq* schemes are evaluated using only the Euclidean distance, aiming to underline the importance of a sequence matching algorithm on top of a zoning scheme.

The parameters of the mPOG descriptor (n_ϕ, n_θ and n_c) are already defined in Section 4.2, resulting to a feature vector of 504 dimensions. Concerning the fHOG descriptor, a 6×6 cell grid is used with 5 possible (unsigned) orientations (720 dimensions). The parameters for both descriptors are selected using a grid search approach. The number of generated query instances, used in the MISM approach, is set to $n_l = 7$. Furthermore, after extracting the descriptor sequences of the images, PCA is performed over local descriptors to reduce the length of each descriptor to 60.

The performance on the GW1 setup of the aforementioned variations along with the state-of-the-art methods of Almazan et al. [10], Retsinas et al. [8] and Sfikas et al. [9], is presented in Table 2. The main observations are summarized below:

(i) *Sequence-based approaches outperform holistic approaches.* The gain in performance between Holistic-mPOG +Eucl and PSeq-mPOG+MISM (over 15 percent MAP) highlights the effectiveness of the proposed method. Moreover, this gap in performance is also observed between PSeq-mPOG+MISM and the state-of-the-art holistic methods presented in [8] and [9], which have similar performance to the holistic mPOG descriptor. The necessity of a sequence matching algorithm is verified by comparing the performance of Euclidean matching on descriptor sequences (Seq-mPOG+Eucl, DSeq-mPOG+Eucl) with sequential matching approaches over descriptor sequences (DSeq-mPOG+DTW, PSeq-mPOG+SM and PSeq-mPOG+MISM). The latter category achieves a noticeable increase in performance, as it handles horizontal translations of the image. On the contrary, using the Euclidean distance on standard zoning (Seq), or even dense zoning (DSeq), may result in the accumulation of errors from miss-aligned zones.

(ii) *mPOG outperforms fHOG.* The mPOG descriptor consistently outperforms fHOG supporting our claim that the proposed descriptor is more robust.

(iii) *Comparison with DTW.* Even though DTW could be applied to uneven sequences, such an approach would not produce reliable scores when it is applied on the proposed uneven zoning. Specifically, distances concerning translated image segments would be accumulated into the score value. Therefore, due to the descriptor's sensitivity to translation, the calculated score would include many error terms related with the same image segment. Furthermore, when applying DTW either to dense (DSeq) or uneven (Pseq) sequences, an early mismatch in the unconstrained sequence alignment may propagate a significant error into the final score. On the contrary, the proposed matching algorithm (SM) is consistent with the formulation and constraints of the specific problem. The aforementioned analysis is verified by the experimental results.

(iv) *Importance of augmentation.* The proposed method PSeq-mPOG+MISM which includes the augmentation scheme for multiple main-zone detection, provides a noteworthy boost in performance compared to the PSeq-mPOG-SM version. The success of the PSeq-mPOG+MISM method emphasizes the significance of a robust main-zone detection.

4.4 Time and Memory Requirements

A simple strategy for the reduction of the retrieval time is the application of the sequence matching approach on a subset of the word images. To this end, based on the successful application of the holistic descriptors, the Holistic-mPOG +Eucl version is considered as the first step of a re-ranking procedure comprising the following steps:

1) Perform a typical retrieval scheme using Euclidean distance on the holistic descriptor.
2) Select a subset of the best retrieved words, i.e., with the smallest Euclidean distance, and consequently re-calculate the matching cost of the reduced set using the SM (or MISM) algorithm on the descriptor sequences.

Table 3 presents the impact on the average retrieval time (per query) as well as the retrieval performance with respect

TABLE 3
Average Retrieval Time (Per Query) - Performance Trade-Off for the Proposed Re-Ranking Procedure (Using PSeq-mPOG+SM on GW1)

Used Words (%)	Time (sec)	P@5(%)	MAP(%)
0	0.0062	93.0	71.7
5	0.0322	96.8	81.5
10	0.0584	96.9	82.2
15	0.0846	96.9	82.6
20	0.1109	97.1	82.9
25	0.1373	97.1	83.1
40	0.2145	97.2	83.5
60	0.3114	97.3	83.6
100	0.5237	97.3	83.7

to the percentage of selected words for re-ranking. This experiment was conducted only for the case of PSeq-mPOG +SM method. Note that using 0 percent of words is equivalent to using only the Holistic-mPOG+Eucl approach. It is clear that the retrieval time has linear dependence on the percentage of the selected subset. However, retrieval performance exhibits no significant changes for percentage values greater than 10 percent. At 10 percent we achieve a speed-up close to ×9 and approximately 1 percent drop in MAP compared to applying the PSeq-mPOG+SM method directly on the entire word image collection. Hence, 10 percent is considered as the default value for the percentage of used words in the upcoming experiments.

Time and memory requirements with respect to feature extraction (including the pre-processing step), retrieval time as well as storage requirements are presented in Table 4. The reported retrieval time per query refers to the re-ranking scheme, i.e., it consists of the retrieval time for comparing the holistic descriptors between the query and the words, as well as the retrieval time for comparing a query sequence with the top 10 percent relevant word sequences. It is worth mentioning that the feature extraction procedure was performed on parallel using 4 cores. As it has been already mentioned, one of the main contributions of this work is to shift the extra information (dense zoning) at query level in order to retain the memory requirements of storing a document collection low. This is indicated in Table 4, where each document (≈ 250 words) requires only 0.5 MB including both PSeq-mPOG and Holistic-mPOG descriptors, whereas a dense sequence approach (DSeq) requires n_d times more storage. The retrieval time, which is the most important factor for a real-time KWS application, is noticeable low compared to the performance gain of the PSeq-mPOG+SM method. As it was expected, the multi-instance modification performs better at the cost of a slower retrieval response (the response difference is proportional to the number of generated instances). Nevertheless, the retrieval time of the proposed method is still sufficient low for a real-time KWS application.

TABLE 4
Resource Requirements for Sequential mPOG Features on GW1

Resource Requirements	PSeq+SM	PSeq+MISM
Extraction Time per Word	0.018 sec	0.018 sec
Extraction Time per Query	0.126 sec	0.659 sec
Memory per Document (after PCA)	525.8 KB	525.8 KB
Retrieval Time per Query	0.058 sec	0.294 sec

TABLE 5
MAP Evaluation on (a) GW2 and (b) GW3 & IAM Datasets

(a)

Method	**GW2**
Wang et al. [33]	17.5
Retsinas et al. [8]	39.1
Zagoris et al. [32]	40.1
Zagoris et al. [22]	40.5
Almazan et al. [10]	48.3
Sfikas et al. [9]	60.3
Kovalchuk et al. [7]	66.3
Zagoris et al. [31]	69.2
Aldavert et. al [21]	76.5
Holistic-fHOG+Eucl	65.1
DSeq-fHOG+DTW	75.2
PSeq-fHOG+SM	75.7
PSeq-fHOG+MISM	77.6
Holistic-mPOG+Eucl	66.3
DSeq-mPOG+DTW	76.8
PSeq-mPOG+SM	79.2
PSeq-mPOG+MISM	**81.1**

(b)

Method	**GW3**	**IAM**
Retsinas et al. [8]	37.0	15.2
Almazan et al. [10]	49.4	-
Sfikas et al. [9]	58.3	13.2
DTW [11]	60.6	12.3
FV [11]	62.7	15.7
Holistic-fHOG+Eucl	63.8	18.7
DSeq-fHOG+DTW	72.4	23.9
PSeq-fHOG+SM	72.7	23.3
PSeq-fHOG+MISM	74.5	24.7
Holistic-mPOG+Eucl	63.8	18.8
DSeq-mPOG +DTW	72.3	26.9
PSeq-mPOG+SM	74.0	27.6
PSeq-mPOG+MISM	77.1	28.1
Almazan et al.* [11]	93.0	55.7
Softmax CNN* [12]	78.2	48.7
Finetuned CNN* [34]	-	46.5
Sudholt et al.* [12]	**96.7**	72.5
Krisnan et al.* [30]	94.8	80.6
Krisnan et al.* [29]	94.4	**84.2**
Sudholt et al.** [12]	74.9	3.4

(*) : train and test on the same dataset, (**) : train and test on different datasets.

4.5 Experimental Evaluation on GW & IAM Datasets

Having concluded on the parameter values (Sections 4.2 & 4.4), we proceed to the evaluation of the proposed method using the remaining GW setups (GW2 and GW3) as well as the IAM dataset. In order to further explore the effectiveness of our method, we also report the performance of both descriptors (fHOG and mPOG) using the variations presented in Section 4.3. The experimental results on the GW2 dataset are presented in Table 5a, while in Table 5b we report the results on both GW3 and IAM datasets. Learning-based methods are also included in Table 5b, even though they cannot be directly compared to learning-free methods. We distinguish two evaluation scenarios for learning-based methods: 1) train and test data originate from the same dataset (denoted with⋆) and 2) train and test data originate from different datasets (denoted with**). The second scenario is used for a better comparison between learning-based and learning-free methods since no fine-tuning occurs. The main observations are presented below:

(i) *The proposed method outperforms all learning-free approaches.* The proposed method (PSeq-mPOG+MISM) significantly outperforms any other learning-free method (including the examined variations) on all datasets/setups. Notably, the proposed method surpasses recent works as [31] and [21], reported at GW2 setup, by a significant extent (11.9 and 4.6 percent, respectively). In addition, the retrieval results on both tables support our previous claim that the proposed mPOG descriptor provides superior performance over the fHOG descriptor. Since both descriptors cannot cope with horizontal translations, they have similar performance when they are used as holistic descriptors. However, the robustness of the mPOG

TABLE 6
Evaluation Metrics for Bentham and Modern Datasets on ICFHR14 Competition

	Bentham14		Modern14	
Method	P@5	MAP	P@5	MAP
Kovalchuk et al. [7]	73.8	52.4	58.8	33.8
⋆ Almazan et al.** [11]	72.4	51.3	70.6	**52.3**
Howe [20]	71.8	46.2	56.9	27.8
Zagoris et al. [32]	52.5	34.1	-	-
Zagoris et al. [22]	62.3	39.3	-	-
Aldavert et al. [21]	62.9	46.5	61.9	38.9
Sfikas et al. [9]	76.4	53.6	56.0	32.1
Retsinas et al. [8]	77.1	57.7	61.3	35.5
Zagoris et al. [31]	78.8	60.0	-	-
Holistic-mPOG+Eucl	77.5	60.4	52.3	29.7
PSeq-mPOG+SM	85.1	70.2	71.4	46.8
PSeq-mPOG+MISM	**85.5**	**71.1**	**73.5**	49.1

(**) : train and test on different datasets. (⋆) : competition winner.

TABLE 7
Evaluation Metrics for Bentham Dataset on ICDAR15 Competition

Method	P@5	MAP
⋆ PRG [18]	46.0	42.4
CVC [18]	34.3	30.0
Zagoris et al. [32]	22.4	19.3
Zagoris et al. [22]	26.8	21.7
Almazan et al.** [11]	41.7	36.3
Sfikas et al. [9]	47.0	41.5
Retsinas et al. [8]	48.7	44.5
Zagoris et al. [31]	50.1	44.0
Holistic-mPOG +Eucl	48.1	44.2
PSeq-mPOG+SM	59.5	56.4
PSeq-mPOG+MISM	**61.6**	**58.4**

(**) : train and test on different datasets. (⋆) : competition winner.

descriptor is evident when used as a local descriptor for sequence-based approaches.

(ii) *Success of learning-based approaches.* Learning-based techniques achieve remarkable results, reporting a significant increase in performance compared to learning-free techniques. This increase in performance is expected due to the existence of a training phase, during which the generated model is adjusted to the existing writing styles. In order to achieve the reported performance, learning-based methods require a considerable amount of available training data (training sets consist of 3645 and 30226 word images for GW3 and IAM datasets, respectively, without including possible augmentations).

(iii) *Generalization of learning-based approaches.* An important observation is that the results reported in the literature are dataset-oriented, i.e., the generalization of the created models is not explored or supported. This fact implies that if a model is trained and tested on different datasets (e.g., trained on IAM and tested on GW dataset), it may lead to a significant drop in performance, as shown in [28]. This observation is also verified in our work using the PHOCNet system [12] (Sudholt et al.**) trained on the IAM dataset and tested on GW3 and vice versa. As it was expected, the model trained on the simple GW dataset shows very poor performance on the challenging IAM dataset. On the contrary, the model trained on the IAM dataset, which includes many writing styles, appears to be more robust. Nevertheless, the gap in performance when compared with the model trained in GW3 is significant (nearly 20 percent), while it performs slightly worse compared to the proposed method. In conclusion, even though learning-based approaches are very successful, their generalization to unseen data coming from different collections needs to be further explored.

4.6 Experimental Evaluation on Competition Datasets

We further explore the generalization and efficiency of the proposed method on several competition datasets. Tables 6, 7 and 8 report the results concerning the ICFHR14, ICDAR15 and ICFHR16 competition datasets, respectively. Besides the proposed method (PSeq-mPOG+MISM), we also evaluate Holistic-mPOG+Eucl and PSeq-mPOG+SM variations in order to highlight the increase in performance when considering sequence matching and augmentation procedures. The competitions' participants are reported at the first part of each table, annotating the competitions' winners using the ⋆ symbol on the left side of the method's name. State-of-the-art methods which used the competition datasets for measuring their performance are reported on the second part of each table. The main observations are summarized below:

(i) *The proposed method outperforms all learning-free approaches.* Holistic-mPOG+Eucl approach performs reasonably well for the majority of the datasets, producing results comparable to state-of-the-art learning-free techniques. Nevertheless, sequential approaches (PSeq-mPOG+SM and PSeq-mPOG+MISM) demonstrate an outstanding boost in performance on all datasets. For example, the performance increase on the Modern14 dataset is over 17 percent on both metrics. Another remark is that, although the MISM approach provides a considerable gain in performance, this is not consistent among all datasets. The reason of this inconsistency is correlated with the degree of success of the main-zone detection. Overall, the proposed

TABLE 8
MAP Evaluation on ICFHR16 Competition Datasets

Method	Botany16	Konzils16
CVCDAG* [19]	75.8	77.9
⋆ PRG* [19]	**89.7**	**96.1**
QTOB* [19]	55.0	82.2
TAU [19]	50.6	71.1
Retsinas et al. [8]	46.7	56.5
Sfikas et al. [9]	46.5	59.9
Holistic-mPOG+Eucl	53.2	64.2
PSeq-mPOG+SM	57.0	71.1
PSeq-mPOG+MISM	58.3	76.2

(*) : train and test on the same dataset. (⋆) : competition winner.

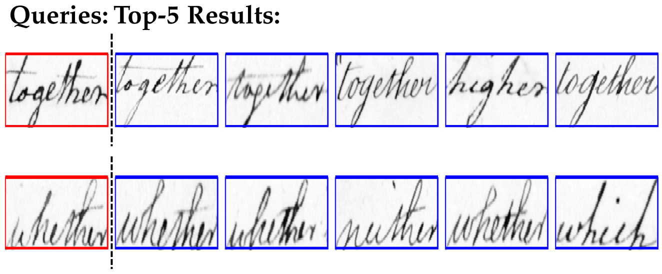


Fig. 8. Two cases of erroneous retrieval: (Top row) 80 percent P@5 (b) 60 percent P@5. Query is separated from the top 5 retrieved words with a dotted line. As it can be observed, PSeq-mPOG+MISM retrieves a set of words that have almost identical prefix or suffix.

method achieves the best performance among every published learning-free method by a notable extent.

(ii) *Learning-based methods trained and tested on different datasets (ICFHR14 and ICDAR15 competitions).* The attribute-based method [11], proposed by Almazan et al., was the competition winner and still outperforms all existing KWS methods on the Modern14 dataset. Even though this method uses a training step, the training was performed on independent datasets that include similar writing styles to the competition's datasets (GW and IAM datasets were selected for the Bentham14 and Modern14 datasets, respectively). The involved training step proved to be advantageous in the case of the Modern14 dataset, which consists of different languages and writers. However, the proposed method surpasses the attribute-based method in terms of P@5 on the Modern14 dataset. Furthermore, it achieves a gain over 20 percent on both Bentham datasets for which method [11] does not perform equally well.

(iii) *Learning-based methods trained and tested on the same dataset (ICFHR16 competition).* Learning-based methods, trained and evaluated on the same collection, report superior performance over learning-free methods, as it can be observed in Table 8 which summarizes the results for the ICFHR16 competition's datasets. This competition allowed participants to use different amount of training data, as an attempt to examine the generalization of their approaches. In addition, the training data was not completely manually annotated, i.e., the word segmentation step is performed automatically. For further details of the competition setup see [19]. It should be noted that the competition winner (*PRG*) used all the available training data. Even though the learning-based methods that participated on this competition are not directly comparable to our learning-free approach, there are cases for which the proposed method achieves similar results to the participants' methods (QTOB on Botany16 and CVCDAG on Konzils16).

4.7 Error Analysis

Even though the proposed method outperforms the existing learning-free keyword spotting techniques, there are cases for which it does not produce adequate results. Two indicative examples are shown in Fig. 8. This erroneous behavior is related to the calculated cost between two sequences (a weighted summation of the distance of zoned segments). Therefore, images of different words with identical segments may have lower cost compared to images of the same word in different writing styles.

5 Conclusions and Future Work

This paper proposes a new learning-free approach for the segmentation-based QbE KWS task. Performance-wise, we focus on assisting the proposed descriptor (mPOG) and addressing its shortcomings in the context of the KWS task. To this end, we have introduced a preprocessing step of critical normalizations, such as contrast and main-zone normalization. At the same time, we deal with possible horizontal translations by constructing sequences of descriptors (extracted on vertical image zones) and calculating their similarity by a sequence matching algorithm. Resource-wise, all steps are designed to be cost-effective. Moreover, storage reduction is achieved by extracting denser sequences of descriptors only from query images. In order to calculate a distance measure between these uneven descriptor sequences a dynamic programming algorithm for sequence matching is proposed. Finally, an augmentation approach with respect to main-zone normalization is adopted in order to handle cases of erroneous main-zone detection.

The proposed method is evaluated on seven publicly available datasets and achieves outstanding performance outperforming all learning-free techniques. Moreover, it reports similar results to learning-based techniques that have been trained and evaluated on independent datasets.

Regarding future work, we plan to explore the enrichment of the augmentation set by including possible variations which are not efficiently addressed in the proposed method such as affine deformations and thinning/thickening of the stroke width. Finally, concerning resources optimization, we aim to further compress the generated descriptors by applying nonlinear (manifold) embedding methods as well as feature quantization techniques.

Acknowledgments

This work has been supported by the EU project READ (Horizon-2020 programme, grant Ref. 674943).

References

[1] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, "A survey of document image word spotting techniques," *Pattern Recognit.*, vol. 68, pp. 310–332, 2017.

[2] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 934–942, 2012.

[3] A. H. Toselli and E. Vidal, "Fast HMM-filler approach for key word spotting in handwritten documents," in *Proc. Int. Conf. Document Anal. Recognit.*, 2013, pp. 501–505.

[4] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 211–224, Feb. 2012.

[5] L. Rothacker and G. A. Fink, "Segmentation-free query-by-string word spotting with Bag-of-Features HMMs," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 661–665.

[6] T . M. Rath and R. Manmatha, "Word spotting for historical documents," *Int. J. Document Anal. Recognit.*, vol. 9, no. 2–4, pp. 139–152, 2007.

[7] A. Kovalchuk, L. Wolf, and N. Dershowitz, "A simple and fast keyword spotting method," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2014, pp. 3–8.

[8] G. Retsinas, G. Louloudis, N. Stamatopoulos, and B. Gatos, "Keyword spotting in handwritten documents using projections of oriented gradients," in *Proc. Workshop Document Anal. Syst.*, 2016, pp. 411–416.

[9] G. Sfikas, G. Retsinas and B. Gatos, "Zoning aggregated hypercolumns for keyword spotting," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2016, pp. 283–288.

[10] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient exemplar word spotting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.

[11] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2552–2566, Dec. 2014.

[12] S. Sudholt and G. A. Fink, "PHOCNet: A deep convolutional neural network for word spotting in handwritten documents," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2016, pp. 277–282.

[13] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," *Pattern Recognit.*, vol. 42, no. 12, pp. 3169–3183, 2009.

[14] R. Manmatha and J. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1212–1225, Aug. 2005.

[15] G. Retsinas, B. Gatos, N. Stamatopoulos, and G. Louloudis, "Isolated character recognition using projections of oriented gradients," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 336–340.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Comput. Vis. Pattern Recognit.*, vol. 2, pp. 886–893, 2005.

[17] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos, "ICFHR 2014 competition on handwritten keyword spotting (H-KWS 2014)," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2014, pp. 814–819.

[18] J. Puigcerver, A. H. Toselli, and E. Vidal, "ICDAR2015 competition on keyword spotting for handwritten documents," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 1176–1180.

[19] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A. H. Toselli, and E. Vidal, "ICFHR2016 handwritten keyword spotting competition (H-KWS 2016)," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2016, pp. 613–618.

[20] N. R. Howe, "Part-structured inkball models for one-shot handwritten keyword spotting," in *Proc. Int. Conf. Document Anal. Recognit.*, 2013, pp. 582–586.

[21] D. Aldavert, M. Rusiñol, R. Toledo, and J. Lladós, "A study of bag-of-visual-words representations for handwritten keyword spotting," *Int. J. Document Anal. Recognit.*, vol. 18, no. 3, pp. 223–234, 2015.

[22] K. Zagoris, I. Pratikakis, and B. Gatos, "Segmentation-based historical handwritten word spotting using document-specific local features," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2014, pp. 9–14.

[23] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 224–236, 2000.

[24] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[25] S. Espana-Boquera, M. J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, "Improving offline handwritten text recognition with hybrid HMM/ANN models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 767–779, Apr. 2011.

[26] J. Pastor-Pellicer, S. Espana-Boquera, M. J. Castro-Bleda, and F. Zamora-Martinez, "A combined convolutional neural network and dynamic programming approach for text line normalization," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 341–345.

[27] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010

[28] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Handwritten word spotting with corrected attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1017–1024.

[29] P. Krishnan, K. Futta, and C. V. Jawahar, "Deep feature embedding for accurate recognition and retrieval of handwritten text," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, 2016, pp. 289–294.

[30] P. Krishnan and C. V. Jawahar, "Matching handwritten document images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[31] K. Zagoris, I. Pratikakis, and B. Gatos, "Unsupervised word spotting in historical handwritten document images using document-oriented local features," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4032–4041, Aug. 2017.

[32] K. Zagoris, K. Ergina, and N. Papamarkos, "Image retrieval systems based on compact shape descriptor and relevance feedback information," *J. Visual Commun. Image Representation*, vol. 22, no. 5, pp. 378–390, 2011.

[33] P. Wang, V. Eglin, C. Garcia, C. Largeron, J. Llados, and A. Fornes, "A novel learning-free word spotting approach based on graph representation," in *Proc. Workshop Document Anal. Syst.*, 2014, pp. 207–211.

[34] A. Sharma and K. P. Sankar, "Adapting off-the-shelf CNNs for word spotting & recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 986–990.

George Retsinas received the diploma from the School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), in 2014. Currently, he is working toward the PhD degree at the same university while working with the Institute of Informatics and Telecommunications in the National Center for Scientific Research "Demokritos" in Athens, Greece. His research work is mainly focused on image processing, machine learning and document image analysis.

Georgios Louloudis received the bachelor's degree in Informatics and Telecommunications, in 2000, the master's and PhD degrees from the Department of Informatics and Telecommunications, UoA, in 2002, 2009, respectively. He is currently working as a research associate at NCSR "Demokritos". His main research interests are in pattern recognition, image processing and document image analysis (http://users.iit.demokritos.gr/~louloud/).

Nikolaos Stamatopoulos received the computer science diploma, in 2006 and the PhD degree from the Department of Informatics and Telecommunications, UoA, in 2011. His PhD thesis is on optical process and analysis of historical documents. He is working as research associate with the NCSR "Demokritos". His main research interests are in pattern recognition, image processing and document image analysis (http://users.iit.demokritos.gr/~nstam/).

Basilis Gatos received the electrical engineering diploma and the PhD degree, both from the Electrical and Computer Engineering Department, Democritus University of Thrace, Xanthi, Greece. He is currently working as a researcher with the Institute of Informatics and Telecommunications of the NCSR "Demokritos", Athens, Greece. His main research interests are in image processing and document image analysis, OCR and pattern recognition (http://users.iit.demokritos.gr/~bgat/).