Keyword Spotting in Handwritten Documents using Projections of Oriented Gradients

George Retsinas[1,2], Georgios Louloudis[1], Nikolaos Stamatopoulos[1] and Basilis Gatos[1]

[1] Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-15310 Athens, Greece
{georgeretsi,louloud,nstam,bgat}@iit.demokritos.gr

[2] School of Electrical and Computer Engineering
National Technical University of Athens
GR-15773 Athens, Greece

***Abstract*—In this paper, we present a novel approach for segmentation-based handwritten keyword spotting. The proposed approach relies upon the extraction of a simple yet efficient descriptor which is based on projections of oriented gradients. To this end, a global and a local word image descriptors, together with their combination, are proposed. Retrieval is performed using to the euclidean distance between the descriptors of a query image and the segmented word images. The proposed methods have been evaluated on the dataset of the ICFHR 2014 Competition on handwritten keyword spotting. Experimental results prove the efficiency of the proposed methods compared to several state-of-the-art techniques.**

***Keywords*—Word Spotting, Feature Extraction, Projections of Oriented Gradients**

I. Introduction

Indexing of digitized documents is hindered by the lack of annotations and transcriptions and the ineffectiveness of using human resources in order to produce them. In particular, historical document indexing is a challenging task, as the more generic handwriting recognition approaches do not perform well. However, these documents can be exploited efficiently by trying to explicitly search/retrieve information in word level. Hence, a practical alternative to handwritten text recognition is keyword spotting (KWS), a very active area of research. Given a query word, either as a string (Query by String) or as an example image (Query by Example), KWS can be defined as the task of identifying the locations on an unindexed document image which have high probability to contain an instance of the query. As final output, a keyword spotting system returns a ranked list of word images.

In the literature, KWS techniques are divided into two main categories based on the considered search space: *Segmentation-Based Approaches*, which assume that each word image of the document is provided through a previously applied word segmentation procedure. Therefore, queries can be directly compared to the segmented word images [1]. *Segmentation-Free Approaches*, that do not have a prior information about the document layout and resemble template matching approaches, which try to find parts of the document that match a model/template of the query [2]. Furthermore, taking into account the existence of a training phase, keyword spotting techniques are also categorized into *learning-based* [3],[4] and *learning-free* [5],[6] approaches. In this work, we address the Query by Example keyword spotting problem with a learning-free, segmentation-based approach.

In handwritten documents, KWS is a challenging task due to the vast variability of different writing styles. This problem is even more noticeable in learning-free approaches, because, ideally, these approaches try to simulate all possible variations from a single word instance (i.e. the query). A common system for a segmentation-based keyword spotting task consists of the following steps. First, a preprocessing is applied to each word image, where many of the aforementioned variabilities are absorbed. Next, a feature extraction step is performed, where each normalized image is represented by a set of descriptive features. Finally, a ranking scheme is used for retrieval, where a similarity measure between the features of the query image and the features of the segmented word images is introduced. It is obvious that the selection of a similarity measure is heavily dependent on the nature of the previously extracted features.

Focusing on the feature extraction step, we can distinguish two main approaches: *shape/appearance features* and *structural features*. In [7], the authors proposed that each word image can be described as a sequence of simple geometric (statistical) features, computed at each column of the image, followed by the application of Dynamic Time Warping algorithm between sequences for producing the ranking. Other approaches extract fixed-sized descriptors for each image, using shape-appearance techniques (e.g. HOG and LBP descriptors [5]), and perform the retrieval task using a simple distance measure. These approaches provide efficient retrieval time, which is essential for large collections of documents, due to their simplicity. Another approach is to extract structural characteristics, leading to descriptive models of the inner structure of a word (e.g. the inkball model presented in [6]). These models are, in theory, more robust to variations of the

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.61

writing style and can efficiently describe deformations at the expense of larger time requirements for retrieval due to the complexity of the involved distance measures.

The proposed keyword spotting method is based on Projection of Oriented Gradients (POG), which has been successfully used in character recognition and displayed robustness to handwritten character variations [8]. POG is a projection-based descriptor that tries to encode crucial information about the edges in a global manner. In this paper, two different approaches are considered in the direction of constructing a word descriptor using POGs.

First, we propose a global approach by adjusting the detail of information retained in each projection in accordance with the (more prolonged) shape of a word, i.e. horizontal projections will be more descriptive than vertical ones. Additionally, we also propose a local approach by dividing the word image into a fixed number of overlapping segments and performing a character-like POG feature extraction for each segment. The main idea of this approach is to segment a word image at character level, in order to use the POG implementation of [8]. It is obvious that the proposed descriptors are of fixed size and thus the retrieval is performed by nearest neighbor search using the euclidean distance. Finally we consider a fusion, performed in the ranking step, of the aforementioned approaches.

The remainder of the paper is organized as follows. In Section II, two novel methods for KWS, as well as their fusion, are proposed, while experimental results are presented in Section III. Finally, conclusions and future directions are drawn in Section IV.

II. Proposed Method

Two approaches for extracting word image descriptors using projections of oriented gradients are proposed. In the upcoming subsections we describe the preprocessing step, we provide a brief description of Projections of Oriented Gradients and finally we present the proposed methods. The retrieval output of our system for a query is a sorted list of the euclidean distances between the query and the segmented word images.

A. Preprocessing

The preprocessing step consists of image binarization, skew correction and height normalization. A normalization of the word image to a fixed size is not necessary, because the POG descriptor is size invariant.

Due to the fact that the POG method is currently restricted to binary images, a *binarization* is imperative for non-binary word images. For the image binarization procedure, we choose Sauvola's method [9].

Furthermore, we employ a *robust regression* method in order to find the main zone of the word. The algorithm that we use for this task is a regression procedure based on iteratively reweighted least squares, using a bisquare weighting function [10]. Essentially, the algorithm finds a set of inliers foreground pixels (and the corresponding outliers) that best describe the word as a line. The result of this procedure consists of a line model, i.e. the parameters a, b of the fitted line $y = ax + b$, as well as the corresponding width of the line, as a variance parameter. Therefore, the word's main zone is the area within the resulting thickened line, which excludes outliers that, ideally, correspond to the ascenders and descenders of the word. Given the line model and its main zone (denoted by a baseline and an upperline as shown in Fig. 1(a),(b),(c)), we perform the following normalization steps:

- **Skew Correction:** Given the parameters of the fitted line, it is trivial to compute the rotation angle of the main zone and subsequently deskew the image.
- **Height Normalization:** We place the main zone in the center of the generated normalized image (y-axis), which will promote the global approach of the proposed feature extraction scheme. Additionally, in order to avoid extreme ascenders and descenders, that contain no useful information, we crop the image using a threshold of the vertical distance from the main zone. The threshold is dependent to the width of the main zone w_l, i.e. the normalized image has a margin of $1.4 \times w_l$ pixels under and over the main zone (see Fig. 1(d),(e),(f)).

Overall, the preprocessing step consists of simple and cost-effective stages in order to obtain normalized word images with respect to the spatial distribution of the word pixels into the image (centralized and deskewed words).

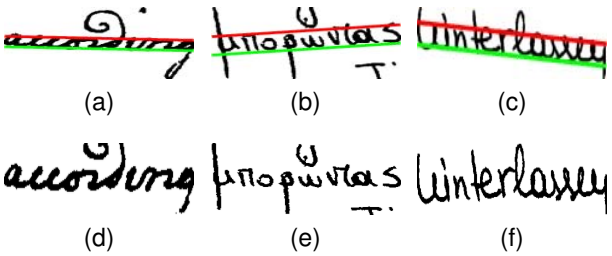

Fig. 1. Word image normalization: (a),(b),(c) are the initial word images and their main zone marked with a baseline and a upperline and (d),(e),(f) are the normalized images after skew correction and height normalization

B. Projections Of Oriented Gradients

Projections of Oriented Gradients, that we presented in [8], have performed well in the character classification task and therefore were chosen as the main descriptor of our proposed methodology in order to describe a word image. The overview of the POG method is depicted in Fig. 2. The steps for extracting the POG descriptor are briefly described below.

Gradient Orientation: This stage is essential for capturing informative details of the image, such as the change in the direction of edges. The edge information is captured with the use of the directional gradients along x-axis (horizontal) G_x and y-axis (vertical) G_y of the image $I(x, y)$, which are computed using the filter kernels $[-1\,0\,1]$ and $[-1\,0\,1]^T$. In order to compute the gradient orientation at each pixel, a transformation into polar coordinates is performed through

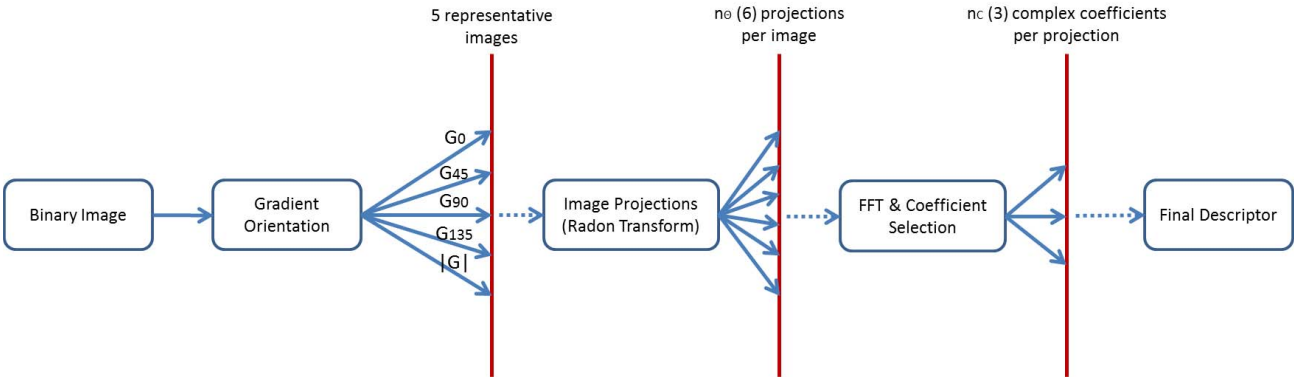

Fig. 2. Overview of the Projections of Oriented Gradients. First, 5 representative images are generated from each binary image corresponding to the gradient orientation. Then, for each image, angular projections are extracted in a Radon-like procedure. Finally, each projection is encoded with a set of low-frequency complex coefficients of its FFT. The final descriptor is the concatenation of all the extracted coefficients.

Equations 1,2. A wrapping is performed so that the orientation values lie on the interval $[0, 180°)$.

$$\|G(x,y)\|_2 = \sqrt{(G_x^2(x,y) + G_y^2(x,y))} \quad (1)$$

$$\angle G(x,y) = arctan\Big(\frac{G_y(x,y)}{G_x(x,y)}\Big) \quad (2)$$

The possible orientations, due to the fact that the selected gradient filter is applied to a binary image, are equal to four (0°, 45°, 90° and 135°) and thus four binary images, that represent the gradient orientation, are constructed: G_0, G_{45}, G_{90} and G_{135}, where $G_\theta = (\angle G(x,y) = \theta°)$. The gradient orientation images along with the gradient magnitude, $\|G\|$, are called "representative" images.

Projections: The basic concept of the projection-based feature extraction is to decompose the binary image into several projections under selected angles, imitating the Radon transform. The projections are n_θ in total, sampled every $180°/n_\theta$, thus the projection angles are:

$$\theta_k = k\frac{180°}{n_\theta}, \quad k \in [0, n_\theta - 1] \quad (3)$$

FFT & Coefficient Selection: To simplify each projection, we keep only the descriptive information corresponding to smoothed regions of relatively high pixel concentration, or equivalently, to the low frequency components of the projection. 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, while the remaining are discarded. Subsequently, a normalization with regard to the number of pixels in each image is applied by dividing each Fourier coefficient by N (i.e. the number of foreground pixels which corresponds to c_0). Selecting a subset of the Fourier coefficients results to projection length independence and, consequently, to image size independence. The final descriptor of a projection is the concatenation of the real and imaginary parts of the following complex feature vector:

$$f_j = c_j/N\,, \quad j = 1, \ldots, n_c \quad (4)$$

Final Descriptor: The final feature vector is the concatenation of the coefficients for every projection of the representation images (G_0,G_{45},G_{90},G_{135} and $\|G\|$), as it is depicted in Fig. 2. Overall, the length of the descriptor is: 5 (images) $\times$ n_θ (projections) $\times$ $2n_c$ (Fourier coefficients).

It should be noted that we can reconstruct approximately the original (representative) image from the POG descriptor via the inverse Radon transform, after interpolating each projection to a specific length using the inverse Fourier transform. Hence, a normalized approximation of the image is generated. Examples of this visualization are depicted at the bottom row of the Figure 3.

C. Word Image Descriptors based on POG

Having described the Projections of Oriented Gradients approach, we will propose two simple adjustments of this method in order to cope with the wider word images compared to character images.

1) Global POG word descriptor (*gPOG*) The first proposed method is a global approach, very similar to the simple POG method, where we adjust the detail of the retained information in each projection. This adjustment is implemented by keeping a different number of coefficients for each projection, i.e we modify only the *coefficient selection* step of the POG method. We choose to retain the number of projections to 6, as in [8], i.e. the projections are extracted at the angles $\{0°, 30°, 60°, 90°, 120°, 150°\}$. Furthermore, we assume that the vertical projection corresponds to the height of a character, thus selecting 3 coefficients [8], while 6 coefficients are selected for describing the horizontal projection. As a result, the numbers of selected coefficients, which correspond to the projection angles, are chosen as {6,7,5,3,5,7}, retaining the fluctuation of the corresponding projection lengths. The largest number of coefficients (7) are selected at 30° (and 150°), because this projection is most probably close to the diagonal of the image.

2) k-segmented POG descriptor (*lPOG*): An alternative to the global approach of the aforementioned method is a

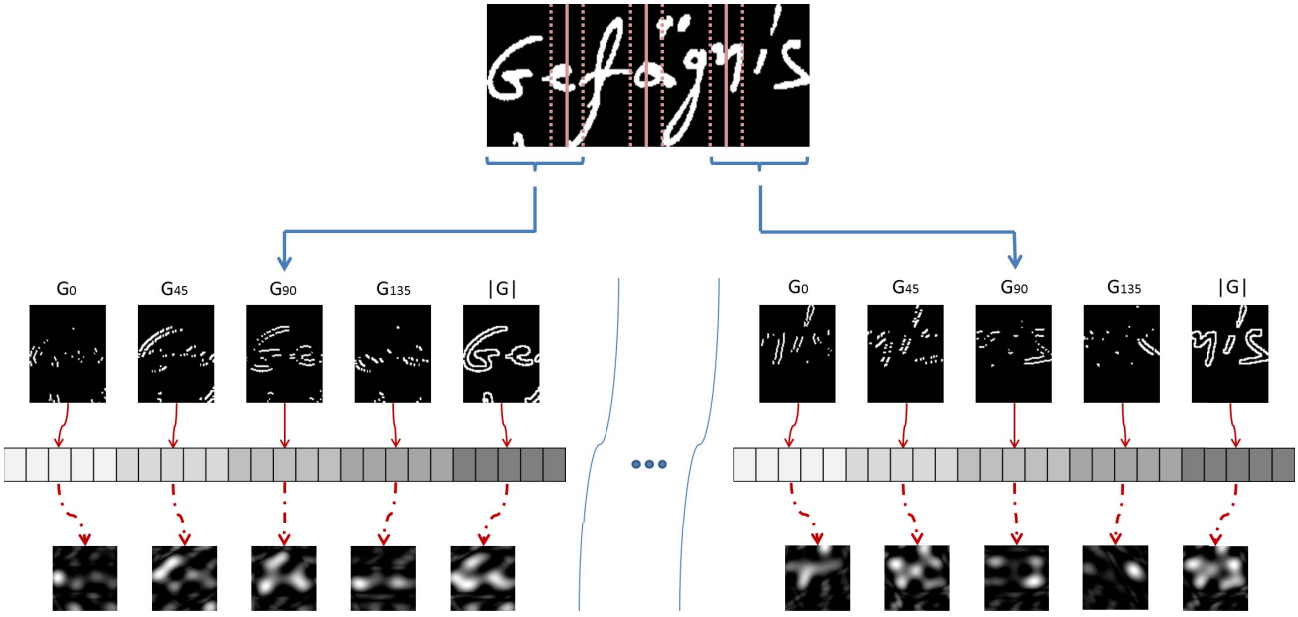

Fig. 3. Overview of the Projections of Oriented Gradient feature extraction for the case of the k-segmented image approach. Each segmented part is used for extracting a conventional POG descriptor. The last row corresponds to the generated images after the reconstruction and visualizes/highlights the stored information in the proposed descriptor.

local approach, where we divide the initial word image into k smaller sub-images. The image is divided along the horizontal axis into a fixed number of overlapping images (20% overlap between neighboring sub-images), trying to reproduce a draft character segmentation in order to apply a simple POG feature extraction to each sub-image. For the experimental part, we choose $k = 4$. An overview of this approach is depicted in Figure 3.

3) Fusion (*fPOG*): The last proposed variation is a combination of the aforementioned descriptors, trying to exploit the benefits of both the global and the local proposed approaches. After the extraction of the descriptors which correspond to the aforementioned methods, the retrieval sorted list is computed using a combination of the respective distances between a query descriptor and a word descriptor. We assume that the distance values for each descriptor contribute equivalently to the final result, thus the combined distance measure $d(q, w)$ of a query image q and a word image w is defined as:

$$d(q,w) = 0.5 \times \frac{\|q1 - w1\|_2}{N1} + 0.5 \times \frac{\|q2 - w2\|_2}{N2} \qquad (5)$$

where q_1, w_1 are the descriptors of N1 features generated from the first approach (gPOG) and $q2$,$w2$ the descriptors of N2 features generated from the second approach (sPOG). The division of the distances with the corresponding length of the descriptors (N1, N2) provides a length normalization in order to achieve comparable distance values for each method.

III. Experimental Results

A. Data Setup

The evaluation of the proposed methodology is performed on both datasets from the ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014) [11] for the segmentation-based track. The datasets are briefly described below:

Bentham Dataset: It consists of handwritten manuscripts in English written by Jeremy Bentham himself as well as by Bentham's secretarial staff [12]. For the segmentation-based track of the competition this dataset consists of 320 image queries and ∼10.000 segmented word images from 50 document images.

Modern Dataset: It consists of modern handwritten documents from the ICDAR 2009 Handwritten Segmentation Contest [13] written in four languages (English, French, German and Greek). For the segmentation-based track of the competition this dataset consists of 300 image queries and ∼15.000 segmented word images from 100 document images (25 for each language).

B. Compared Methods

In order to evaluate the efficiency of the proposed methodology, we compare it to the methods that competed in the segmentation-based track of the H-KWS 2014 contest. These three methods are briefly described below:

G1: (Kovalchuk et al., [5]) This method is based on the extraction of HOG and LBP features, after each word image is resized into a fixed rectangle, resulting to a very large descriptor. A cosine similarity operator and maximum pooling

TABLE I
EVALUATION METRICS FOR BOTH BENTHAM AND MODERN DATASETS

	Bentham Dataset				Modern Dataset			
Method	**P@5**	**MAP**	**BNDCG**	**NDCG**	**P@5**	**MAP**	**BNDCG**	**NDCG**
G1	0.738	0.524	0.742	0.762	0.588	0.338	0.611	0.612
G2	0.724	0.513	0.744	0.764	0.706	0.523	0.757	0.757
G3	0.718	0.462	0.638	0.657	0.569	0.278	0.484	0.485
gPOG	0.758	0.553	0.773	0.7749	0.569	0.328	0.629	0.629
lPOG	0.768	0.564	0.780	0.782	0.591	0.326	0.632	0.632
fPOG	0.771	0.577	0.789	0.791	0.613	0.355	0.654	0.654

is used to reduce the dimensionality of the features to a rather small descriptor (250D). Retrieval is performed by ranking the target words with respect to their euclidean distance.

G2: (Almazan et al., [4]) A Fisher Vector descriptor is extracted from each word image, while the transcription of each word is encoded into a pyramidal histogram of characters (PHOC). The image descriptors and the transcription descriptors are used to learn a projection to an attribute vector space. Finally, Canonical Correlation Analysis is utilized to further improve the efficiency of their approach. It should be noted, that a training set is required, where each word image is to be supplied with a transcription, thus the training phase is performed in an independent dataset with similar writing style.

G3: (Howe, [6]) This method is based on a flexible inkball (template) model which allows deformed template matching. Query models are fitted to the target words and each target word is converted to such a model for reverse verification. Retrieval is performed by ranking the target words according to the two-way match scores.

C. Evaluation Metrics

The measures chosen for the performance evaluation of the proposed methods correspond to the metrics used in H-KWS 2014 competition [11]. In more detail, we used the following evaluation measures:

- **P@5:** Precision at top 5 retrieved words. P@k measure is defined as:

$$\text{P@k} = \frac{|\{\text{relevant words}\} \cap \{\text{k retrieved words}\}|}{|\{\text{k retrieved words}\}|} \quad (6)$$

- **MAP:** Mean Average Precision. The MAP for a set of queries is the mean of the average precision scores for each query. The Average Precision for a query is defined as:

$$\text{AP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{|\{\text{relevant words}\}|} \quad (7)$$

  where rel(k) takes the value 1 if a word is relevant at rank k and the value 0 otherwise.
- **NDCG:** Normalized Discounted Cumulative Gain. The main concept of this metric is to introduce a penalty when highly relevant words appear lower in the retrieval list. The NDCG is defined as:

$$\text{NDCG} = \frac{DCG}{IDCG}, \quad \text{DCG} = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2(i)} \quad (8)$$

  where rel_i is the relevance judgment at position i and IDGG is the ideal DCG which corresponds to the groundtruth.
- **BNDCG:** Binary Normalized Discounted Cumulative Gain. Same as NDCG, except that the relevances are binary, i.e. either 0 or 1.

D. Experimental Results

Following the ICFHR 2014 H-KWS Competition evaluation, we experimented on both Bentham and Modern datasets, using the three proposed approaches and we compare them to the competing methods. The retrieval results are presented in Table I, along with the evaluation metrics of the compared methods, while in Fig. 4 the Precision-Recall Curves are shown, using only the proposed fusion method, which performed best amongst the proposed approaches.

In both datasets, the *lPOG* method outperforms the *gPOG* method and their combination *fPOG* exhibits the best performance over all evaluation metrics. These results demonstrate the effectiveness of using a combination of the two alternative approaches, i.e. a global and a local descriptor.

It can be observed that in the Bentham dataset, both proposed methods and their fusion outperform the compared methods on all the evaluation metrics. These results prove the efficiency of the proposed approaches, even compared to more complex descriptors (G1), flexible description models (G3) and methods that include a training phase (G2). However, the Modern dataset is more challenging, as it consists of many writers and writing styles in different languages. This leads to a general drop on the evaluation metrics compared to the Bentham dataset (with the exception of G2 method). Only the fusion method exhibits steadily better performance from G1 and G3 methods in the Modern dataset, which are the directly comparable methods, as they are learning-free techniques. The G2 method, proposed by Almazan, displays far better results in the Modern dataset. However, Almazan uses a training phase and thus this approach is more powerful than learning-free approaches. Even though the training was performed on a different dataset (for the case of Modern dataset it was trained on the IAM dataset), still this method is not directly comparable to the other participating methods.

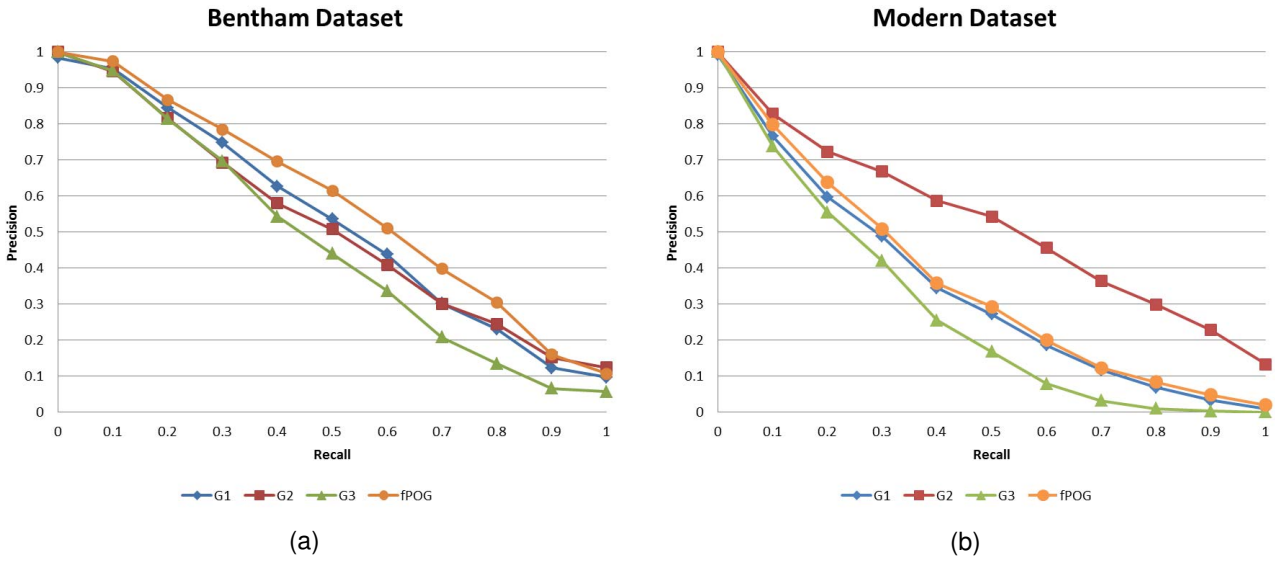

Fig. 4. Precision-Recall Curves for (a) Bentham and (b) Modern Datasets

It should be noted that if the proposed fusion technique (fPOG) was participating in the H-KWS competition,following the same score evaluation (i.e. summing the ranking positions over all evaluation metrics), it would be ranked first.

IV. Conclusion

In this paper, we present two novel methods and their fusion for learning-free, segmentation-based keyword spotting. The proposed approach relies upon the extraction of a simple yet efficient descriptor which is based on projections of oriented gradients. To this end, a global and a local word image descriptors, together with their combination, are proposed. As the experimental results indicate, the proposed word descriptors outperform other learning-free state-of-the-art techniques for the task of keyword spotting. A possible future extension of the presented work, will be the adjustment of the proposed method for the task of segmentation-free keyword spotting, as well as the addition of a training phase in order to further utilize the effectiveness of the proposed descriptors.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 600707 (project tranScriptorium). This work has been also supported by the OldDocPro project (ID 4717) funded by the GSRT.

References

[1] M. Greibus and L. Telksnys, "Speech Keyword Spotting with Rule Based Segmentation", 19th International Conference on Information and Software Technologies, pp. 186-197, Lithuania, 2013.

[2] M. Rusinol, D. Aldavert, R. Toledo and J. Llados, "Efficient segmentation-free keyword spotting in historical document collections", Pattern Recognition, Vol. 48, No. 2, 2015, pp. 545-555.

[3] A. H. Tosseli and E. Vidal, "Word-Graph Based Handwriting Key-Word Spotting: Impact of Word-Graph Size on Performance", 11th International Workshop on Document Analysis Systems, pp. 176-180, France, 2014.

[4] J. Almazan, A. Gordo, A. Fornes and E. Valveny, "Handwritten Word Spotting with Corrected Attributes", 15th International Conference on Computer Vision, pp. 1017-1024, Australia, 2013.

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

[6] N. R. Howe, "Part-structured inkball models for one-shot handwritten keyword spotting", 12th International Conference on Document Analysis and Recognition, pp. 582-586, USA, 2013.

[7] T . M. Rath and R. Manmatha, "Word spotting for historical documents", International Journal on Document Analysis and Recognition, Vol. 9, No. 2-4, pp. 139-152, 2007.

[8] G. Retsinas, B. Gatos, N. Stamatopoulos and G. Louloudis, "Isolated Character Recognition using Projections of Oriented Gradients", 13th International Conference on Document Analysis and Recognition, pp. 336-340, France, 2015.

[9] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", Pattern Recognition, Vol. 33, No. 2, pp. 224-236, 2000.

[10] P. W. Holland and R. E. Welsch, "Robust Regression Using Iteratively Reweighted Least-Squares", Communications in Statistics: Theory and Methods, pp. 813-827, 1977.

[11] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis and N. Stamatopoulos, "ICFHR 2014 Competition on Handwritten KeyWord Spotting (H-KWS 2014)", 14th International Conference on Frontiers in Handwriting Recognition, pp. 814-819, Greece, 2014.

[12] D. G. Long, "The manuscripts of Jeremy Bentham: a chronological index to the collection in the Library of University College, London: based on the catalogue by A. Taylor Milne", 1981.

[13] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR2009 handwriting segmentation contest", International Journal on Document Analysis and Recognition, Vol.14, No. 1, pp. 25-33, 2011.