ISSN 1213-6972 Journal of WSCG http://www.wscg.eu/ Vol.26, 2018, No.1

Radial Line Fourier Descriptor for Historical Handwritten Text Representation

Anders Hast and Ekta Vats
Department of Information Technology
Uppsala University
SE-751 05 Uppsala, Sweden
anders.hast@it.uu.se; ekta.vats@it.uu.se

ABSTRACT

Automatic recognition of historical handwritten manuscripts is a daunting task due to paper degradation over time. Recognition-free retrieval or word spotting is popularly used for information retrieval and digitization of the historical handwritten documents. However, the performance of word spotting algorithms depends heavily on feature detection and representation methods. Although there exist popular feature descriptors such as Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF), the invariant properties of these descriptors amplify the noise in the degraded document images, rendering them more sensitive to noise and complex characteristics of historical manuscripts. Therefore, an efficient and relaxed feature descriptor is required as handwritten words across different documents are indeed similar, but not identical. This paper introduces a Radial Line Fourier (RLF) descriptor for handwritten word representation, with a short feature vector of 32 dimensions. A segmentation-free and training-free handwritten word spotting method is studied herein that relies on the proposed RLF descriptor, takes into account different keypoint representations and uses a simple preconditioner-based feature matching algorithm. The effectiveness of the RLF descriptor for segmentation-free handwritten word spotting is empirically evaluated on well-known historical handwritten datasets using standard evaluation measures.

Keywords

Radial Line Fourier descriptor, word spotting, feature matching

1 INTRODUCTION

Automatic recognition of poorly degraded handwritten text is challenging due to complex layouts and paper degradations over time. Typically, an old manuscript suffers from degradations such as paper stains, faded ink and ink bleed-through. There is variability in writing style, and the presence of text and symbols written in an unknown language. This hampers the document readability, and renders the task of searching a word in a set of non-indexed documents i.e. word spotting, to be more difficult.

In literature [Gio17], word spotting approaches can either be segmentation-based where the search space consists of a set of segmented word images, or segmentation-free with the complete document image in the search space. This paper focuses on segmentation-free word spotting, which is typically preferred over segmentation-based methods when dealing with heavily degraded document images [Zag17]. However, the performance of word spotting algorithms significantly depends on the appropriate selection of feature detection and representation methods [Gio17]. In general, feature descriptors represent a region with distinct feature in a document image, coded into a numerical feature vector, which is subsequently compared with the feature vector of a reference image to perform matching.

Efforts have been made in the recent past towards research on feature detection and representation methods. Some popular methods include Scale Invariant Feature Transform (SIFT) [Low04], Speeded Up Robust Features (SURF) [Bay08] and Histograms of oriented Gradients (HoG) [Dal05]. SIFT and HoG contributed significantly towards the progress of several visual recognition systems in the last decade [Gir14]. However, these local descriptors were mainly designed for the representation of natural scene images, that possess structurally different characteristics from the document images. For example, the detection of the most important edges using pyramid scaling in SIFT creates local interest points between the text lines [Zag17]. The invariant properties of these descriptors amplify the noise in the degraded document images, rendering them

https://doi.org/10.24132/JWSCG.2018.26.1.4

more sensitive to noise and complex characteristics of historical manuscripts [Zag17]. The work by [Ley09] analyzed that the rotation-invariant features are more sensitive to noise in a document image, and perform poorly as compared to rotation-dependent features.

Since the existing descriptors are found to be unsuitable for representing handwritten text with high levels of degradations [Zag17, Ley09], it is important to design a descriptor to address this issue. This paper introduces a Radial Line Fourier (RLF) descriptor which is tailor-made for word spotting applications with fast feature representation and robustness to degradations. RLF is a fast and short-length feature vector of 32 dimensions, based on log-polar sampling followed by computing a few elements of the Discrete Fourier Transform (DFT) along each radial line. It does not require any orientation information from the feature detectors, and simple feature detectors can be used without compromising the descriptor and word spotting performance.

This paper is organized as follows. Section 2 reviews the state-of-art methods used in word spotting pipeline, with main focus on interest point detection and feature representation methods. Section 3 presents the proposed method based on the RLF descriptor for segmentation-free handwritten word spotting. Section 4 demonstrates the efficacy of the proposed method on well-known historical datasets using standard evaluation measures. Section 5 concludes the paper.

2 RELATED WORK

Appropriate selection of interest points (keypoints) and feature descriptors is indispensable for the performance of a word spotting system. This section discusses some popular interest point detection and feature representation methods with reference to word spotting systems. It is important to note that the segmentation-free word spotting framework presented herein is training-free, therefore training-based methods such as deep learning are not considered.

2.1 Interest Point Detection

Feature detection, or interest point detection refers to finding keypoints in an image that contain crucial information. There exist several interest point detectors in literature. For example, the Harris corner detector [Har88] is popularly used for corner points detection. It computes a combination of eigenvalues of the structure tensor such that the corners are located in an image. Shi-Tomasi corner detector [Shi94] is a modified version of Harris detector. The minimum of two eigenvalues is computed and a point is considered as a corner point if this minimum value exceeds a certain threshold. The Maximally Stable Extremal Regions (MSER) [Mat02] detector detects keypoints such that all pixels inside the extremal region are either darker or brighter than all the outer boundary pixels.

Typically, interest point based feature matching is performed by using a single interest point detector type. SIFT and SURF are the most popular detectors that capture the blob type of features in the image. SIFT uses the Difference of Gaussians (DoG) that computes the difference between Gaussian blurred images using different values of σ, where σ defines the Gaussian blur from a continuous point of view. SURF computes the Determinant of the Hessian (DoH) matrix, that defines the product of the eigenvalues. In principle, any combination of different keypoint detectors can be selected depending upon the application. This work uses a combination of four types of keypoint detectors for handwritten text representation, that consists of corner detectors, dark and bright blobs, saddle points, and the edges of text strokes.

2.2 Feature Representation

After a set of interest points has been detected, a suitable representation of their values has to be defined to perform word matching. In general, a feature descriptor is constructed from the pixels in the local neighborhood of each interest point. Fixed length feature descriptors are most commonly used that generate a fixed length feature vector, which can be easily compared using standard distance metrics (e.g. the Euclidean distance). Sometimes, fixed length feature vectors are computed directly from the extracted features without the need of a learning step [Gio17].

Gradient-based feature descriptors tend to be superior, and include SIFT [Low04], HoG [Dal05] and SURF [Bay08] descriptors. The 128-dimensional SIFT descriptor is formed from histograms of local gradients. SIFT is both scale and rotation invariant, and includes an intricate underlying framework to ensure this. Similarly, HoG computes a histogram of gradient orientations in a certain local region. An important difference between SIFT and HoG is that HoG normalizes the histograms in overlapping blocks, and creates a redundant expression. SURF descriptor is generally faster than SIFT, and is created by concatenating Haar wavelet responses in sub-regions of an oriented square window. SIFT and SURF are invariant to both scale and rotation changes. There are several variants of these descriptors that have been employed for word spotting [Rod08, Gio17].

Many feature descriptors use local image content in square areas around each interest point to form a feature vector [Has16]. Both scale and rotation invariance can be obtained in different ways [Gau11]. The Fourier transform has been used to compute descriptors that is illumination and rotation invariant, and scale-invariant to a certain extent [Car02, Car03]. In order

to overcome dimensionality issues that may arise in a high-dimensional space, binary descriptors are introduced that are faster, but less precise, for example the Binary Robust Invariant Scalable Keypoints (BRISK) descriptor [Leu11] and Fast Retina Keypoint (FREAK) descriptor [Ala12].

However, these descriptors with strict invariance properties are not suitable for handwritten document representation. This is mainly because the invariance property renders them more sensitive to noise in a degraded document, as has been carefully studied in [Zag17, Ley09].

A method for searching handwritten Arabic documents based on a set of binary shape features is presented in [Sri05], where a correlation distance based matching technique has been employed. However, it was argued by [Gau11] that the features that are dependent on word shape characteristics are not effective in dealing with multi-writer document collections. Instead, the texture information in a spatial context is considered more reliable than the shape information, as suggested in [Gau11, Lla12].

In [Ley07], the image zones representing the most informative parts in a document image are detected based on the gradient orientation computed by taking convolution of the image with the first and second derivatives of the Gaussian kernel. However, this method was found to be inefficient for short words with less than four characters, and therefore an improved version was proposed in [Ley09]. The feature matching algorithm in [Ley09] was found to be very sensitive to variations in handwriting and font sizes, and the overall matching process was too slow for processing large datasets. An interesting block-based document image descriptor was presented by [Gat09] where the query image was scaled and rotated to produce different word instances, and for each instance, a different set of feature vectors was computed. However, several versions of queries generated significant amount of noise in the final merging state, rendering the method inefficient for handling large writing style and font variations.

Inspired by Bag-of-Visual Words (BoVW) model, a patch-based framework that uses SIFT for local feature representation was presented in [Rus11]. The codebook generation step of BoVW model is expensive, and this method is also found to be unsuitable for handling query font size and handwriting variations [Zag17]. The performance of popular word descriptors in a BoVW context was evaluated in [Lla12], and it was suggested that the statistical BoVW approach generates the best result, but with significant increase in overhead in terms of memory requirements to store the descriptors.

The winning algorithm, [Kov14], for segmentation-free track of ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14], employed HoG and Local Binary Patterns (LBP) descriptors, and the word retrieval is performed using the nearest neighbour search, followed by a simple oppression of extra overlapping candidates. The work by [Zag17] outperformed the winning algorithms from ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14], and ICDAR2015 Handwritten Keyword Spotting Competition [Pui15]. They proposed a new approach towards handwritten word spotting, where the spatial information representing the current location of a feature point is taken into account, and is based on the texture information. However, it is unclear how well this method performs in challenging cases where a a word shares several letters with other different words. The RLF descriptor based method proposed herewith handles this issue by dividing a word into several parts (depending upon the size of the word) to eliminate false-positives, and perform reliable keypoint-based feature matching.

The performance of different features for word spotting applications was evaluated using Dynamic Time Warping (DTW) [Rod08] and Hidden Markov Models (HMMs) [Rod09]. It was found that the local gradient histogram features outperform other geometrical or profile-based features. These methods generally match features from evenly distributed locations over normalized words where no nearest neighbor search is necessary. This is because each point in a word has its corresponding point in some other word located in the very same position. Recently, a method based on feature matching of keypoints derived from the words was proposed [Has16], which requires a nearest neighbor search. In this case, a relaxed descriptor is required that is not over-precise, since the handwritten words are not normalized. This is due to complex characteristic of handwritten words, unlike simple Optical character recognition (OCR) text. Handwritten words across different documents are similar, but not identical due to variability in writing styles.

In an endeavor to address the issues discussed above, this work proposes the RLF descriptor, which is tailor-made for handwritten words representation. The main highlights of this work are as follows: (a) a segmentation-free and training word spotting approach is studied; (b) the proposed method uses a combination of different keypoint detectors to capture different characteristics in a handwritten document, which consists of both lines, corners and blobs; (c) the RLF descriptor is designed, which is a fast and short-length feature vector of 32 dimensions with several advantages; (d) a simple preconditioner-based feature matching algorithm is presented. Advantages of RLF descriptor include faster word spotting (due to short length of feature vector), robustness to degradations, flexibility to be employed with existing feature detectors, efficient memory utilization, and no increase

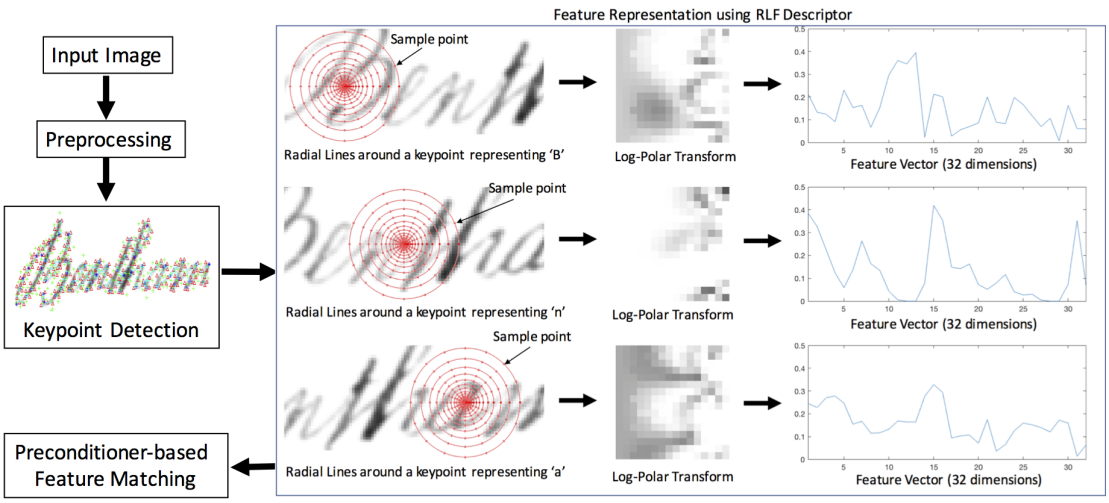

Figure 1: Radial Line Fourier (RLF) descriptor for feature representation in a word spotting framework. Each keypoint detected is represented using log-polar sampling scheme with 16 sampling points per ring. Each radial line, originating in the center and traversing each ring, is used to obtain a square (16 x 16) transformed image representation. In the next step, DFT is applied along each row (corresponding to radial lines) to compute the amplitude of a few elements for each row that constitute the feature vector. Finally, the feature vector generated is presented where x-axis denotes the feature vector length (i.e. 32), and y-axis denotes the amplitude of DFT.

in overhead for feature orientation estimation. The proposed methodology is discussed as follows.

3 METHODOLOGY

The pipeline of the word spotting framework is as follows. For an input document image, preprocessing is performed to remove background noise using two band-pass filtering approach [Vat17]. This is followed by keypoints detection, feature representation using RLF descriptor, and preconditioner-based feature matching. The framework of the proposed approach is pictorially described in Figure 1.

3.1 Preprocessing

Preprocessing is the initial step of the word spotting algorithm where the background noise is removed using a simple two band-pass filtering approach, as proposed in [Vat17]. A high frequency band-pass filter is used to separate the fine detailed text from the background, and a low frequency band-pass filter is used for masking and noise removal. The background removal is performed in such a way that the gray-level information crucial for the feature extraction is not affected. This allows the keypoint detector and the RLF descriptor to be more informative.

3.2 Keypoint Detection

To begin with, keypoints are detected for the document image and the query word. A combination of four different types of keypoint detectors is used to capture a variety of features that represent a handwritten document, and consists of lines, corners and blobs. Figure 2 presents the keypoint detectors used herein using an example image of a smoothed query word, *Bentham*. *Blue* * represents the Harris corner detector [Har88], *green* + represents the result of using the square of the Determinant of Hessian (DoH), which captures both dark and bright blobs, *red* Δ represents negative of DoH (-DoH) and finds the saddle points, and *cyan* + represents the result of an edge detector ($Assymetric^2$) [Has14b].

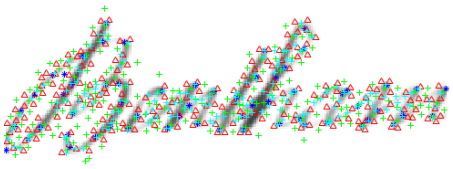

Figure 2: An example of a query word *Bentham* depicting four different types of keypoints. Blue * is a corner detector, green + finds the dark and bright blobs, red Δ finds the saddle points, and cyan + finds the edges of the text strokes.

3.3 Radial Line Fourier Descriptor

Radial Line Fourier (RLF) descriptor is a short-length feature vector of 32 dimensions, presented in this work for representation of handwritten words. RLF is inspired from a variant of Scale Invariant Descriptor (SID) [Kok08], known as SID-Rot [Tru13], and the idea is to perform log-polar sampling in a circular neighborhood around each keypoint. SID is a scale and rotation invariant descriptor, whereas SID-Rot is scale-invariant but rotation-sensitive descriptor. Typically, the Fourier transform can be applied over scales only to obtain a scale-invariant and rotation-dependent

descriptor, or a rotation-invariant and scale-sensitive descriptor. The Fourier transformation over scales render the SID-Rot to be rotation-sensitive, and the scale invariance is achieved by sampling over a large radius with a descriptor length of 3360. This method works well in representing natural scene images with scale changes and no rotations. However, strict invariance properties amplify noise in degraded document images [Zag17], and may lead to loss of useful information. Therefore, a relaxed feature descriptor, such as RLF, is required.

RLF descriptor computes a feature vector representation of an image feature, and is based on log-polar sampling followed by computing a few elements of the DFT along each radial line. It characterizes an image region as a whole using a single feature vector of fixed size, and no learning step is involved. Figure 1 presents the general framework of the RLF descriptor for feature representation in a word spotting pipeline, and discussed in detail as follows.

After the keypoints representing a document image have been detected, log-polar sampling is performed at each keypoint, where each radial line (going from the center, traversing each ring around the center along a line) is transformed into a square representation, as highlighted in Figure 1. The log-polar transform resampling resolution is set to 16 sample points per ring to obtain a square (16 x 16) transformed image. When sampling is done in a log-polar fashion, certain interpolation is required as the pixel coordinates are seldom in the center. One could for instance use a bilinear interpolation to achieve higher accuracy. In this work, sub-pixel sampling is computed using the Gaussian interpolation in a 3x3 neighborhood. In the next step, DFT is used to compute the amplitude of a few elements that constitute the feature vector.

The Fast Fourier Transform (FFT) performed efficiently in [Has14a] for creating descriptors that are relaxed. However, it was found to be impractical for high level applications with large amount of data [Has16]. This is because the FFT is rather slow in computations, such as computing the distance measures (i.e. phase correlation). In general, the FFT requires $O(N\,log(N))$ computations for a discrete series $f(n)$ with N elements. Therefore, this work improves and simplifies the computations needed to generate a faster feature representation, still benefiting from the advantages of the Fourier transform. We propose to use just a few elements from DFT of the sampled elements $f(n)$ along the radial line, and the computation required (using Euler's formula) is

$$\mathscr{F}[f(n)](k) = \sum_{n=0}^{N-1} f(n)\cos(2\pi nk/N) - i(f(n)\sin(2\pi nk/N)). \tag{1}$$

The value of k determines the frequency used to compute the Fourier element, where $k \in 0,2,4...$ Typically, noise in a document image has higher frequency as compared to the main text in the document image, therefore the second ($k = 2$) and third ($k = 4$) elements of the Fourier transform are selected to form the feature descriptor. DFT requires only $O(N)$ computations per element. Note that the Discrete Cosine (DC) component is obtained for $k = 0$ and is less informative. The trigonometric functions in the DFT do not have to be computed for each step, and the computation requires simple mathematical operations using the Chebyshev recurrence relation, same as the original Fourier Transform.

The RLF descriptor is thus constructed by computing the amplitude of a few elements of DFT:

$$|\mathscr{F}[f(n)](k)| = \sqrt{\Re(\mathscr{F}[f(n)](k))^2 + \Im(\mathscr{F}[f(n)](k))^2}. \tag{2}$$

The descriptor computation using only $k = 2$ suffices well for handwritten word representation under the test settings, and most importantly the descriptor is very short (length 32) with fast feature representation. However, experimentally it was found that by adding a second element for $k = 4$, the quality of the subsequent matching improved, even though the feature vector thus generated is twice as long. The advantage is that it makes it possible to sample in a smaller neighborhood, while still getting the same number of corresponding matches, with better accuracy. Nevertheless, adding a third element for $k = 6$ did not improve the accuracy significantly, and is found to be not worth the extra computational effort. This work uses RLF descriptor with length 32 for experimental analysis, taking into account the trade-off between computational cost and accuracy.

The RLF feature vector thus generated is presented in Figure 1, where x-axis denotes the feature vector length (i.e. 32 dimensions), and y-axis denote the frequency amplitudes of DFT. The advantages of the RLF descriptor are many-fold. The RLF descriptor computes a fast and short-length feature vector, to be able to perform quick feature matching in the nearest neighbor search. The RLF descriptor emphasizes on the pixels closer to the feature center, making it less sensitive to erroneous feature size estimation. It is resistant to high frequency changes, such as due to residuals from neighboring words, as it is based on the low frequency content in the local neighborhood. Nevertheless, it is insensitive to small differences in form and shape, as long as they are almost same, i.e. the low frequencies are sufficiently similar.

3.4 Feature Matching

A segmentation-free and training-free word spotting method based on the proposed RLF descriptor is studied herein. In general, no prior information is available about the potential word in the document that is to

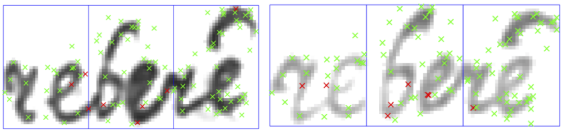

Figure 3: Matched points represented using the RLF descriptor for a sample word *reberé*. The matching keypoints (inliers) are in green, and the matches discarded (outliers) by the preconditioner are in red.

be matched with the query word. By using the RLF descriptor, the word matching problem is reduced to a much faster search problem. In this work, a simple preconditioner-based feature matching algorithm is employed.

To begin with, words are partitioned into several parts in order to avoid confusion between similar words and reduce false positives. This is to overcome the drawback of keypoint-based matching techniques [Zag17], where parts of the retrieved words may be very similar to some part of the query word, or where a word shares several letters with other different words, hence generate false positives. In the experiments, words are divided into several parts depending upon the length of the query word. For example, in Figure 3, a sample query word *reberé* and its corresponding retrieved word are divided into three parts, and the preconditioner-based matching is performed in the respective three different parts of both the words.

After the partitioning step, a nearest neighbor part-based search is performed in an optimal sliding window within the subgroups of the detected keypoints. The keypoint matching algorithm computes the extent of the the matching points in a word, and therefore is able to capture words that are partially outside the sliding window. Consequently, the matched points are removed from the set of points when a word is found, to avoid finding the same word again.

The resultant correspondences between the query word and the retrieved word in the sliding window obtained after a simple keypoint matching consists of many outliers and needs further refinement. A common approach is to use Random sample consensus (RANSAC) [Fis87] to learn transformations between the words. However, it is important to have a relaxed transformation instead, because the same word at different locations in a document can differ with small variations in font sizes, or even larger variations in a multi-writer scenario. Therefore, a deterministic preconditioner (inspired from [Has12a]) is used in this work that eliminates the need to use RANSAC and helps in removing the false matches. In [Has16], preconditioner had been used along with Putative Match Analysis (PUMA) [Has12b], which is found to be computationally expensive and increases overhead in computing false positives. To keep the matching algorithm simple yet effective, this work uses a matching algorithm that is solely based on the preconditioner.

Figure 4: Sample document images from the BH2M dataset [Fer14].

The preconditioner creates a cluster of corresponding matches in a two-dimensional space as positional vectors. This means that the correspondences between the query word and the retrieved word in the sliding window with same length and direction are potential inliers, that forms a two-dimensional cluster. However, the clusters are expected to be slightly scattered due to complex characteristics of words (e.g. words can differ in font and style), therefore the threshold must be relaxed or loosely set. The preconditioner finds the inliers efficiently and removes the outliers with fast computation speed. Figure 3 represents the matched points obtained from the proposed method, where the matching keypoints or inliers are highlighted in green and the outliers discarded by the preconditioner are in red. The preconditioner-based matching efficiently captures complex variations in handwriting by estimating the core text dimensions on-the-fly. The effectiveness of the proposed method has been experimentally demonstrated in the next section.

4 EXPERIMENTAL RESULTS

This section describes the datasets used in the experiments, and empirically evaluates the proposed method.

4.1 Datasets

For experimental analysis, the Barcelona Historical Handwritten Marriages dataset, and the Bentham dataset in two variants are taken into account. The former is heavily degraded, posing challenges for the word spotter, and the latter in both variants demonstrate multi-writer handwriting variations to a certain extent, along with document degradations. The datasets are discussed as follows:

- *Barcelona Historical Handwritten Marriages Dataset (BH2M)*: It consists of historical handwritten marriage records stored in the archives of Barcelona cathedral, written between 1617 and 1619 by a single writer in old Catalan. Figure 4 presents sample document images from the BH2M dataset. The reader is referred to [Fer14] for a deeper understanding of the dataset.

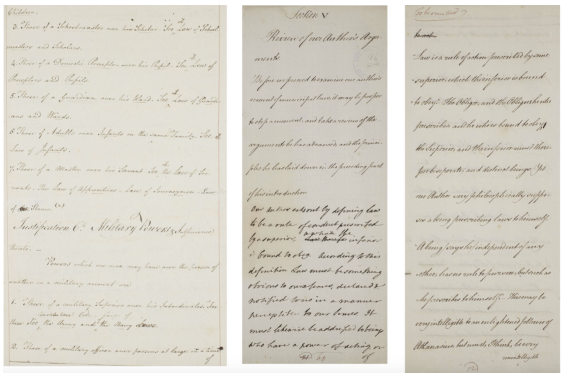

Figure 5: Sample document images from the Bentham dataset used in ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14].

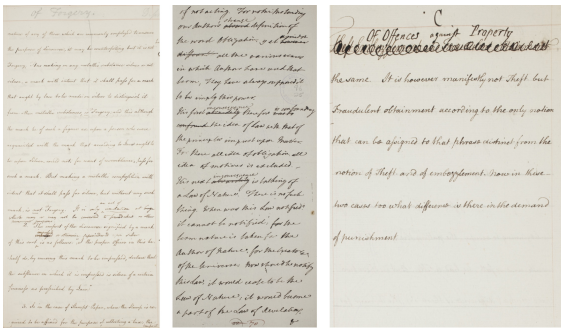

Figure 6: Sample document images from the Bentham dataset used in ICDAR 2015 Handwritten Keyword Spotting Competition [Pui15].

- *Bentham Dataset*: It consists of handwritten document pages from the Bentham collection, which have been prepared in the *tranScriptorium* project. The Bentham collection consists of manuscripts on law and moral philosophy handwritten by Jeremy Bentham (1748-1832) over a period of 60 years, and some handwritten documents from his secretarial staff. This dataset in first variant was used in ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14], and the second variant in ICDAR 2015 Handwritten Keyword Spotting Competition [Pui15]. Figure 5 and 6 present sample images from the Bentham dataset used in ICFHR 2014 and ICDAR 2015 competitions, respectively. For the experiments, all pages from both variants of Bentham dataset used in the competitions are employed, which have been written by different authors in different styles, font-sizes, and contains crossed-out words.

4.2 Results

The performance of the proposed method is empirically evaluated against the winning algorithms of ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14], and ICDAR 2015 Handwritten Keyword Spotting Competition [Pui15], along with the other state-of-the-art methods such as [Zag17, Ley09]. The evaluation measure used is the classic mean Average Precision (mAP) metric popularly used in document word spotting. In general, the retrieved regions of all the document pages are combined and re-ranked according to the score obtained. If a region overlaps more than 50% of the area of the ground truth corpora, it is classified as a positive region. The Precision and Recall values are first computed, and since a single value is preferable for comparison across different methods, the mAP of each method is calculated as the final result. A higher value of mAP is more desirable.

Method	**mAP**
[Alm12b]	0.513
[Zag17]	0.530
Proposed method	**0.783**

Table 1: Experimental results for BH2M dataset.

Method	**mAP**
[Ley09]	0.221
[How13]	0.409
[Kov14]	0.423
[Zag17]	0.517
Proposed method	**0.490**

Table 2: Experimental results for Bentham dataset used in ICFHR 2014 competition.

Method	**mAP**
PRG, TU Dortmund	0.293
CVC, Spain	0.116
[Zag17]	0.326
Proposed method	**0.786**

Table 3: Experimental results for Bentham dataset used in ICDAR 2015 competition.

Tables 1-3 present the segmentation-free handwritten word spotting results for various methods. In Table 1, the performance of the proposed method is evaluated on the BH2M dataset against the methods proposed in [Alm12b] and [Zag17]. The method proposed in [Alm12b] is based on exemplar-SVM framework for word spotting, and the method presented in [Zag17] is based on Document-oriented Local Features (DoLF). It is observed from Table 1 that the proposed method achieves higher mAP as compared to [Alm12b] and [Zag17]. This is mainly because the performance of [Alm12b] and [Zag17] is found to be weaker for challenging cases where a a word shares several letters with other different words. Typically, a higher mAP is achieved when search is performed on a long query word (e.g. *habitant*), as there is less possibility of finding the query word as part of other similar word. However, in an ideal scenario it is highly possible for a query word to share several characters with

other words, even with a longer word. A simple example of a query word from the BH2M dataset is *donsella*, where some characters are common with query words *fill* and *filla*. A much challenging case observed is the sequence of overlapping characters in the query words *fill* and *filla*, where *fill* is retrieved while searching for *filla*. The proposed method handles this effectively by dividing a word into several parts depending upon the length of the word, and then perform part-based keypoint matching. This simple approach reduces the false-positives by a significant margin, as is evident from the results in Table 1.

Table 2 presents the results obtained using different methods on the Bentham dataset from ICFHR 2014 Handwritten Keyword Spotting Competition [Pra14]. The performance of the proposed method is empirically evaluated against the state-of-the-art methods such as [Ley09], [How13], [Kov14] (i.e. winner of ICFHR 2014 competition), and [Zag17]. It is observed from Table 2 that the proposed method achieves higher mAP as compared to [Ley09], [How13] and [Kov14], and performs comparable against [Zag17] for all test images under the experimental settings. This is mainly because the relaxed nature of RLF allows it to capture more details in a degraded document image as compared to descriptors with stricter invariance properties that render them more sensitive to noise. This is important as the same query word at different locations in a document can differ with small variations in font sizes, or even larger variations in a multi-writer scenario. However, even though the proposed approach is observed to perform significantly in comparison with other methods discussed in Table 2, a mAP of 0.490 suggests further investigation. It is observed that the document images in the Bentham dataset from ICFHR 2014 competition consists of handwritten text from two or more authors, where the core text size in a document page differs across different locations in the same document page. This pose challenges for the algorithm in estimating the average core text size for each document page, as the normalization of text size might result in loss of information. The authors aim at investigating this issue further and working towards the improvement of the proposed algorithm as future work.

Table 3 evaluates the performance of the proposed method on the second variant of Bentham dataset introduced in the ICDAR 2015 Handwritten Keyword Spotting Competition [Pui15]. Unlike the first variant of the Bentham dataset discussed above, this dataset does not significantly suffer from the problem of highly variable core text size across a document page. This is evident from the higher mAP value achieved in Table 3. It is observed that the proposed method achieves higher accuracy in comparison with the winner algorithms from the competition, as well as a recent method [Zag17]. The RLF descriptor with relaxed feature description takes into account the handwriting variations to a considerable extent, and the standard core text size is estimated for each document page without significant errors.

Method	**mAP**
SIFT [Low04]	0.115
SURF [Bay08]	0.106
BRISK [Leu11]	0.035
ORB [Rub11]	0.098
KAZE [Alc12]	0.283
DoLF [Zag17]	0.517
Proposed RLF	**0.490**

Table 4: Performance evaluation of feature representation methods on Bentham dataset used in ICFHR 2014 competition.

Method	**mAP**
HoG [Alm12a]	0.584
Loci [Fer11]	0.419
Graph-based [Wan14]	0.565
FFT [Has16]	0.771
Proposed RLF	**0.783**

Table 5: Performance evaluation of feature representation methods on BH2M dataset.

In order to highlight the importance of the proposed RLF descriptor, a comparison is done with the existing feature representation methods such as SIFT [Low04], SURF [Bay08], BRISK [Leu11], Oriented FAST and Rotated BRIEF (ORB) [Rub11], KAZE [Alc12], DoLF [Zag17], HoG [Alm12a], Loci features [Fer11], graph-based [Wan14] and FFT [Has16]. Table 4 presents the experimental results to evaluate the feature representation methods used in the word spotting framework for the Bentham dataset (ICFHR 2014 competition), as an example. This is with reference to the mAP values published in a recent work [Zag17] under the given experimental set up. It is observed from Table 4 that the RLF descriptor achieves higher mAP in comparison with SIFT, SURF, BRISK, ORB and KAZE, and performs comparable against DoLF. Table 5 validates the performance of the RLF descriptor with respect to the BH2M dataset, and the experiments are performed under the same test settings where the matching algorithm is same for all feature representation methods. The RLF descriptor performs significantly in comparison with other methods, because of the advantages inherited from relaxed feature representation and efficient algorithm design. Nevertheless, with reference to the three historical handwritten datasets used in the experiments, the proposed method is observed to be most consistent and stable with high mAP.

5 CONCLUSION

This paper presented a fast and robust Radial Line Fourier descriptor, with a short feature vector of 32 dimensions, for segmentation-free and training-free handwritten word spotting. A simple preconditioner-based feature matching algorithm is employed, and the experimental results on a variety of historical document images from well-known datasets demonstrate the effectiveness of the proposed method. Under the experimental settings, the proposed RLF descriptor based method outperformed the state-of-the-art methods, including the winners of the popular keyword spotting competitions. As future work, the ideas presented herein will be scaled to aid word feature representation for heavily degraded archival databases with improvements using query expansion.

6 ACKNOWLEDGMENTS

This work was supported by the Swedish strategic research programme eSSENCE and the Riksbankens Jubileumsfond (Dnr NHS14-2068:1). The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017/7-97.

7 REFERENCES

[Ala12] Alahi, A., Ortiz, R., Vandergheynst, P. Freak: Fast retina keypoint, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 510-517, 2012.

[Alc12] Alcantarilla, P.F., Bartoli, A., Davison, A.J. Kaze features, in European Conference on Computer Vision, Springer, pp. 214-227, 2012.

[Alm12a] Almazán, J., Fernández, D., Fornés, A., Lladós, J., Valveny, E. A coarse-to-fine approach for handwritten word spotting in large scale historical documents collection, in IEEE International Conference on Frontiers in Handwriting Recognition, pp. 455-460, 2012.

[Alm12b] Almazá, J., Gordo, A., Fornés, A., Valveny, E. Efficient exemplar word spotting, 2012.

[Bay08] Baya, H., Essa, A., Tuytelaarsb, T., Van Goola, L. Speeded-up robust features (surf). Computer Vision and Image Understanding, 110(3), 346-359, 2008.

[Car02] Carneiro, G., Jepson, A.D. Phase-based local features, in European Conference on Computer Vision, Springer. pp. 282-296, 2002.

[Car03] Carneiro, G., Jepson, A.D. Multi-scale phase-based local features, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 736-743, 2003.

[Dal05] Dalal, N., Triggs, B. Histograms of oriented gradients for human detection, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[Fer11] Fernández, D., Lladós, J., Fornés, A. Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure, in Iberian Conference on Pattern Recognition and Image Analysis, Springer, pp. 628-635, 2011.

[Fer14] Fernández-Mota, D., Almazán, J., Cirera, N., Fornés, A., Lladós, J. Bh2m: The barcelona historical, handwritten marriages database, in 22nd IEEE International Conference on Pattern Recognition, pp. 256-261, 2014.

[Fis87] Fischler, M.A., Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Readings in computer vision, 726-740, 1987.

[Gat09] Gatos, B., Pratikakis, I. Segmentation-free word spotting in historical printed documents, in 10th IEEE International Conference on Document Analysis and Recognition, pp. 271-275, 2009.

[Gau11] Gauglitz, S., Höllerer, T., Turk, M. Evaluation of interest point detectors and feature descriptors for visual tracking. International Journal of Computer Vision, 94, 335-360, 2011.

[Gio17] Giotis, A.P., Sfikas, G., Gatos, B., Nikou, C. A survey of document image word spotting techniques. Pattern Recognition, 68, 310-332, 2017.

[Gir14] Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation, in IEEE conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.

[Har88] Harris, C., Stephens, M. A combined corner and edge detector, in fourth alvey vision conference, pp. 147-151, 1988.

[Has12a] Hast, A., Marchetti, A. An efficient preconditioner and a modified ransac for fast and robust feature matching, in International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'2012), pp. 11-18, 2012.

[Has12b] Hast, A., Marchetti, A. Putative match analysis: a repeatable alternative to ransac for matching of aerial images, in International Conference on Computer Vision Theory and Applications, pp. 341-344, 2012.

[Has14a] Hast, A., 2014. Robust and invariant phase based local feature matching, in 22nd IEEE International Conference on Pattern Recognition, pp. 809-814, 2014.

[Has14b] Hast, A., Marchetti, A. Invariant interest point detection based on variations of the spinor tensor, in 22nd International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'2014), pp. 49-56, 2014.

[Has16] Hast, A., Fornés, A. A segmentation-free handwritten word spotting approach by relaxed feature matching, in 12th IAPR Workshop on Document Analysis Systems, pp. 150-155, 2016.

[How13] Howe, N.R. Part-structured inkball models for one-shot handwritten word spotting, in 12th IEEE International Conference on Document Analysis and Recognition, pp. 582-586, 2013.

[Kok08] Kokkinos, I., Yuille, A. Scale invariance without scale selection, in IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[Kov14] Kovalchuk, A., Wolf, L., Dershowitz, N. A simple and fast word spotting method, in 14th IEEE International Conference on Frontiers in Handwriting Recognition, pp. 3-8, 2014.

[Leu11] Leutenegger, S., Chli, M., Siegwart, R.Y. Brisk: Binary robust invariant scalable keypoints, in IEEE International Conference on Computer Vision, pp. 2548-2555, 2011.

[Ley07] Leydier, Y., Lebourgeois, F., Emptoz, H. Text search for medieval manuscript images, Pattern Recognition, 40, 3552-3567, 2007.

[Ley09] Leydier, Y., Ouji, A., LeBourgeois, F., Emptoz, H. Towards an omnilingual word retrieval system for ancient manuscripts, Pattern Recognition, 42, 2089-2105, 2009.

[Lla12] Lladós, J., Rusinól, M., Fornés, A., Fernández, D., Dutta, A. On the influence of word representations for handwritten word spotting in historical documents, International Journal of Pattern Recognition and Artificial Intelligence, 26, 1263002, 2012.

[Low04] Lowe, D.G. Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60, 91-110, 2004.

[Mat02] Matas, J., Chum, O., Urban, M., Pajdla, T. Robust wide baseline stereo from maximally stable extremal regions, in British Machine Vision Conference, pp. 36.1-36.10, 2002.

[Pra14] Pratikakis, I., Zagoris, K., Gatos, B., Louloudis, G., Stamatopoulos, N. ICFHR 2014 competition on handwritten keyword spotting (h-kws 2014), in 14th IEEE International Conference on Frontiers in Handwriting Recognition, pp. 814-819, 2014.

[Pui15] Puigcerver, J., Toselli, A.H., Vidal, E. ICDAR2015 competition on keyword spotting for handwritten documents, in 13th IEEE International Conference on Document Analysis and Recognition, pp. 1176-1180, 2015.

[Rod08] Rodriguez, J.A., Perronnin, F. Local gradient histogram features for word spotting in unconstrained handwritten documents, in 1st International Conference on Frontiers in Handwriting Recognition, pp. 7-12, 2008.

[Rod09] Rodríguez-Serrano, J.A., Perronnin, F. Handwritten word-spotting using hidden markov models and universal vocabularies, Pattern Recognition, 42(9), 2106-2116, 2009.

[Rub11] Rublee, E., Rabaud, V., Konolige, K., Bradski, G. Orb: An efficient alternative to sift or surf, in IEEE International Conference on Computer Vision, pp. 2564-2571, 2011.

[Rus11] Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J. Browsing heterogeneous document collections by a segmentation-free word spotting method, in IEEE International Conference on Document Analysis and Recognition, pp. 63-67, 2011.

[Shi94] Shi, J., et al. Good features to track, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593-600, 1994.

[Sri05] Srihari, S., Srinivasan, H., Babu, P., Bhole, C. Handwritten arabic word spotting using the cedarabic document analysis system, in Symposium on Document Image Understanding Technology, pp. 123-132, 2005.

[Tru13] Trulls, E., Kokkinos, I., Sanfeliu, A., Moreno-Noguer, F. Dense segmentation-aware descriptors, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2890-2897, 2013.

[Vat17] Vats, E., Hast, A., Singh, P., 2017. Automatic document image binarization using bayesian optimization, in 4th International Workshop on Historical Document Imaging and Processing, pp. 89-94, 2017.

[Wan14] Wang, P., Eglin, V., Garcia, C., Largeron, C., Lladós, J., Fornés, A. A coarse-to-fine word spotting approach for historical handwritten documents based on graph embedding and graph edit distance, in 22nd IEEE International Conference on Pattern Recognition, pp. 3074-3079, 2014.

[Zag17] Zagoris, K., Pratikakis, I., Gatos, B. Unsupervised word spotting in historical handwritten document images using document-oriented local features, IEEE Transactions on Image Processing, 26(8), 4032-4041, 2017.